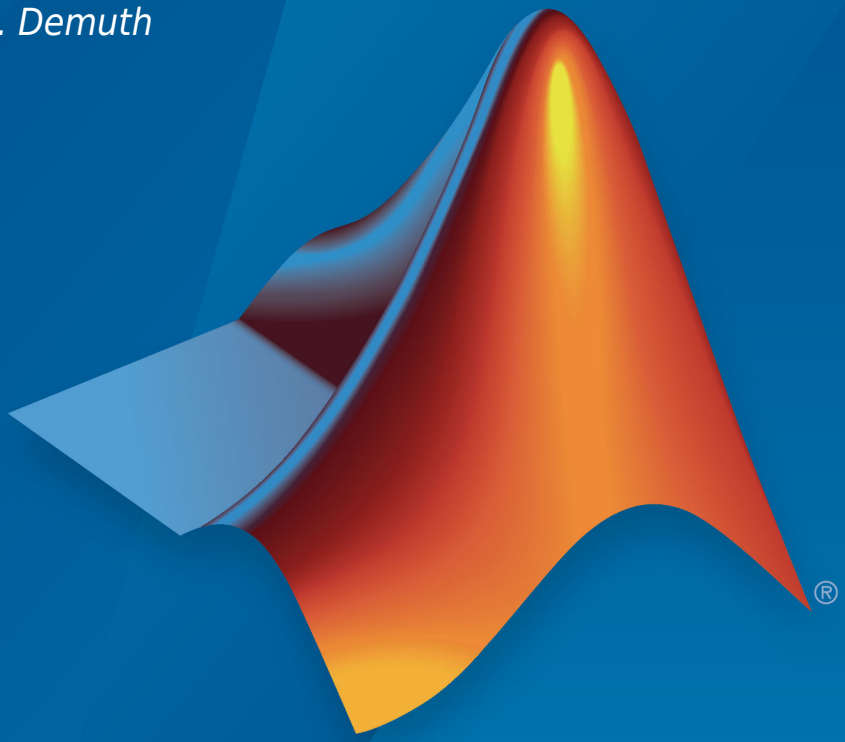
Deep Learning Toolbox™

Getting Started Guide

Mark Hudson Beale
Martin T. Hagan
Howard B. Demuth



MATLAB®

R2019**a**

MathWorks®

How to Contact MathWorks

Latest news: `www.mathworks.com`

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

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

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

Phone: 508-647-7000

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

Deep Learning Toolbox™ Getting Started Guide

© COPYRIGHT 1992–2019 by The MathWorks, Inc.

The software described in this document is furnished under a license agreement. The software may be used or copied only under the terms of the license agreement. No part of this manual may be photocopied or reproduced in any form without prior written consent from The MathWorks, Inc.

FEDERAL ACQUISITION: This provision applies to all acquisitions of the Program and Documentation by, for, or through the federal government of the United States. By accepting delivery of the Program or Documentation, the government hereby agrees that this software or documentation qualifies as commercial computer software or commercial computer software documentation as such terms are used or defined in FAR 12.212, DFARS Part 227.72, and DFARS 252.227-7014. Accordingly, the terms and conditions of this Agreement and only those rights specified in this Agreement, shall pertain to and govern the use, modification, reproduction, release, performance, display, and disclosure of the Program and Documentation by the federal government (or other entity acquiring for or through the federal government) and shall supersede any conflicting contractual terms or conditions. If this License fails to meet the government's needs or is inconsistent in any respect with federal procurement law, the government agrees to return the Program and Documentation, unused, to The MathWorks, Inc.

Trademarks

MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See `www.mathworks.com/trademarks` for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

Patents

MathWorks products are protected by one or more U.S. patents. Please see `www.mathworks.com/patents` for more information.

Revision History

June 1992	First printing	
April 1993	Second printing	
January 1997	Third printing	
July 1997	Fourth printing	
January 1998	Fifth printing	Revised for Version 3 (Release 11)
September 2000	Sixth printing	Revised for Version 4 (Release 12)
June 2001	Seventh printing	Minor revisions (Release 12.1)
July 2002	Online only	Minor revisions (Release 13)
January 2003	Online only	Minor revisions (Release 13SP1)
June 2004	Online only	Revised for Version 4.0.3 (Release 14)
October 2004	Online only	Revised for Version 4.0.4 (Release 14SP1)
October 2004	Eighth printing	Revised for Version 4.0.4
March 2005	Online only	Revised for Version 4.0.5 (Release 14SP2)
March 2006	Online only	Revised for Version 5.0 (Release 2006a)
September 2006	Ninth printing	Minor revisions (Release 2006b)
March 2007	Online only	Minor revisions (Release 2007a)
September 2007	Online only	Revised for Version 5.1 (Release 2007b)
March 2008	Online only	Revised for Version 6.0 (Release 2008a)
October 2008	Online only	Revised for Version 6.0.1 (Release 2008b)
March 2009	Online only	Revised for Version 6.0.2 (Release 2009a)
September 2009	Online only	Revised for Version 6.0.3 (Release 2009b)
March 2010	Online only	Revised for Version 6.0.4 (Release 2010a)
September 2010	Tenth printing	Revised for Version 7.0 (Release 2010b)
April 2011	Online only	Revised for Version 7.0.1 (Release 2011a)
September 2011	Online only	Revised for Version 7.0.2 (Release 2011b)
March 2012	Online only	Revised for Version 7.0.3 (Release 2012a)
September 2012	Online only	Revised for Version 8.0 (Release 2012b)
March 2013	Online only	Revised for Version 8.0.1 (Release 2013a)
September 2013	Online only	Revised for Version 8.1 (Release 2013b)
March 2014	Online only	Revised for Version 8.2 (Release 2014a)
October 2014	Online only	Revised for Version 8.2.1 (Release 2014b)
March 2015	Online only	Revised for Version 8.3 (Release 2015a)
September 2015	Online only	Revised for Version 8.4 (Release 2015b)
March 2016	Online only	Revised for Version 9.0 (Release 2016a)
September 2016	Online only	Revised for Version 9.1 (Release 2016b)
March 2017	Online only	Revised for Version 10.0 (Release 2017a)
September 2017	Online only	Revised for Version 11.0 (Release 2017b)
March 2018	Online only	Revised for Version 11.1 (Release 2018a)
September 2018	Online only	Revised for Version 12.0 (Release 2018b)
March 2019	Online only	Revised for Version 12.1 (Release 2019a)

Contents

Acknowledgments

Acknowledgments . . . . . **viii**

1 Getting Started

Deep Learning Toolbox Product Description **1-2**
Key Features **1-2**

Get Started with Deep Network Designer **1-3**

Try Deep Learning in 10 Lines of MATLAB Code **1-12**

Classify Image Using Pretrained Network **1-15**

Get Started with Transfer Learning **1-18**

Create Simple Image Classification Network **1-22**

Create Simple Sequence Classification Network **1-26**

Shallow Networks for Pattern Recognition, Clustering and Time Series **1-30**
Shallow Network Apps and Functions in Deep Learning Toolbox **1-31**
Deep Learning Toolbox Applications **1-32**
Shallow Neural Network Design Steps **1-34**

Fit Data with a Shallow Neural Network **1-35**
Defining a Problem **1-35**

Using the Neural Network Fitting App 1-36
Using Command-Line Functions 1-51

Classify Patterns with a Shallow Neural Network 1-61
Defining a Problem 1-61
Using the Neural Network Pattern Recognition App 1-62
Using Command-Line Functions 1-79

Cluster Data with a Self-Organizing Map 1-86
Defining a Problem 1-86
Using the Neural Network Clustering App 1-86
Using Command-Line Functions 1-99

Shallow Neural Network Time-Series Prediction and Modeling 1-106
Defining a Problem 1-106
Using the Neural Network Time Series App 1-107
Using Command-Line Functions 1-123

Train Shallow Networks on CPUs and GPUs 1-133
Parallel Computing Toolbox 1-133
Parallel CPU Workers 1-134
GPU Computing 1-134
Multiple GPU/CPU Computing 1-135
Cluster Computing with MATLAB Parallel Server 1-135
Load Balancing, Large Problems, and Beyond 1-135

Sample Data Sets for Shallow Neural Networks 1-137

Shallow Neural Networks Glossary

Acknowledgments

Acknowledgments

The authors would like to thank the following people:

Joe Hicklin of MathWorks for getting Howard into neural network research years ago at the University of Idaho, for encouraging Howard and Mark to write the toolbox, for providing crucial help in getting the first toolbox Version 1.0 out the door, for continuing to help with the toolbox in many ways, and for being such a good friend.

Roy Lurie of MathWorks for his continued enthusiasm.

Mary Ann Freeman of MathWorks for general support and for her leadership of a great team of people we enjoy working with.

Rakesh Kumar of MathWorks for cheerfully providing technical and practical help, encouragement, ideas and always going the extra mile for us.

Alan LaFleur of MathWorks for facilitating our documentation work.

Stephen Vanreusel of MathWorks for help with testing.

Dan Doherty of MathWorks for marketing support and ideas.

Orlando De Jesús of Oklahoma State University for his excellent work in developing and programming the dynamic training algorithms described in “Time Series and Dynamic Systems” and in programming the neural network controllers described in “Neural Network Control Systems”.

Martin T. Hagan, **Howard B. Demuth**, and **Mark Hudson Beale** for permission to include various problems, examples, and other material from Neural Network Design, January, 1996.

Getting Started

- “Deep Learning Toolbox Product Description” on page 1-2
- “Get Started with Deep Network Designer” on page 1-3
- “Try Deep Learning in 10 Lines of MATLAB Code” on page 1-12
- “Classify Image Using Pretrained Network” on page 1-15
- “Get Started with Transfer Learning” on page 1-18
- “Create Simple Image Classification Network” on page 1-22
- “Create Simple Sequence Classification Network” on page 1-26
- “Shallow Networks for Pattern Recognition, Clustering and Time Series” on page 1-30
- “Fit Data with a Shallow Neural Network” on page 1-35
- “Classify Patterns with a Shallow Neural Network” on page 1-61
- “Cluster Data with a Self-Organizing Map” on page 1-86
- “Shallow Neural Network Time-Series Prediction and Modeling” on page 1-106
- “Train Shallow Networks on CPUs and GPUs” on page 1-133
- “Sample Data Sets for Shallow Neural Networks” on page 1-137

Deep Learning Toolbox Product Description

Create, analyze, and train deep learning networks

Deep Learning Toolbox provides a framework for designing and implementing deep neural networks with algorithms, pretrained models, and apps. You can use convolutional neural networks (ConvNets, CNNs) and long short-term memory (LSTM) networks to perform classification and regression on image, time-series, and text data. Apps and plots help you visualize activations, edit network architectures, and monitor training progress.

For small training sets, you can perform transfer learning with pretrained deep network models (including SqueezeNet, Inception-v3, ResNet-101, GoogLeNet, and VGG-19) and models imported from TensorFlow™-Keras and Caffe.

To speed up training on large datasets, you can distribute computations and data across multicore processors and GPUs on the desktop (with Parallel Computing Toolbox™), or scale up to clusters and clouds, including Amazon EC2® P2, P3, and G3 GPU instances (with MATLAB® Parallel Server™).

Key Features

- Convolutional neural network (CNN), long short-term memory (LSTM), and directed acyclic graph (DAG) network topologies
- Transfer learning with pretrained models and models imported from TensorFlow-Keras and Caffe
- Import and export of ONNX™ models for interoperability with other deep learning frameworks
- App for graphically designing and analyzing network architectures
- Training and inference with CPUs and GPUs on desktops, clusters, and clouds (including Amazon EC2 P2, P3, and G3 GPU instances)
- Unsupervised learning algorithms for shallow networks, including self-organizing maps and competitive layers
- Supervised learning algorithms for shallow networks, including multilayer, radial basis, learning vector quantization (LVQ), time-delay, and nonlinear autoregressive (NARX)

Get Started with Deep Network Designer

This example shows how to fine-tune a pretrained GoogLeNet network to classify a new collection of images. This process is called transfer learning and is usually much faster and easier than training a new network, because you can apply learned features to a new task using a smaller number of training images. To interactively prepare a network for transfer learning, use Deep Network Designer.

Load Pretrained Network

Load a pretrained GoogLeNet network. If you need to download the network, use the download link.

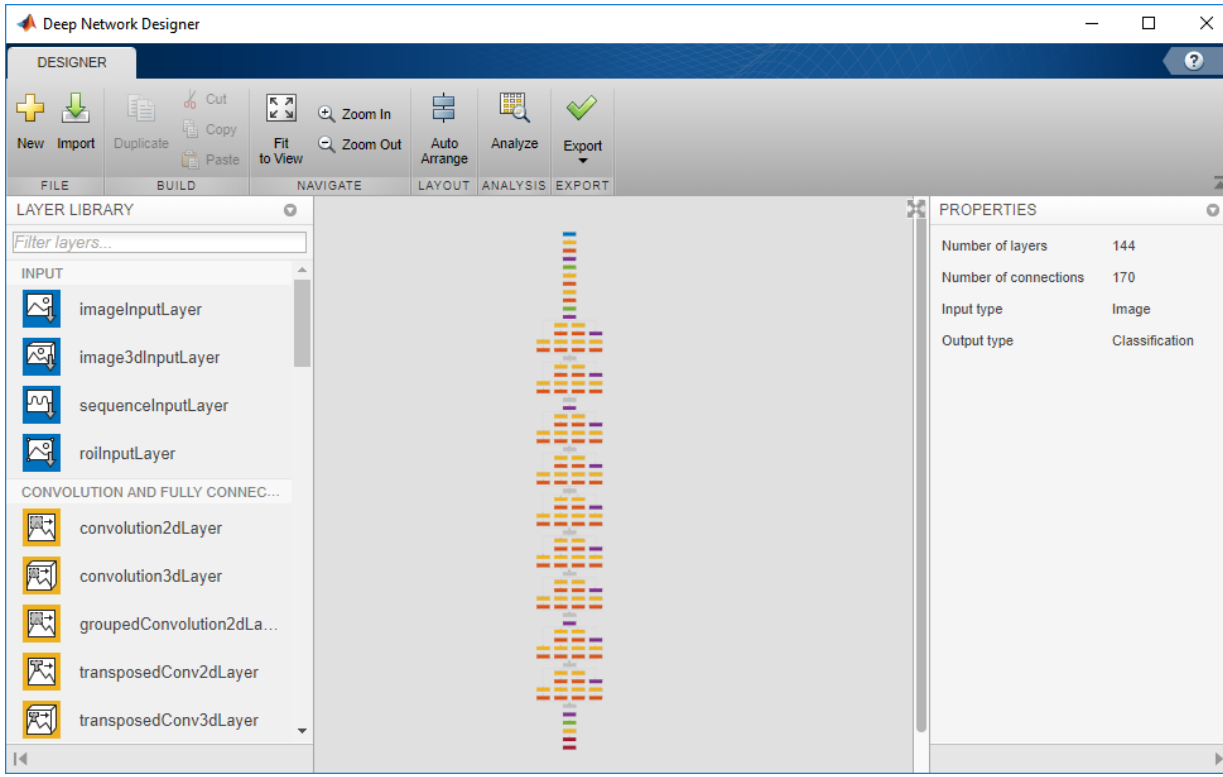
```
net = googlenet;
```

Import Network into Deep Network Designer

Open Deep Network Designer.

```
deepNetworkDesigner
```

Click **Import** and select the network from the workspace. Deep Network Designer displays a zoomed out view of the whole network. Explore the network plot. To zoom in with the mouse, use **Ctrl**+scroll wheel.

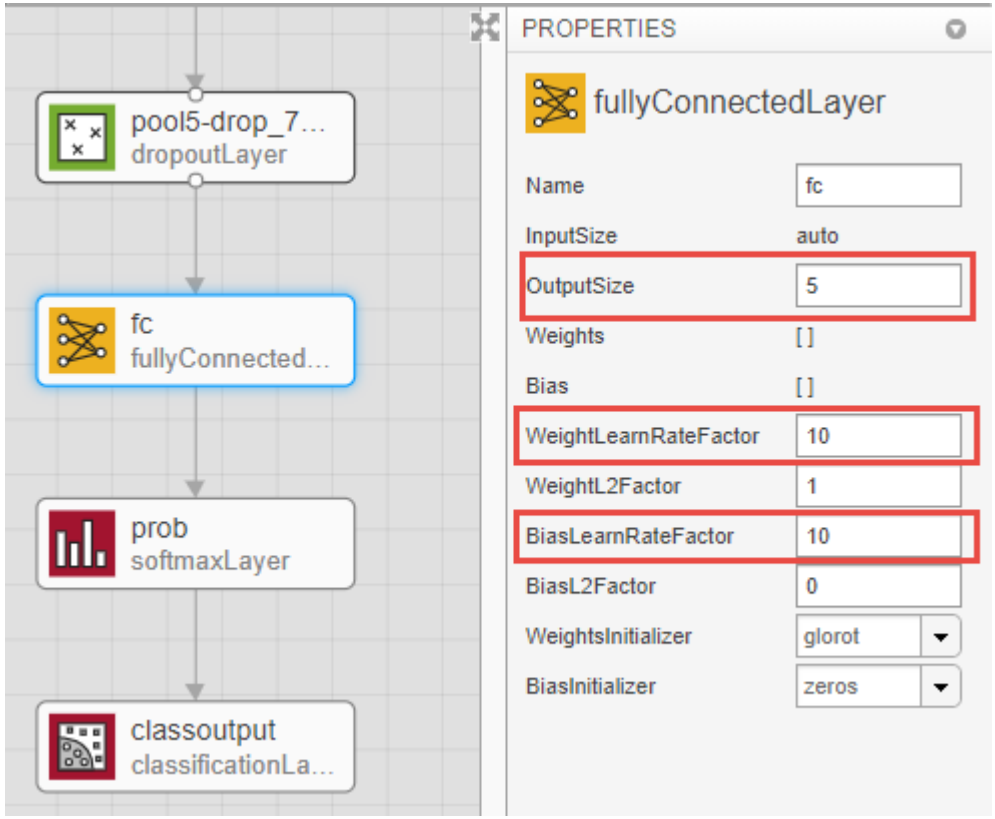


Edit Network for Transfer Learning

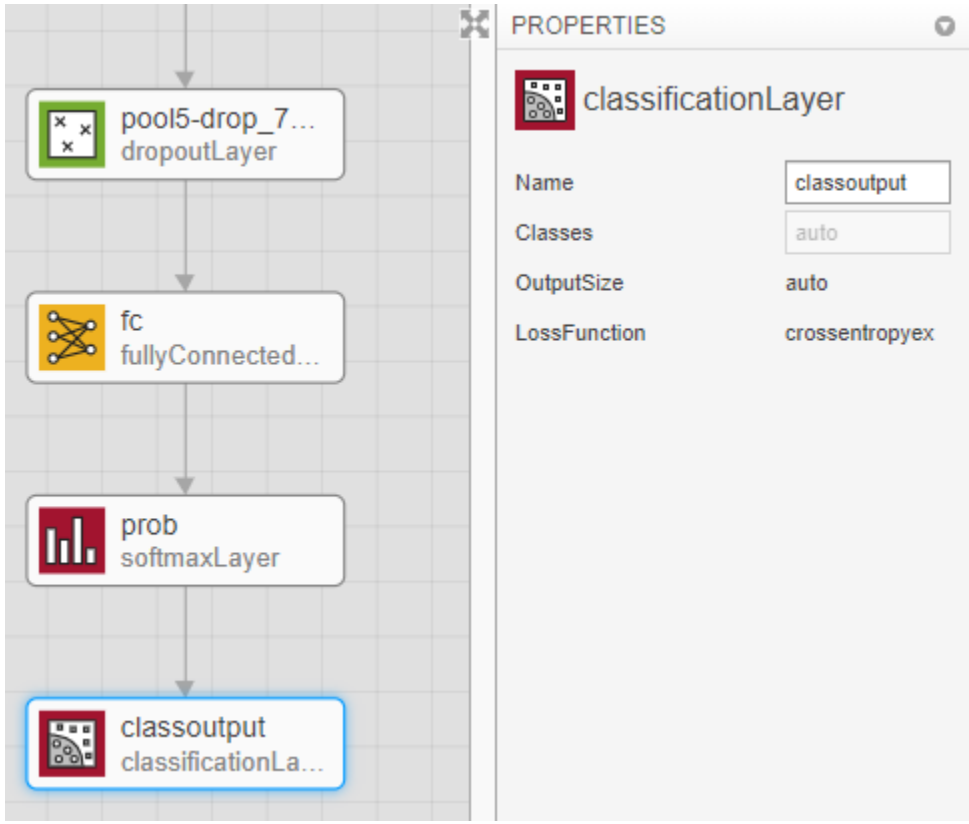
To retrain a pretrained network to classify new images, replace the final layers with new layers adapted to the new data set. You must change the number of classes to match your data.

Drag a new **fullyConnectedLayer** from the **Layer Library** onto the canvas. Edit the `OutputSize` to the number of classes in the new data, in this example, 5.

Edit learning rates to learn faster in the new layers than in the transferred layers. Set `WeightLearnRateFactor` and `BiasLearnRateFactor` to 10. Delete the last fully connected and connect up your new layer instead.

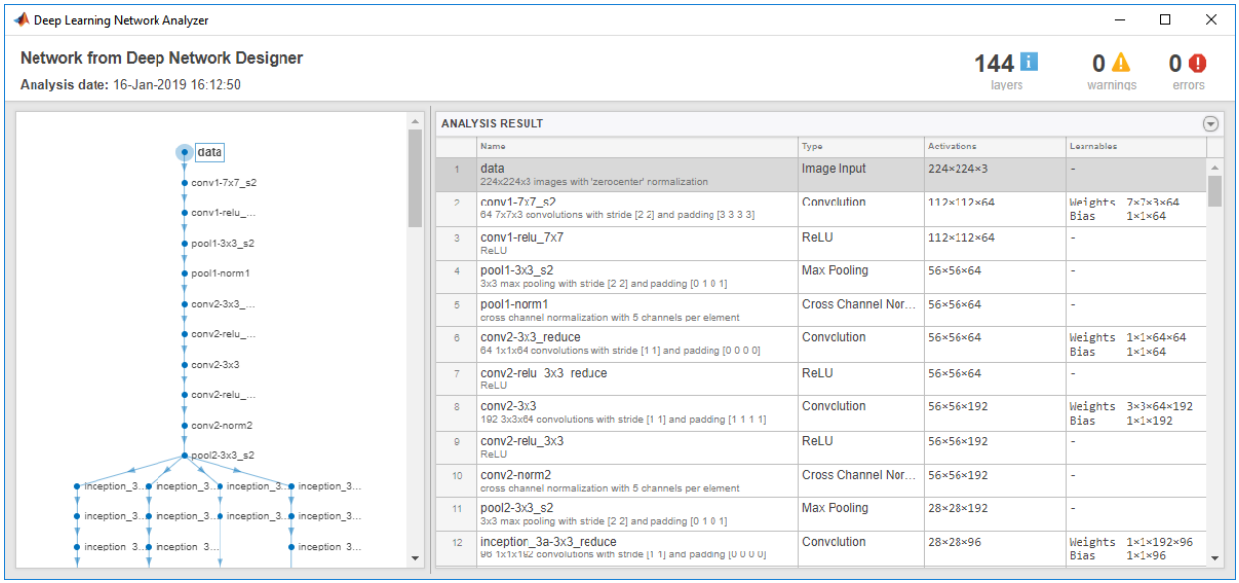


Replace the output layer. Scroll to the end of the **Layer Library** and drag a new **classificationLayer** onto the canvas. Delete the original `output` layer and connect up your new layer instead.



Check Network

To make sure your edited network is ready for training, click **Analyze**, and ensure the Deep Learning Network Analyzer reports zero errors.



Export Network for Training

Return to the Deep Network Designer and click **Export**. Deep Network Designer exports the network to a new variable called `lgraph_1` containing the edited network layers. You can now supply the layer variable to the `trainNetwork` function. You can also generate MATLAB® code that recreates the network architecture and returns it as a `layerGraph` object or a `Layer` array in the MATLAB workspace.

Load Data and Train Network

Unzip and load the new images as an image datastore. Divide the data into 70% training data and 30% validation data.

```
unzip('MerchData.zip');
imds = imageDatastore('MerchData','IncludeSubfolders',true,'LabelSource','foldernames');
[imdsTrain,imdsValidation] = splitEachLabel(imds,0.7,'randomized');
```

Resize images to match the pretrained network input size.

```
augimdsTrain = augmentedImageDatastore([224 224],imdsTrain);
augimdsValidation = augmentedImageDatastore([224 224],imdsValidation);
```

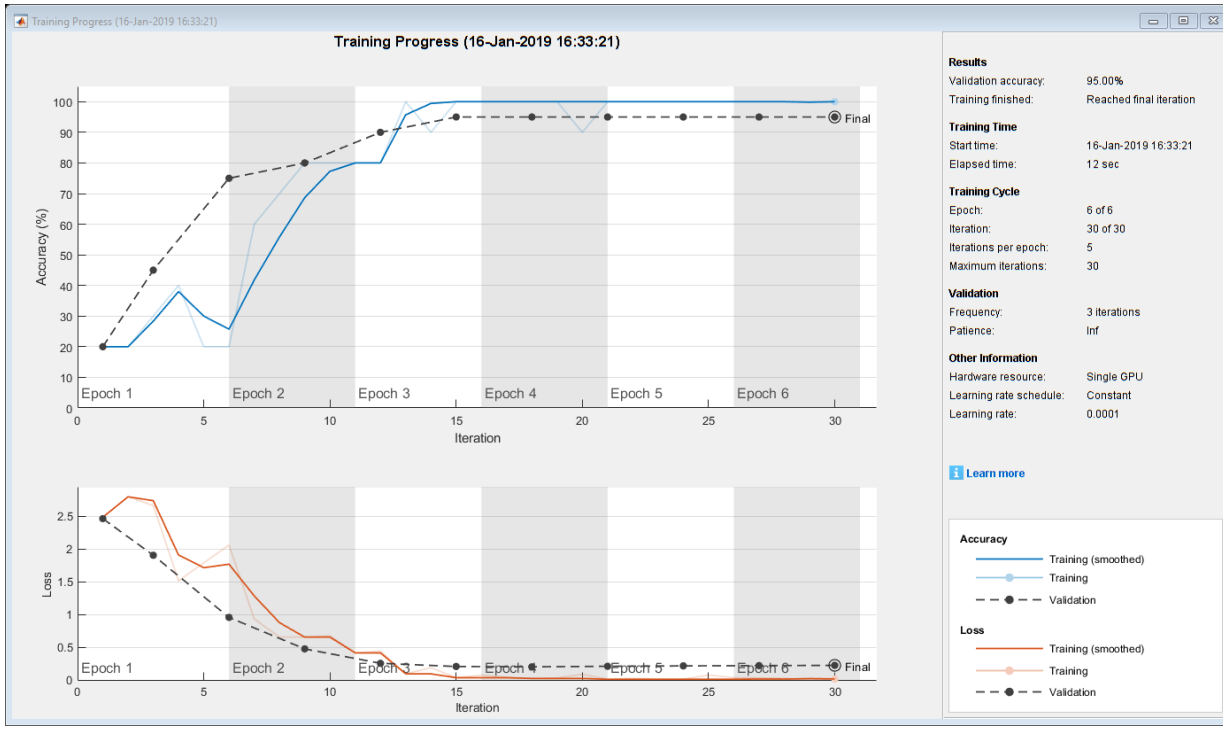
Specify training options.

- Specify the mini-batch size, that is, how many images to use in each iteration.
- Specify a small number of epochs. An epoch is a full training cycle on the entire training data set. For transfer learning, you do not need to train for as many epochs. Shuffle the data every epoch.
- Set `InitialLearnRate` to a small value to slow down learning in the transferred layers.
- Specify validation data and a small validation frequency.
- Turn on the training plot to monitor progress while you train.

```
options = trainingOptions('sgdm', ...
    'MiniBatchSize',10, ...
    'MaxEpochs',6, ...
    'Shuffle','every-epoch', ...
    'InitialLearnRate',1e-4, ...
    'ValidationData',augimdsValidation, ...
    'ValidationFrequency',6, ...
    'Verbose',false, ...
    'Plots','training-progress');
```

To train the network, supply the layers exported from the app, `lgraph_1`, the training images, and options, to the `trainNetwork` function. By default, `trainNetwork` uses a GPU if available (requires Parallel Computing Toolbox™). Otherwise, it uses a CPU. Training is fast because the data set is so small.

```
netTransfer = trainNetwork(augimdsTrain,lgraph_1,options);
```



Test Trained Network

Classify the validation images using the fine-tuned network, and calculate the classification accuracy.

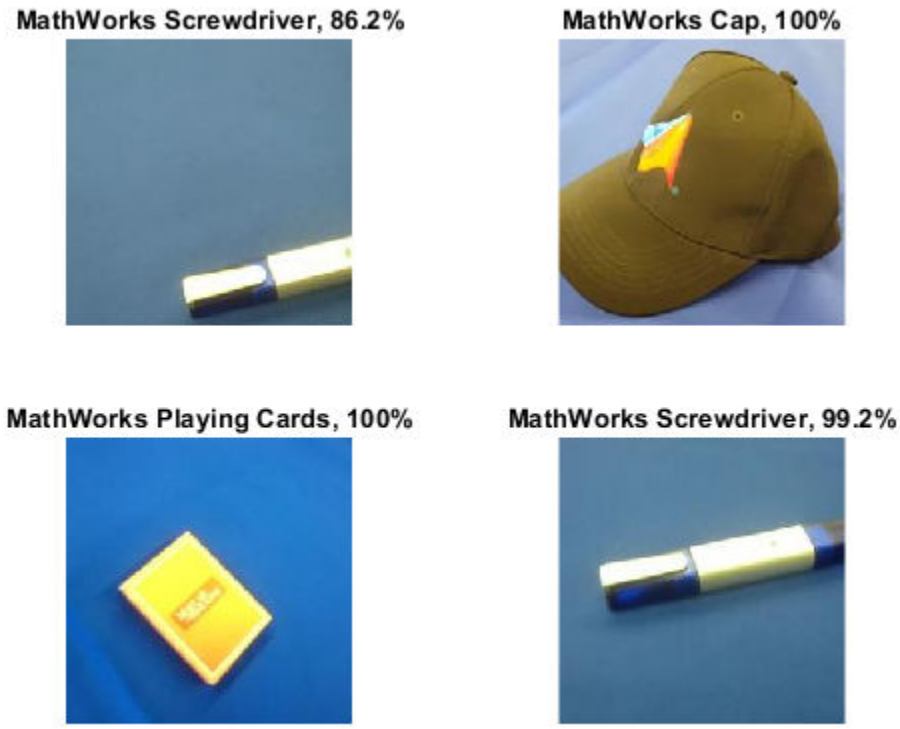
```
[YPred,probs] = classify(netTransfer,augimdsValidation);
accuracy = mean(YPred == imdsValidation.Labels)
```

```
accuracy = 1
```

Display four sample validation images with predicted labels and predicted probabilities.

```
idx = randperm(numel(augimdsValidation.Files),4);
figure
for i = 1:4
    subplot(2,2,i)
    I = readimage(imdsValidation,idx(i));
    imshow(I)
    label = YPred(idx(i));
```

```
    title(string(label) + ", " + num2str(100*max(probs(idx(i),:)),3) + "%");
end
```



To learn more and try other pretrained networks, see Deep Network Designer.

See Also

Deep Network Designer

More About

- “Build Networks with Deep Network Designer”

- “Deep Learning Tips and Tricks”
- “List of Deep Learning Layers”

Try Deep Learning in 10 Lines of MATLAB Code

This example shows how to use deep learning to identify objects on a live webcam using only 10 lines of MATLAB code. Try the example to see how simple it is to get started with deep learning in MATLAB.

1. Run these commands to get the downloads if needed, connect to the webcam, and get a pretrained neural network.

```
camera = webcam; % Connect to the camera
net = alexnet;   % Load the neural network
```

If you need to install the `webcam` and `alexnet` add-ons, a message from each function appears with a link to help you download the free add-ons using Add-On Explorer. Alternatively, see Deep Learning Toolbox Model *for AlexNet Network* and MATLAB Support Package for USB Webcams.

After you install Deep Learning Toolbox Model *for AlexNet Network*, you can use it to classify images. AlexNet is a pretrained convolutional neural network (CNN) that has been trained on more than a million images and can classify images into 1000 object categories (for example, keyboard, mouse, coffee mug, pencil, and many animals).

2. Run the following code to show and classify live images. Point the webcam at an object and the neural network reports what class of object it thinks the webcam is showing. It will keep classifying images until you press **Ctrl+C**. The code resizes the image for the network using `imresize`.

```
while true
    im = snapshot(camera);       % Take a picture
    image(im);                   % Show the picture
    im = imresize(im,[227 227]); % Resize the picture for alexnet
    label = classify(net,im);    % Classify the picture
    title(char(label));          % Show the class label
    drawnow
end
```

In this example, the network correctly classifies a coffee mug. Experiment with objects in your surroundings to see how accurate the network is.

To watch a video of this example, see Deep Learning in 11 Lines of MATLAB Code.

To learn how to extend this example and show the probability scores of classes, see "Classify Webcam Images Using Deep Learning".

For next steps in deep learning, you can use the pretrained network for other tasks. Solve new classification problems on your image data with transfer learning or feature extraction. For examples, see "Start Deep Learning Faster Using Transfer Learning" and "Train Classifiers Using Features Extracted from Pretrained Networks". To try other pretrained networks, see "Pretrained Deep Neural Networks".

See Also

`alexnet` | `trainNetwork` | `trainingOptions`

More About

- “Classify Webcam Images Using Deep Learning”
- “Classify Image Using Pretrained Network” on page 1-15
- “Get Started with Transfer Learning” on page 1-18
- “Transfer Learning with Deep Network Designer”
- “Create Simple Image Classification Network” on page 1-22
- “Create Simple Sequence Classification Network” on page 1-26

Classify Image Using Pretrained Network

This example shows how to classify an image using the pretrained deep convolutional neural network GoogLeNet.

GoogLeNet has been trained on over a million images and can classify images into 1000 object categories (such as keyboard, coffee mug, pencil, and many animals). The network has learned rich feature representations for a wide range of images. The network takes an image as input, and then outputs a label for the object in the image together with the probabilities for each of the object categories.

Load Pretrained Network

Load the pretrained GoogLeNet network. You can also choose to load a different pretrained network for image classification. This step requires the Deep Learning Toolbox™ Model *for GoogLeNet Network* support package. If you do not have the required support packages installed, then the software provides a download link.

```
net = googlenet;
```

Read and Resize Image

The image that you want to classify must have the same size as the input size of the network. For GoogLeNet, the network input size is the `InputSize` property of the image input layer.

Read the image that you want to classify and resize it to the input size of the network. This resizing slightly changes the aspect ratio of the image.

```
I = imread("peppers.png");
inputSize = net.Layers(1).InputSize;
I = imresize(I,inputSize(1:2));
```

Classify and Display Image

Classify and display the image with the predicted label.

```
label = classify(net,I);
figure
imshow(I)
title(string(label))
```

For a more detailed example showing how to also display the top predictions with their associated probabilities, see “Classify Image Using GoogLeNet”.

For next steps in deep learning, you can use the pretrained network for other tasks. Solve new classification problems on your image data with transfer learning or feature extraction. For examples, see “Start Deep Learning Faster Using Transfer Learning” and “Train Classifiers Using Features Extracted from Pretrained Networks”. To try other pretrained networks, see “Pretrained Deep Neural Networks”.

References

1 Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015.

2 *BVLC GoogLeNet Model*. https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet

See Also

`classify` | `googlenet`

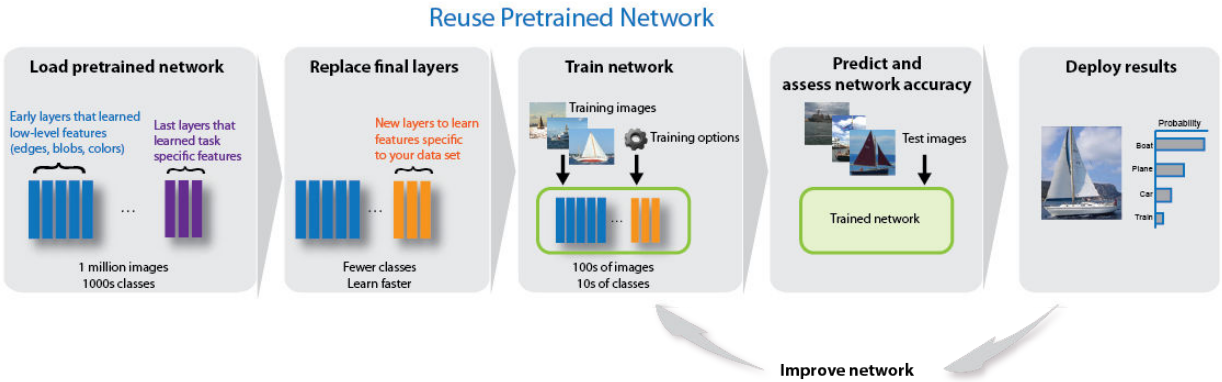
More About

- “Classify Image Using GoogLeNet”
- “Try Deep Learning in 10 Lines of MATLAB Code” on page 1-12
- “Get Started with Transfer Learning” on page 1-18
- “Transfer Learning with Deep Network Designer”
- “Create Simple Image Classification Network” on page 1-22
- “Create Simple Sequence Classification Network” on page 1-26

Get Started with Transfer Learning

This example shows how to use transfer learning to retrain ResNet-18, a pretrained convolutional neural network, to classify a new set of images. Try this example to see how simple it is to get started with deep learning in MATLAB®.

Transfer learning is commonly used in deep learning applications. You can take a pretrained network and use it as a starting point to learn a new task. Fine-tuning a network with transfer learning is usually much faster and easier than training a network with randomly initialized weights from scratch. You can quickly transfer learned features to a new task using a smaller number of training images.



Load Data

Unzip and load the new images as an image datastore. Divide the data into training and validation data sets. Use 70% of the images for training and 30% for validation.

```
unzip('MerchData.zip');
imds = imageDatastore('MerchData','IncludeSubfolders',true,'LabelSource','foldernames')
[imdsTrain,imdsValidation] = splitEachLabel(imds,0.7,'randomized');
```

Load Pretrained Network

Load the pretrained ResNet-18 network. If Deep Learning Toolbox™ Model *for ResNet-18 Network* is not installed, then the software provides a download link. ResNet-18 has been trained on over a million images and can classify images into 1000 object categories (such as keyboard, coffee mug, pencil, and many animals). The network has learned rich feature representations for a wide range of images. The network takes an image as input and

outputs a label for the object in the image together with the probabilities for each of the object categories. To perform transfer learning using different pretrained networks, see "Train Deep Learning Network to Classify New Images".

```
net = resnet18;
```

Replace Final Layers

To retrain ResNet-18 to classify new images, replace the last fully connected layer and the final classification layer of the network. In ResNet-18, these layers have the names `'fc1000'` and `'ClassificationLayer_predictions'`, repectively. Set the new fully connected layer to have the same size as the number of classes in the new data set (5, in this example). To learn faster in the new layers than in the transferred layers, increase the learning rate factors of the fully connected layer.

```
numClasses = numel(categories(imdsTrain.Labels));
lgraph = layerGraph(net);

newFCLayer = fullyConnectedLayer(numClasses,'Name','new_fc','WeightLearnRateFactor',10,
lgraph = replaceLayer(lgraph,'fc1000',newFCLayer);

newClassLayer = classificationLayer('Name','new_classoutput');
lgraph = replaceLayer(lgraph,'ClassificationLayer_predictions',newClassLayer);
```

Train Network

The network requires input images of size 224-by-224-by-3, but the images in the image datastore have different sizes. Use an augmented image datastore to automatically resize the training images. You can also use an `imageDataAugmenter` to specify additional augmentation operations to perform on the training images to help prevent the network from overfitting.

```
inputSize = net.Layers(1).InputSize;
augimdsTrain = augmentedImageDatastore(inputSize(1:2),imdsTrain);
augimdsValidation = augmentedImageDatastore(inputSize(1:2),imdsValidation);
```

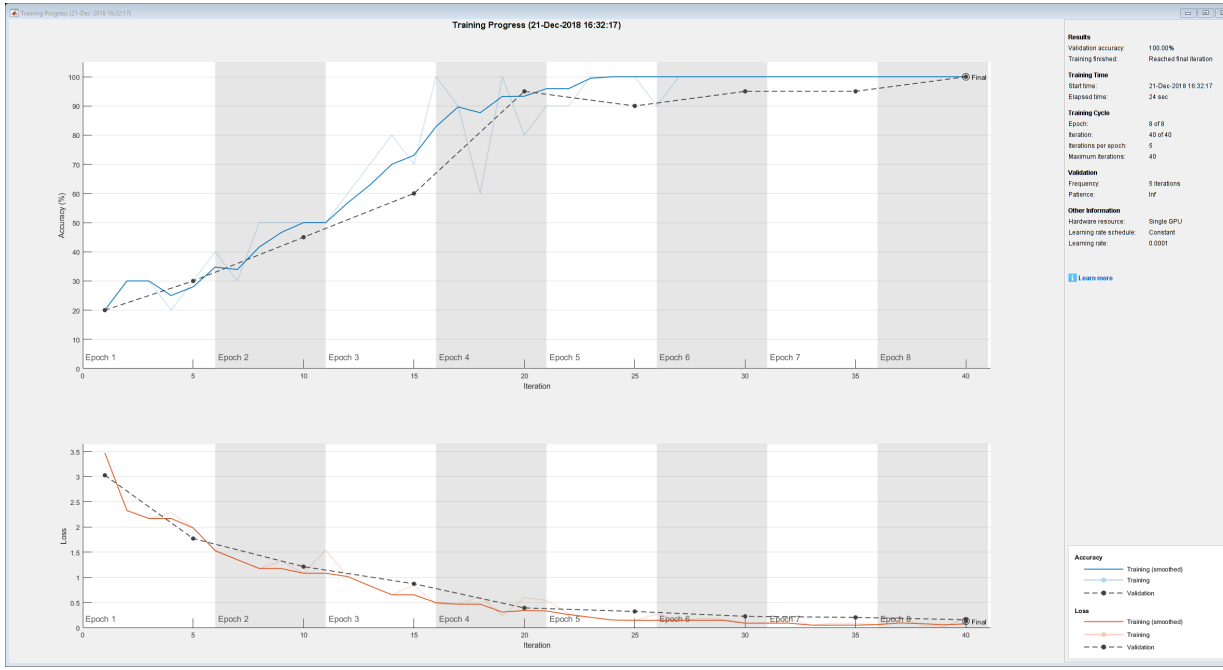
Specify the training options, including mini-batch size and validation data. Set `InitialLearnRate` to a small value to slow down learning in the transferred layers. In the previous step, you increased the learning rate factors for the fully connected layer to speed up learning in the new final layers. This combination of learning rate settings results in fast learning only in the new layers and slower learning in the other layers.

```
options = trainingOptions('sgdm', ...
    'MiniBatchSize',10, ...
```

```
    'MaxEpochs',8, ...
    'InitialLearnRate',1e-4, ...
    'Shuffle','every-epoch', ...
    'ValidationData',augimdsValidation, ...
    'ValidationFrequency',5, ...
    'Verbose',false, ...
    'Plots','training-progress');
```

Train the network using the training data. By default, `trainNetwork` uses a GPU if one is available (requires Parallel Computing Toolbox™ and a CUDA® enabled GPU with compute capability 3.0 or higher). Otherwise, it uses a CPU.

```
trainedNet = trainNetwork(augimdsTrain,lgraph,options);
```



Classify Validation Images

Classify the validation images using the fine-tuned network, and calculate the classification accuracy.

```
YPred = classify(trainedNet,augimdsValidation);
accuracy = mean(YPred == imdsValidation.Labels)
```

```
accuracy = 1
```

Learn More

To perform transfer learning using different pretrained networks, see “Train Deep Learning Network to Classify New Images”. To try transfer learning using the Deep Network Designer app, see “Transfer Learning with Deep Network Designer”. For a list and comparison of the pretrained networks, see “Pretrained Deep Neural Networks”.

References

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in neural information processing systems*. 2012.

[2] *BVLC AlexNet Model*. https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet

See Also

`alexnet` | `trainNetwork` | `trainingOptions`

More About

- “Transfer Learning Using AlexNet”
- “Try Deep Learning in 10 Lines of MATLAB Code” on page 1-12
- “Classify Image Using Pretrained Network” on page 1-15
- “Transfer Learning with Deep Network Designer”
- “Create Simple Image Classification Network” on page 1-22
- “Create Simple Sequence Classification Network” on page 1-26

Create Simple Image Classification Network

This example shows how to create and train a simple convolutional neural network for deep learning classification. Convolutional neural networks are essential tools for deep learning and are especially suited for image recognition.

The example demonstrates how to:

- Load image data.
- Define the network architecture.
- Specify training options.
- Train the network.
- Predict the labels of new data and calculate the classification accuracy.

Load Data

Load the digit sample data as an image datastore. The `imageDatastore` function automatically labels the images based on folder names.

```
digitDatasetPath = fullfile(matlabroot,'toolbox','nnet','nndemos', ...
    'nndatasets','DigitDataset');

imds = imageDatastore(digitDatasetPath, ...
    'IncludeSubfolders',true, ...
    'LabelSource','foldernames');
```

Divide the data into training and validation data sets, so that each category in the training set contains 750 images, and the validation set contains the remaining images from each label. `splitEachLabel` splits the image datastore into two new datastores for training and validation.

```
numTrainFiles = 750;
[imdsTrain,imdsValidation] = splitEachLabel(imds,numTrainFiles,'randomize');
```

Define Network Architecture

Define the convolutional neural network architecture. Specify the size of the images in the input layer of the network and the number of classes in the fully connected layer before the classification layer. Each image is 28-by-28-by-1 pixels and there are 10 classes.

```
inputSize = [28 28 1];
numClasses = 10;

layers = [
    imageInputLayer(inputSize)
    convolution2dLayer(5,20)
    batchNormalizationLayer
    reluLayer
    fullyConnectedLayer(numClasses)
    softmaxLayer
    classificationLayer];
```

For more information about deep learning layers, see “List of Deep Learning Layers”.

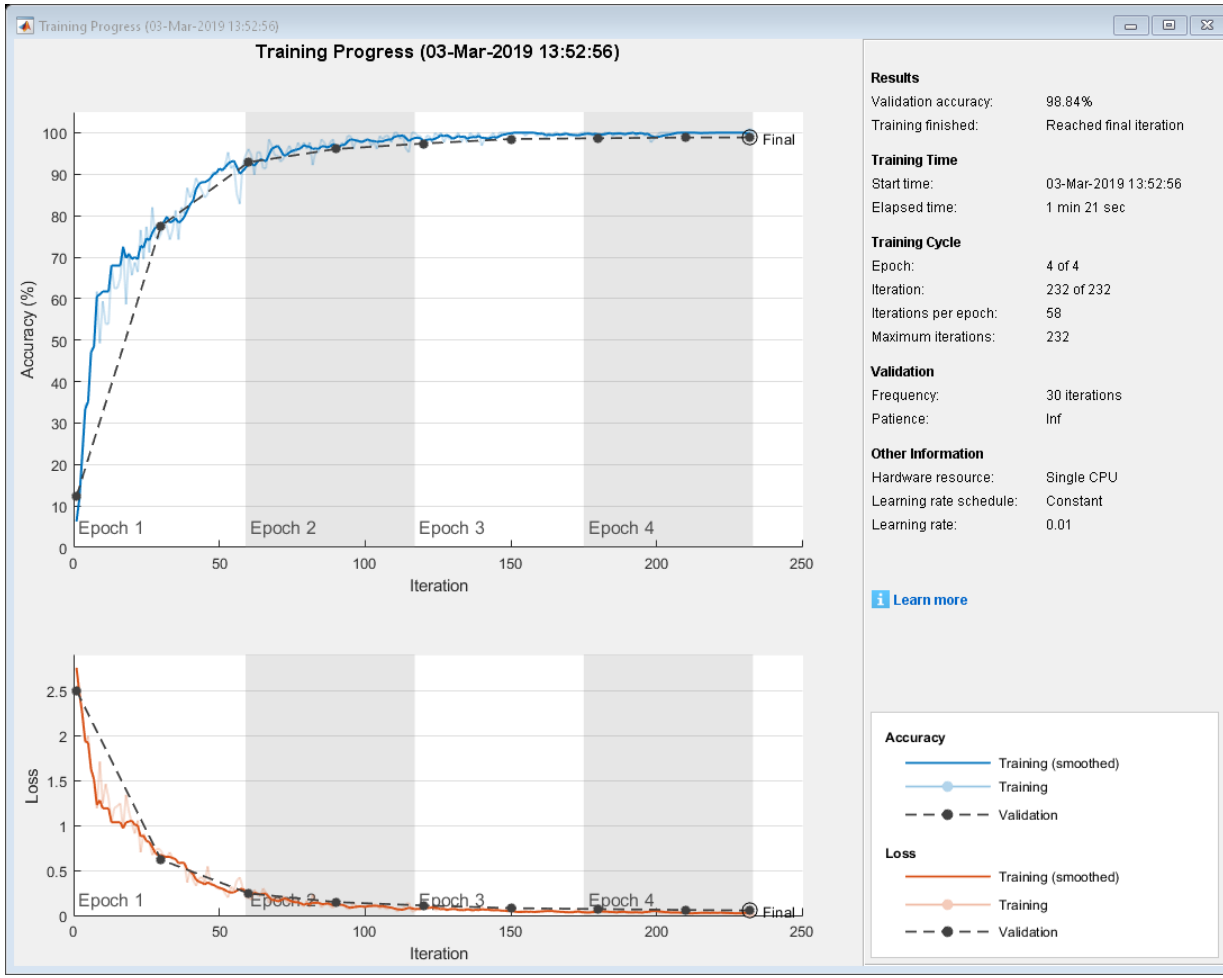
Train Network

Specify the training options and train the network.

By default, `trainNetwork` uses a GPU if one is available (requires Parallel Computing Toolbox™ and a CUDA® enabled GPU with compute capability 3.0 or higher). Otherwise, it uses a CPU. You can also specify the execution environment by using the `'ExecutionEnvironment'` name-value pair argument of `trainingOptions`.

```
options = trainingOptions('sgdm', ...
    'MaxEpochs',4, ...
    'ValidationData',imdsValidation, ...
    'ValidationFrequency',30, ...
    'Verbose',false, ...
    'Plots','training-progress');

net = trainNetwork(imdsTrain,layers,options);
```



For more information about training options, see “Set Up Parameters and Train Convolutional Neural Network”.

Test Network

Classify the validation data and calculate the classification accuracy.

```
YPred = classify(net,imdsValidation);
YValidation = imdsValidation.Labels;
accuracy = mean(YPred == YValidation)
```

```
accuracy = 0.9884
```

For next steps in deep learning, you can try using pretrained network for other tasks. Solve new classification problems on your image data with transfer learning or feature extraction. For examples, see “Start Deep Learning Faster Using Transfer Learning” and “Train Classifiers Using Features Extracted from Pretrained Networks”. To learn more about pretrained networks, see “Pretrained Deep Neural Networks”.

See Also

`trainNetwork` | `trainingOptions`

More About

- “Start Deep Learning Faster Using Transfer Learning”
- “Try Deep Learning in 10 Lines of MATLAB Code” on page 1-12
- “Classify Image Using Pretrained Network” on page 1-15
- “Get Started with Transfer Learning” on page 1-18
- “Transfer Learning with Deep Network Designer”
- “Create Simple Sequence Classification Network” on page 1-26

Create Simple Sequence Classification Network

This example shows how to create a simple long short-term memory (LSTM) classification network.

To train a deep neural network to classify sequence data, you can use an LSTM network. An LSTM network is a type of recurrent neural network (RNN) that learns long-term dependencies between time steps of sequence data.

The example demonstrates how to:

- Load sequence data.
- Define the network architecture.
- Specify training options.
- Train the network.
- Predict the labels of new data and calculate the classification accuracy.

Load Data

Load the Japanese Vowels data set as described in [1] and [2]. The predictors are a cell arrays containing sequences of varying length with a feature dimension of 12. The labels are categorical vectors of labels 1,2,...,9.

```
[XTrain,YTrain] = japaneseVowelsTrainData;
[XValidation,YValidation] = japaneseVowelsTestData;
```

View the sizes of the first few training sequences. The sequences are matrices with 12 rows (one row for each feature) and a varying number of columns (one column for each time step).

```
XTrain(1:5)
```

```
ans = 5x1 cell array
    {12x20 double}
    {12x26 double}
    {12x22 double}
    {12x20 double}
    {12x21 double}
```

Define Network Architecture

Define the LSTM network architecture. Specify the number of features in the input layer and the number of classes in the fully connected layer.

```
numFeatures = 12;
numHiddenUnits = 100;
numClasses = 9;

layers = [ ...
    sequenceInputLayer(numFeatures)
    lstmLayer(numHiddenUnits,'OutputMode','last')
    fullyConnectedLayer(numClasses)
    softmaxLayer
    classificationLayer];
```

Train Network

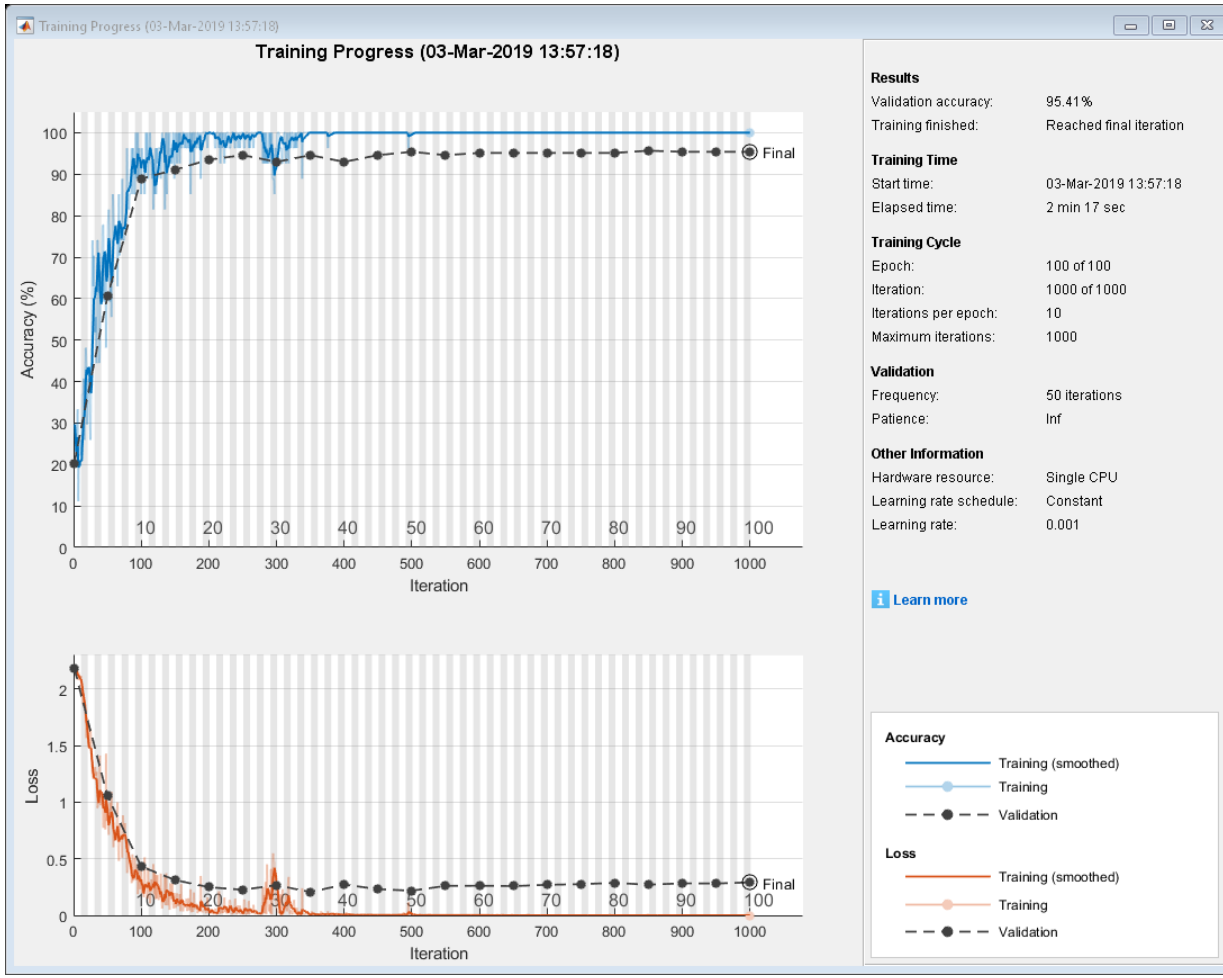
Specify the training options and train the network.

Because the mini-batches are small with short sequences, the CPU is better suited for training. Set `'ExecutionEnvironment'` to `'cpu'`. To train on a GPU, if available, set `'ExecutionEnvironment'` to `'auto'` (the default value).

```
miniBatchSize = 27;

options = trainingOptions('adam', ...
    'ExecutionEnvironment','cpu', ...
    'MaxEpochs',100, ...
    'MiniBatchSize',miniBatchSize, ...
    'ValidationData',{XValidation,YValidation}, ...
    'GradientThreshold',2, ...
    'Shuffle','every-epoch', ...
    'Verbose',false, ...
    'Plots','training-progress');

net = trainNetwork(XTrain,YTrain,layers,options);
```



For more information about specifying training options, see “Set Up Parameters and Train Convolutional Neural Network”.

Test Network

Classify the test data and calculate the classification accuracy. Specify the same mini-batch size used for training.

```
YPred = classify(net,XValidation,'MiniBatchSize',miniBatchSize);
acc = mean(YPred == YValidation)
```

```
acc = 0.9541
```

For next steps, you can try improving the accuracy by using bidirectional LSTM (BiLSTM) layers or by creating a deeper network. For more information, see “Long Short-Term Memory Networks”.

For an example showing how to use convolutional networks to classify sequence data, see “Speech Command Recognition Using Deep Learning”.

References

1. M. Kudo, J. Toyama, and M. Shimbo. "Multidimensional Curve Classification Using Passing-Through Regions." *Pattern Recognition Letters*. Vol. 20, No. 11–13, pages 1103–1111.
2. *UCI Machine Learning Repository: Japanese Vowels Dataset*. https://archive.ics.uci.edu/ml/datasets/Japanese+Vowels

See Also

`lstmLayer` | `trainNetwork` | `trainingOptions`

More About

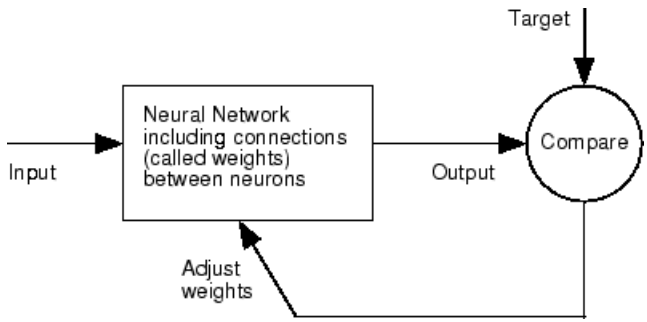
- “Long Short-Term Memory Networks”
- “Try Deep Learning in 10 Lines of MATLAB Code” on page 1-12
- “Classify Image Using Pretrained Network” on page 1-15
- “Get Started with Transfer Learning” on page 1-18
- “Transfer Learning with Deep Network Designer”
- “Create Simple Image Classification Network” on page 1-22

Shallow Networks for Pattern Recognition, Clustering and Time Series

In this section...
“Shallow Network Apps and Functions in Deep Learning Toolbox” on page 1-31
“Deep Learning Toolbox Applications” on page 1-32
“Shallow Neural Network Design Steps” on page 1-34

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. You can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements.

Typically, neural networks are adjusted, or trained, so that a particular input leads to a specific target output. The next figure illustrates such a situation. Here, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target. Typically, many such input/target pairs are needed to train a network.



Neural networks have been trained to perform complex functions in various fields, including pattern recognition, identification, classification, speech, vision, and control systems.

Neural networks can also be trained to solve problems that are difficult for conventional computers or human beings. The toolbox emphasizes the use of neural network paradigms that build up to—or are themselves used in— engineering, financial, and other practical applications.

The following topics explain how to use graphical tools for training neural networks to solve problems in function fitting, pattern recognition, clustering, and time series. Using these tools can give you an excellent introduction to the use of the Deep Learning Toolbox software:

- “Fit Data with a Shallow Neural Network” on page 1-35
- “Classify Patterns with a Shallow Neural Network” on page 1-61
- “Cluster Data with a Self-Organizing Map” on page 1-86
- “Shallow Neural Network Time-Series Prediction and Modeling” on page 1-106

Shallow Network Apps and Functions in Deep Learning Toolbox

There are four ways you can use the Deep Learning Toolbox software.

- The first way is through its tools. You can open any of these tools from a master tool started by the command `nnstart`. These tools provide a convenient way to access the capabilities of the toolbox for the following tasks:
- • Function fitting (`nftool`)
  - Pattern recognition (`nprtool`)
  - Data clustering (`nctool`)
  - Time-series analysis (`ntstool`)
- The second way to use the toolbox is through basic command-line operations. The command-line operations offer more flexibility than the tools, but with some added complexity. If this is your first experience with the toolbox, the tools provide the best introduction. In addition, the tools can generate scripts of documented MATLAB code to provide you with templates for creating your own customized command-line functions. The process of using the tools first, and then generating and modifying MATLAB scripts, is an excellent way to learn about the functionality of the toolbox.
- The third way to use the toolbox is through customization. This advanced capability allows you to create your own custom neural networks, while still having access to the full functionality of the toolbox. You can create networks with arbitrary connections, and you still be able to train them using existing toolbox training functions (as long as the network components are differentiable).
- The fourth way to use the toolbox is through the ability to modify any of the functions contained in the toolbox. Every computational component is written in MATLAB code and is fully accessible.

These four levels of toolbox usage span the novice to the expert: simple tools guide the new user through specific applications, and network customization allows researchers to try novel architectures with minimal effort. Whatever your level of neural network and MATLAB knowledge, there are toolbox features to suit your needs.

Automatic Script Generation

The tools themselves form an important part of the learning process for the Deep Learning Toolbox software. They guide you through the process of designing neural networks to solve problems in four important application areas, without requiring any background in neural networks or sophistication in using MATLAB. In addition, the tools can automatically generate both simple and advanced MATLAB scripts that can reproduce the steps performed by the tool, but with the option to override default settings. These scripts can provide you with templates for creating customized code, and they can aid you in becoming familiar with the command-line functionality of the toolbox. It is highly recommended that you use the automatic script generation facility of these tools.

Deep Learning Toolbox Applications

It would be impossible to cover the total range of applications for which neural networks have provided outstanding solutions. The remaining sections of this topic describe only a few of the applications in function fitting, pattern recognition, clustering, and time series analysis. The following table provides an idea of the diversity of applications for which neural networks provide state-of-the-art solutions.

Industry	Business Applications
Aerospace	High-performance aircraft autopilot, flight path simulation, aircraft control systems, autopilot enhancements, aircraft component simulation, and aircraft component fault detection
Automotive	Automobile automatic guidance system, and warranty activity analysis
Banking	Check and other document reading and credit application evaluation
Defense	Weapon steering, target tracking, object discrimination, facial recognition, new kinds of sensors, sonar, radar and image signal processing including data compression, feature extraction and noise suppression, and signal/image identification

Industry	Business Applications
Electronics	Code sequence prediction, integrated circuit chip layout, process control, chip failure analysis, machine vision, voice synthesis, and nonlinear modeling
Entertainment	Animation, special effects, and market forecasting
Financial	Real estate appraisal, loan advising, mortgage screening, corporate bond rating, credit-line use analysis, credit card activity tracking, portfolio trading program, corporate financial analysis, and currency price prediction
Industrial	Prediction of industrial processes, such as the output gases of furnaces, replacing complex and costly equipment used for this purpose in the past
Insurance	Policy application evaluation and product optimization
Manufacturing	Manufacturing process control, product design and analysis, process and machine diagnosis, real-time particle identification, visual quality inspection systems, beer testing, welding quality analysis, paper quality prediction, computer-chip quality analysis, analysis of grinding operations, chemical product design analysis, machine maintenance analysis, project bidding, planning and management, and dynamic modeling of chemical process system
Medical	Breast cancer cell analysis, EEG and ECG analysis, prosthesis design, optimization of transplant times, hospital expense reduction, hospital quality improvement, and emergency-room test advisement
Oil and gas	Exploration
Robotics	Trajectory control, forklift robot, manipulator controllers, and vision systems
Securities	Market analysis, automatic bond rating, and stock trading advisory systems
Speech	Speech recognition, speech compression, vowel classification, and text-to-speech synthesis
Telecommunications	Image and data compression, automated information services, real-time translation of spoken language, and customer payment processing systems

Industry	Business Applications
Transportation	Truck brake diagnosis systems, vehicle scheduling, and routing systems

Shallow Neural Network Design Steps

In the remaining sections of this topic, you will follow the standard steps for designing neural networks to solve problems in four application areas: function fitting, pattern recognition, clustering, and time series analysis. The work flow for any of these problems has seven primary steps. (Data collection in step 1, while important, generally occurs outside the MATLAB environment.)

1. Collect data
2. Create the network
3. Configure the network
4. Initialize the weights and biases
5. Train the network
6. Validate the network
7. Use the network

You will follow these steps using both the GUI tools and command-line operations in the following sections:

- “Fit Data with a Shallow Neural Network” on page 1-35
- “Classify Patterns with a Shallow Neural Network” on page 1-61
- “Cluster Data with a Self-Organizing Map” on page 1-86
- “Shallow Neural Network Time-Series Prediction and Modeling” on page 1-106

Fit Data with a Shallow Neural Network

Neural networks are good at fitting functions. In fact, there is proof that a fairly simple neural network can fit any practical function.

Suppose, for instance, that you have data from a health clinic. You want to design a network that can predict the percentage of body fat of a person, given 13 anatomical measurements. You have a total of 252 example people for which you have those 13 items of data and their associated percentages of body fat.

You can solve this problem in two ways:

- Use a graphical user interface, `nftool`, as described in “Using the Neural Network Fitting App” on page 1-36.
- Use command-line functions, as described in “Using Command-Line Functions” on page 1-51.

It is generally best to start with the GUI, and then to use the GUI to automatically generate command-line scripts. Before using either method, first define the problem by selecting a data set. Each GUI has access to many sample data sets that you can use to experiment with the toolbox (see “Sample Data Sets for Shallow Neural Networks” on page 1-137). If you have a specific problem that you want to solve, you can load your own data into the workspace. The next section describes the data format.

Defining a Problem

To define a fitting problem for the toolbox, arrange a set of Q input vectors as columns in a matrix. Then, arrange another set of Q target vectors (the correct output vectors for each of the input vectors) into a second matrix (see “Data Structures” for a detailed description of data formatting for static and time series data). For example, you can define the fitting problem for a Boolean AND gate with four sets of two-element input vectors and one-element targets as follows:

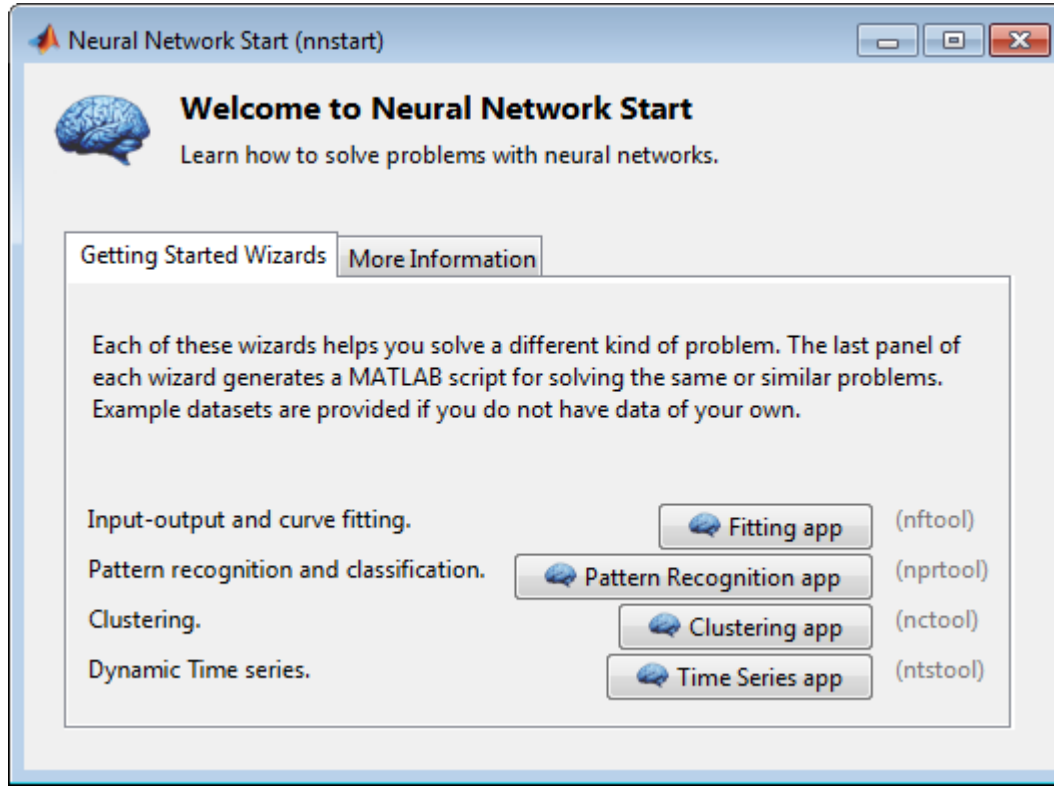
```
inputs = [0 1 0 1; 0 0 1 1];
targets = [0 0 0 1];
```

The next section shows how to train a network to fit a data set, using the neural network fitting app, `nftool`. This example uses the body fat data set provided with the toolbox.

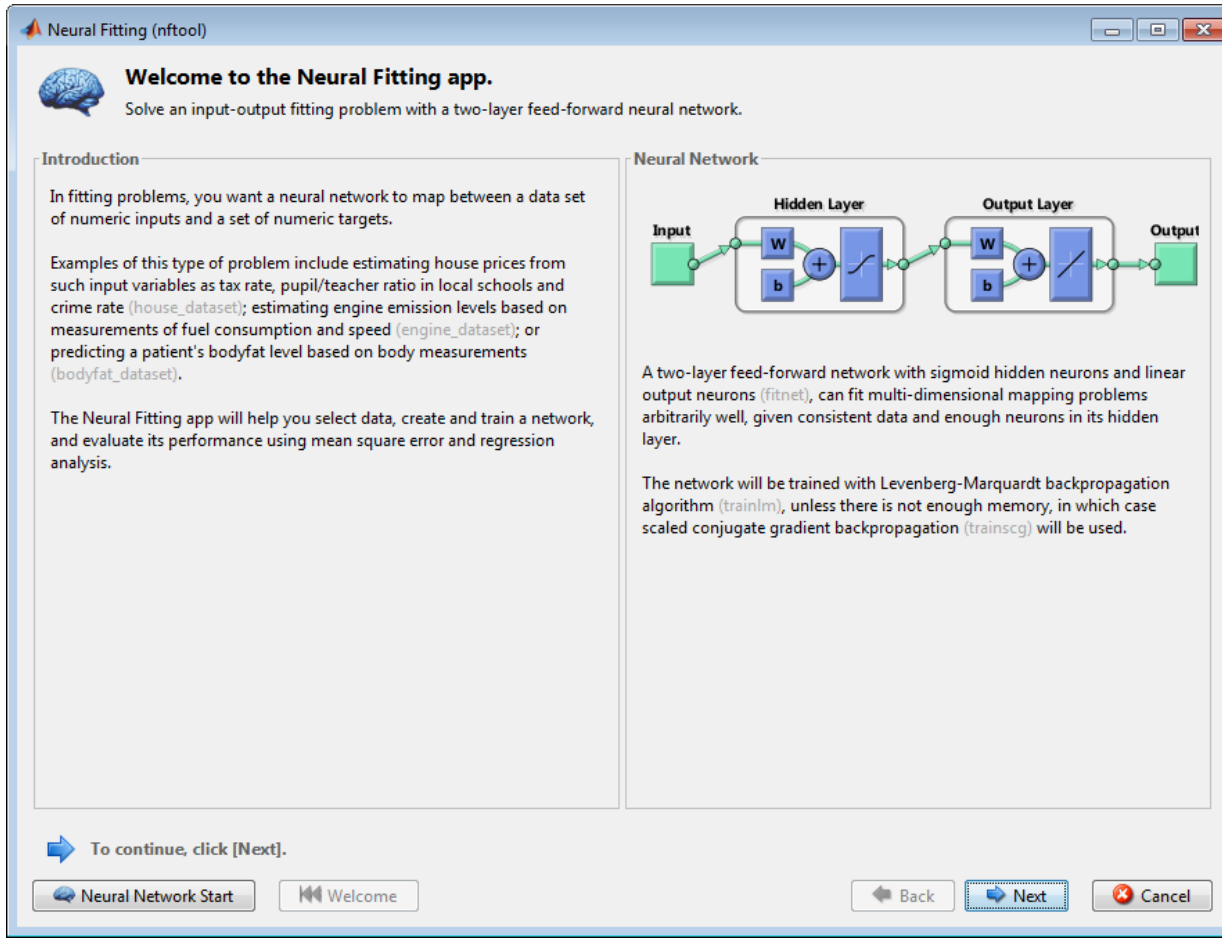
Using the Neural Network Fitting App

1 Open the Neural Network Start GUI with this command:

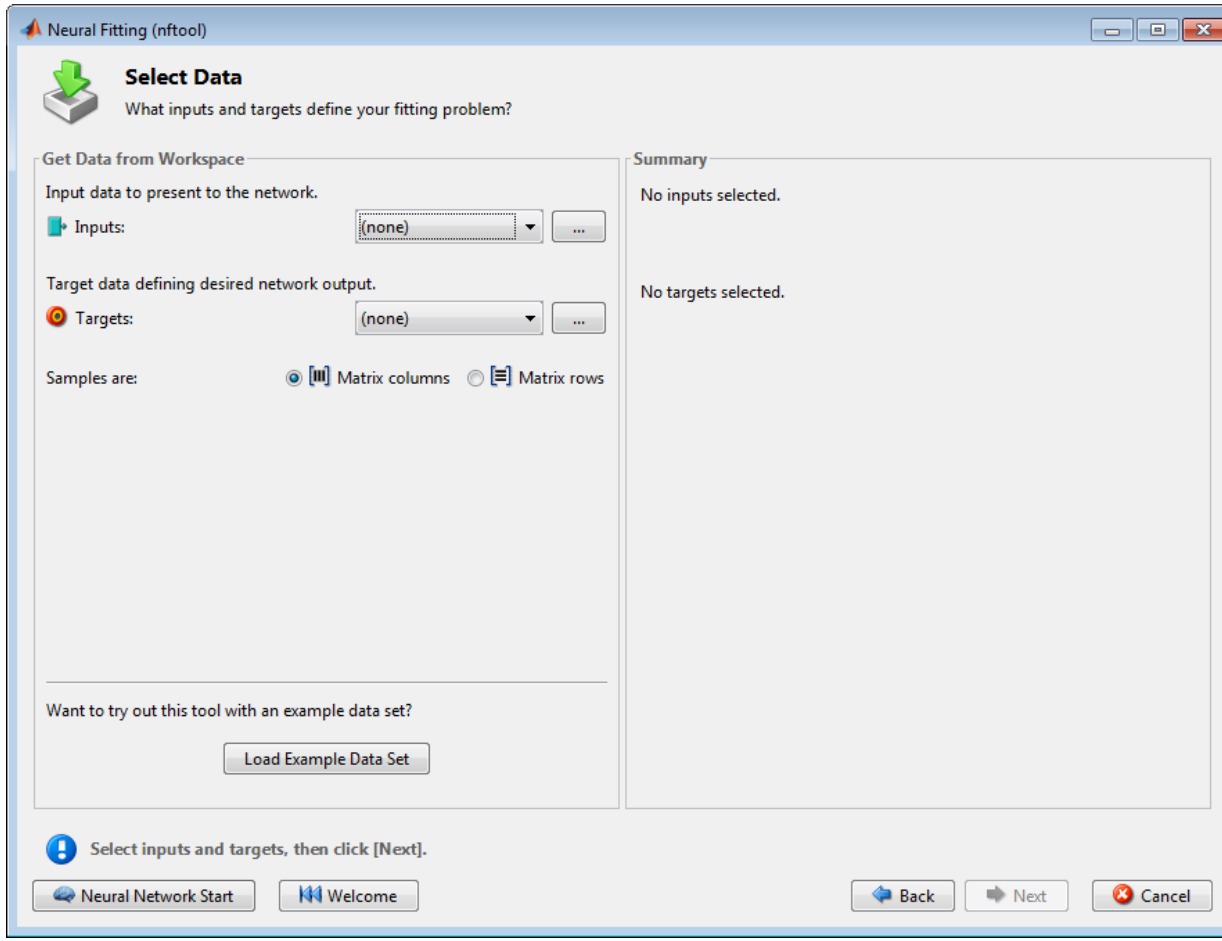
```
nnstart
```



2 Click **Fitting app** to open the Neural Network Fitting App. (You can also use the command `nftool`.)

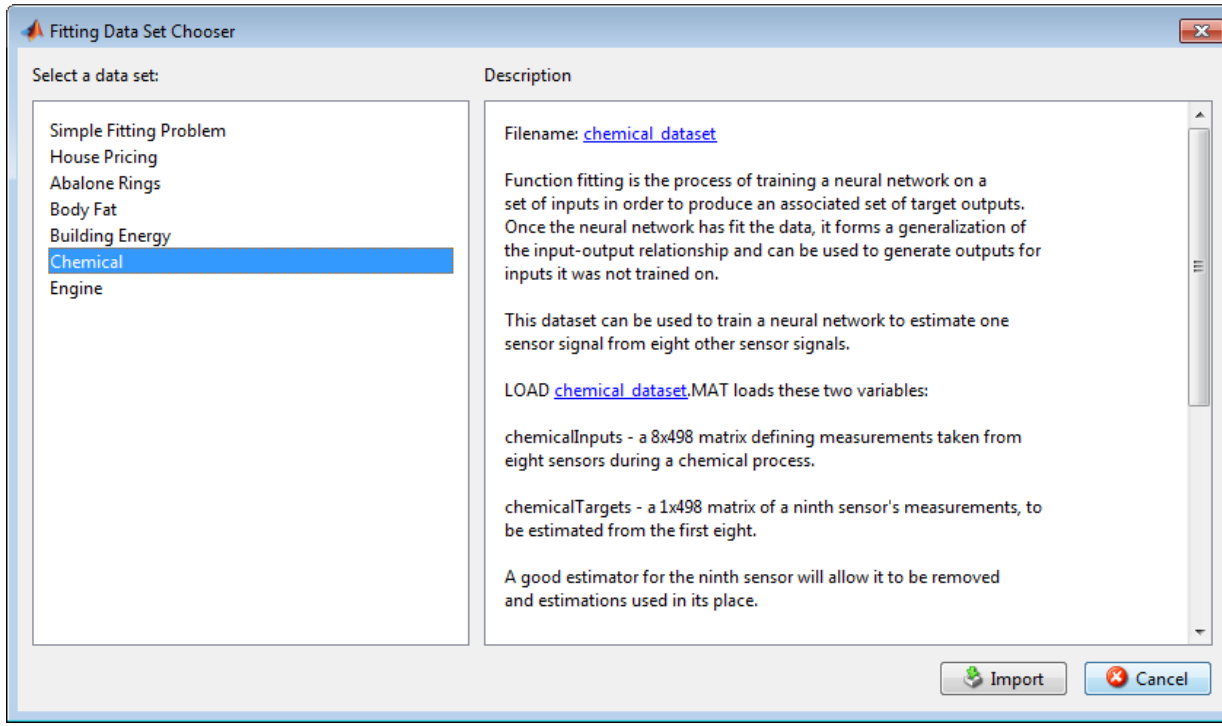


3 Click **Next** to proceed.



4 Click **Load Example Data Set** in the Select Data window. The Fitting Data Set Chooser window opens.

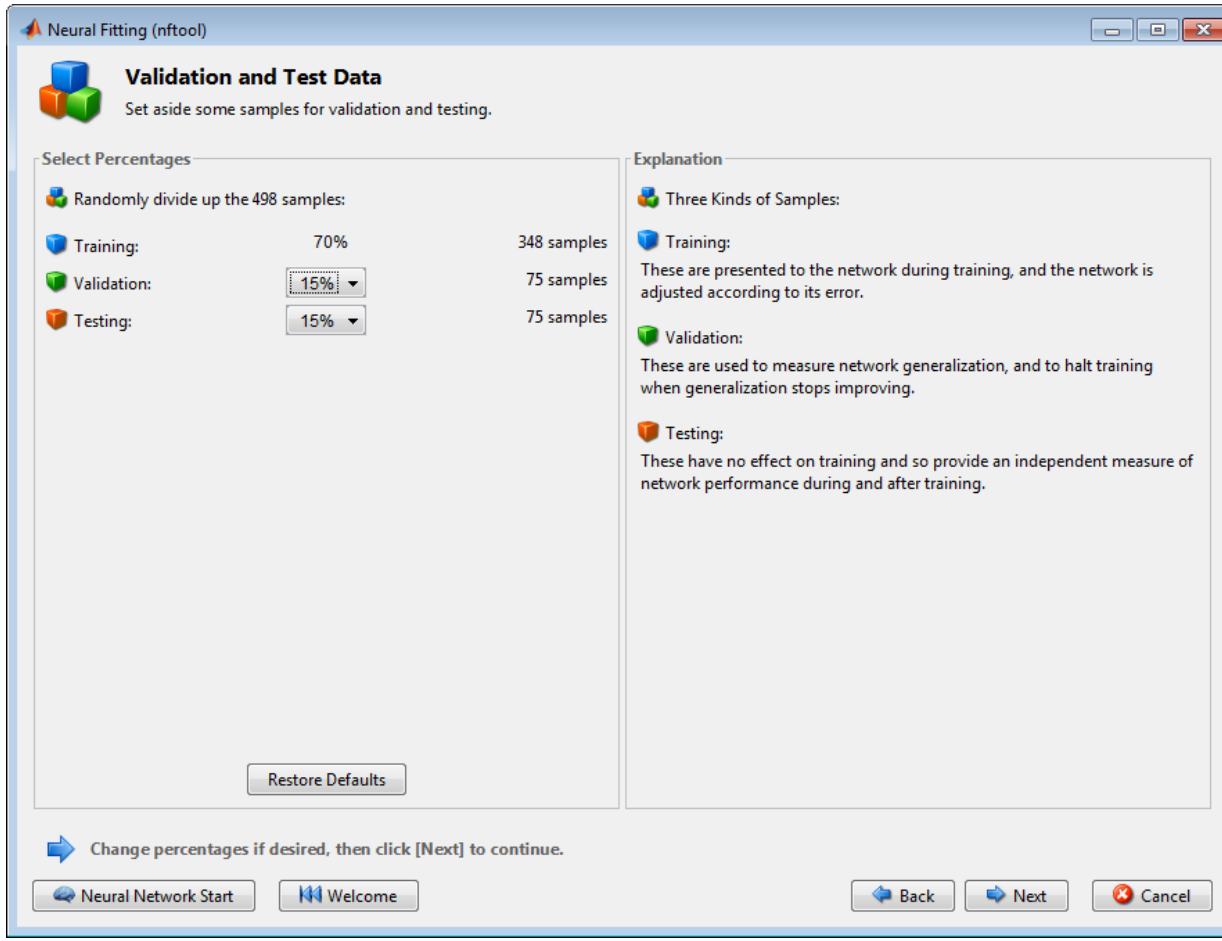
Note Use the **Inputs** and **Targets** options in the Select Data window when you need to load data from the MATLAB workspace.



5 Select **Chemical**, and click **Import**. This returns you to the Select Data window.

6 Click **Next** to display the Validation and Test Data window, shown in the following figure.

The validation and test data sets are each set to 15% of the original data.



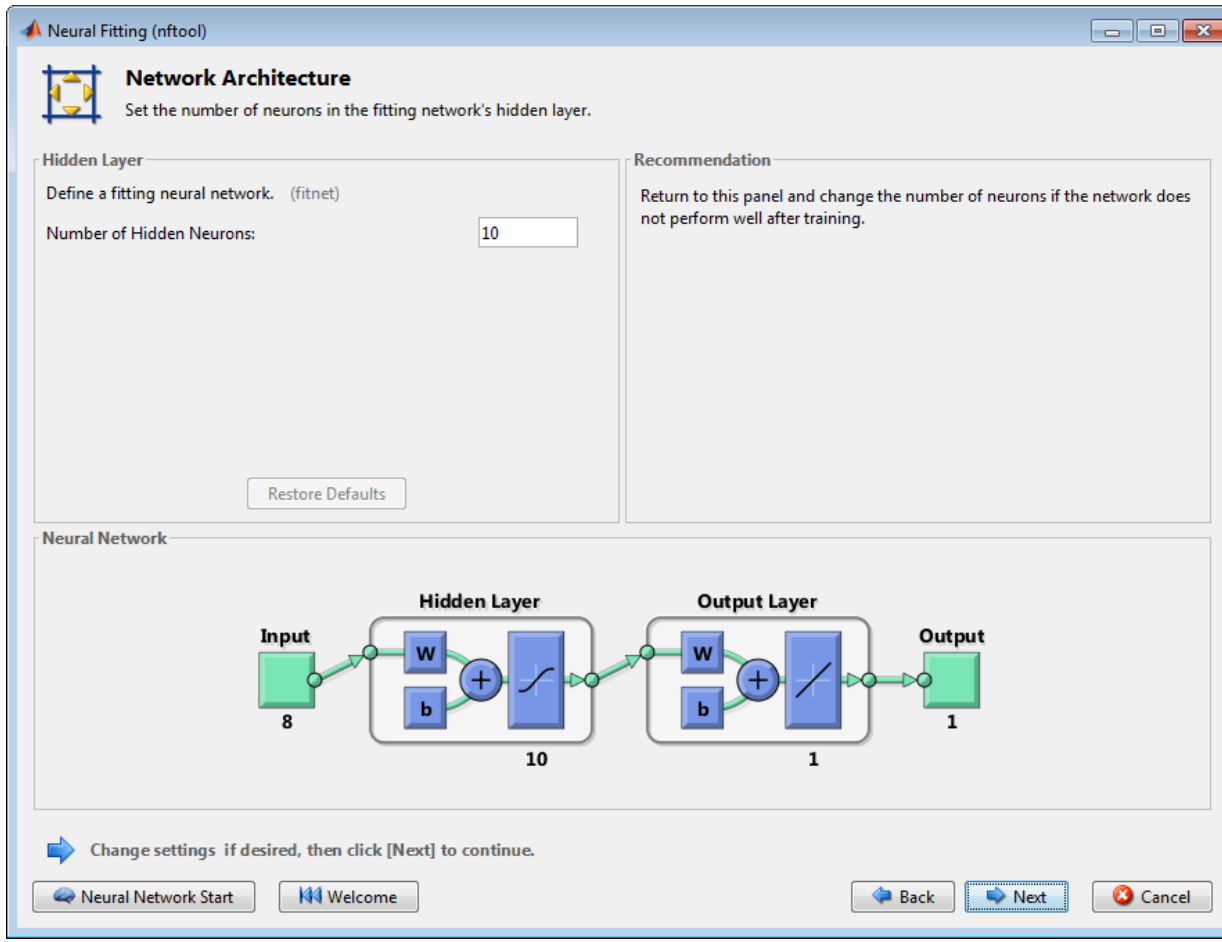
With these settings, the input vectors and target vectors will be randomly divided into three sets as follows:

- 70% will be used for training.
- 15% will be used to validate that the network is generalizing and to stop training before overfitting.
- The last 15% will be used as a completely independent test of network generalization.

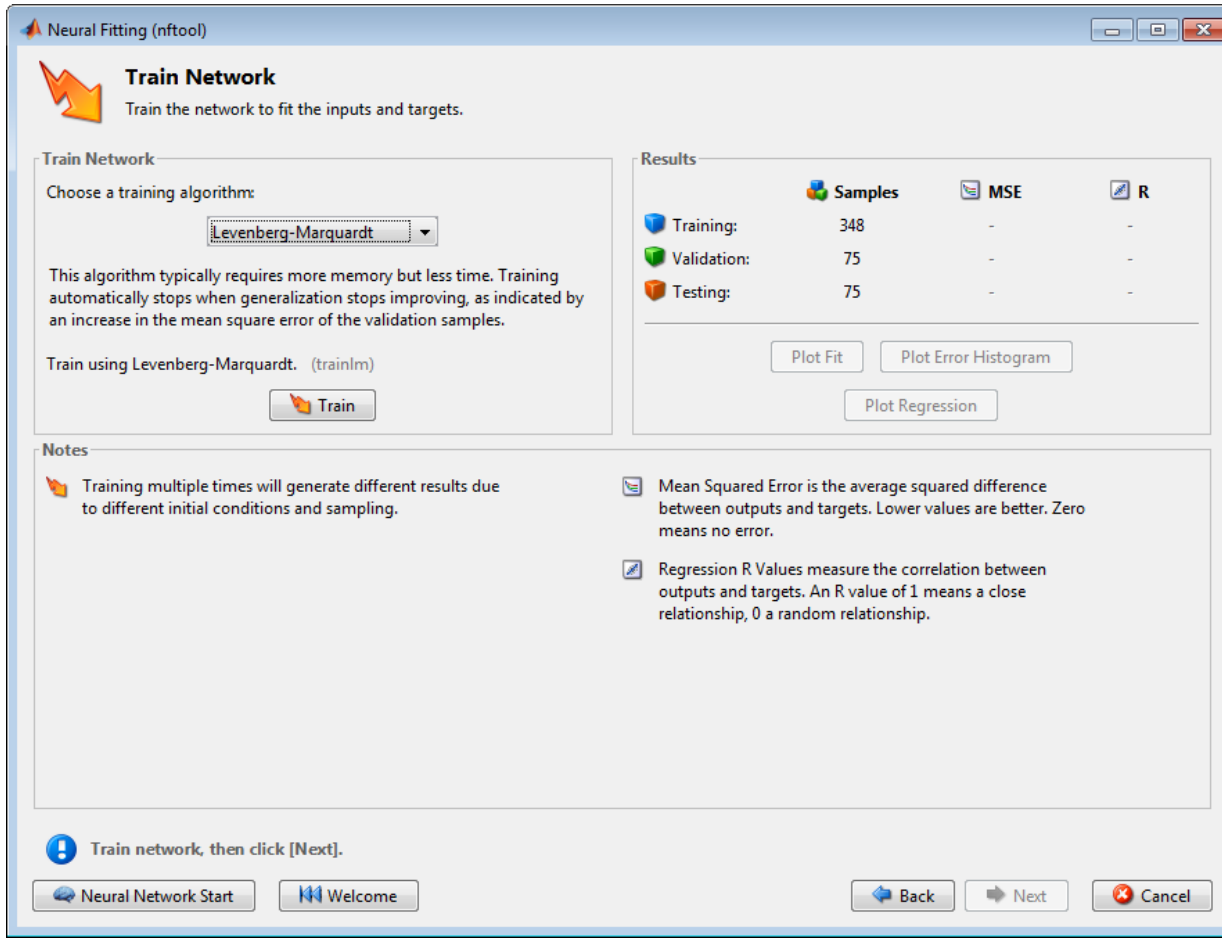
(See “Dividing the Data” for more discussion of the data division process.)

7 Click **Next**.

The standard network that is used for function fitting is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. The default number of hidden neurons is set to `10`. You might want to increase this number later, if the network training performance is poor.

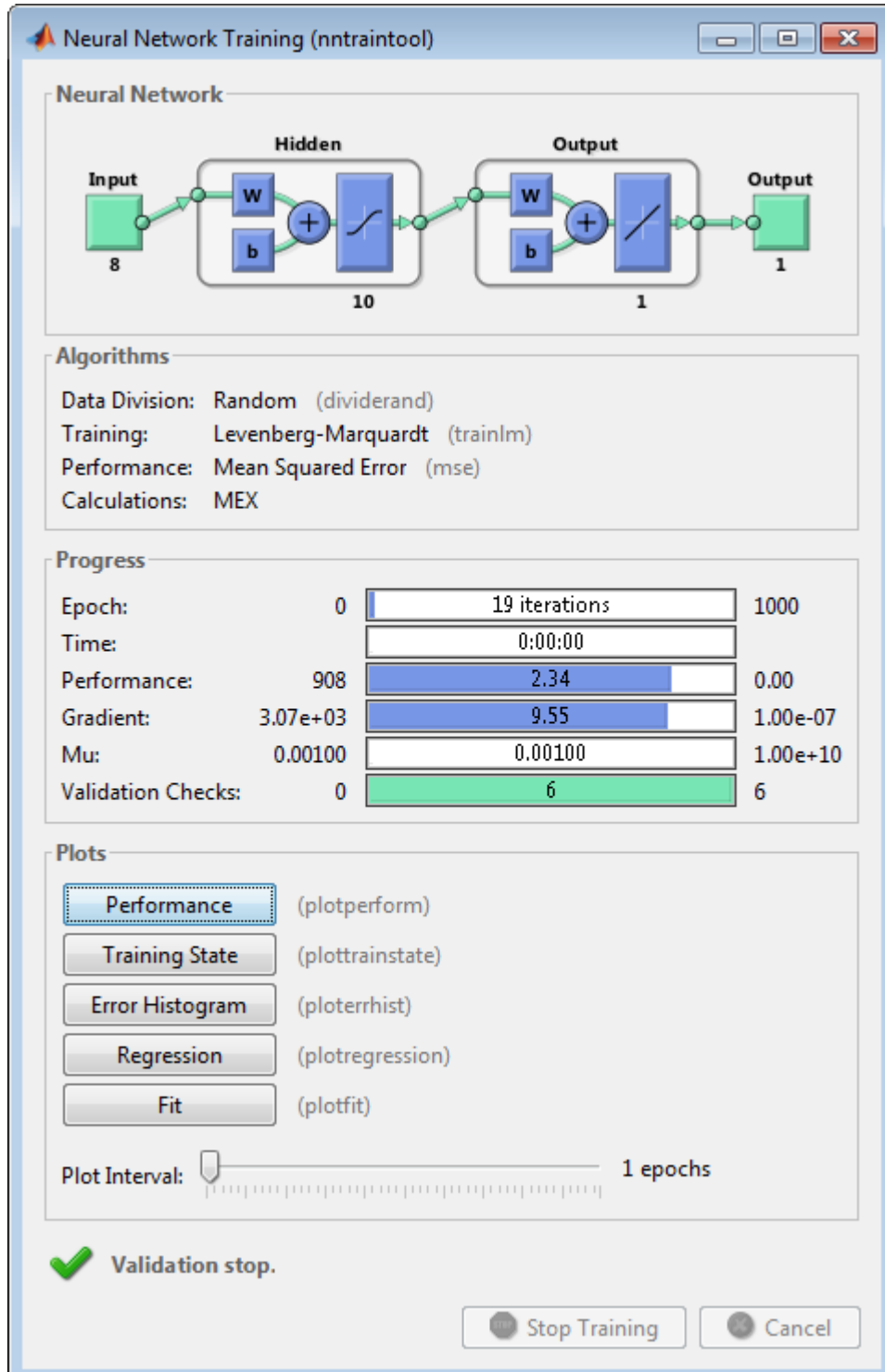


8 Click **Next**.



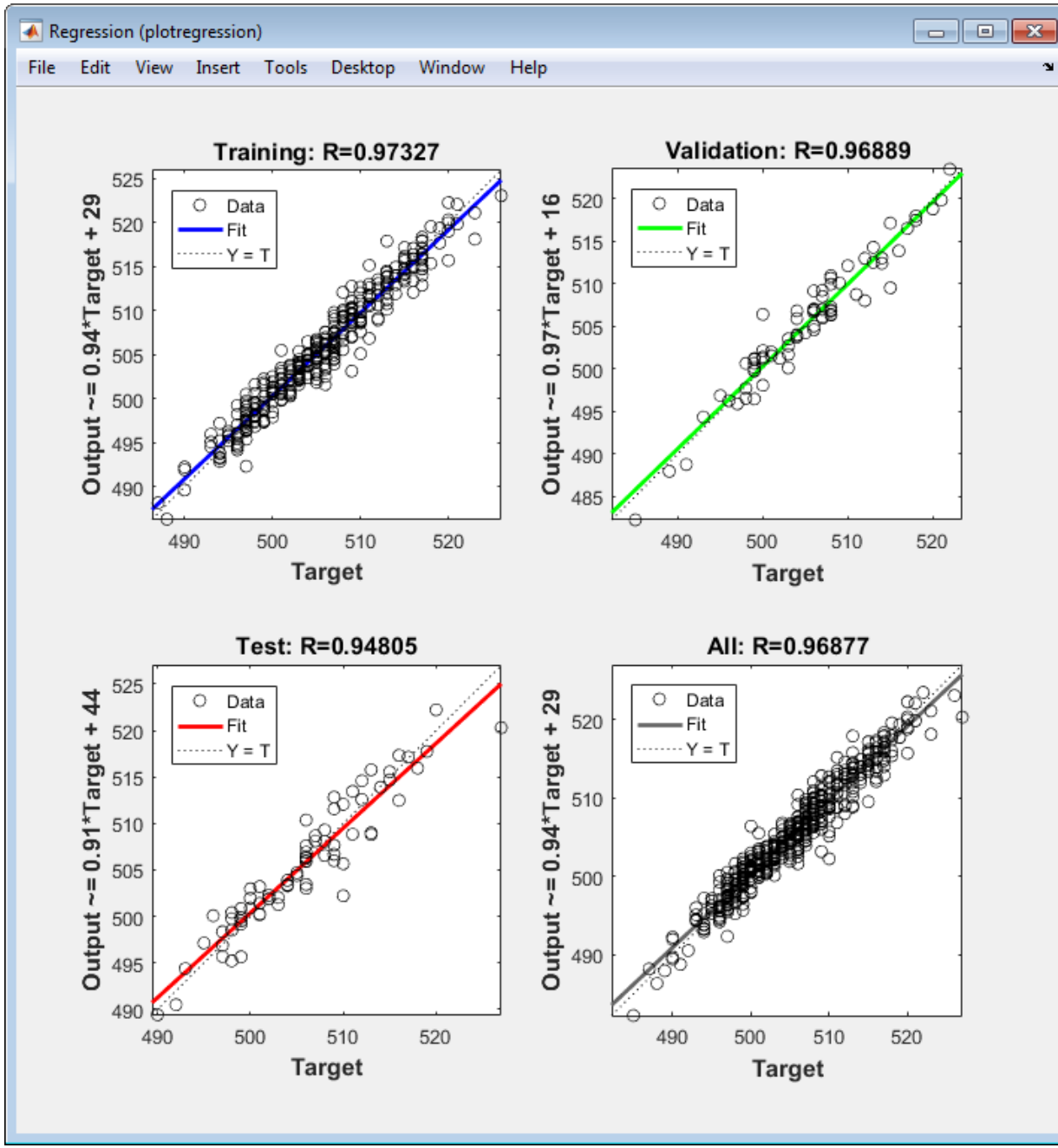
9 Select a training algorithm, then click **Train**. Levenberg-Marquardt (`trainlm`) is recommended for most problems, but for some noisy and small problems Bayesian Regularization (`trainbr`) can take longer but obtain a better solution. For large problems, however, Scaled Conjugate Gradient (`trainscg`) is recommended as it uses gradient calculations which are more memory efficient than the Jacobian calculations the other two algorithms use. This example uses the default Levenberg-Marquardt.

The training continued until the validation error failed to decrease for six iterations (validation stop).

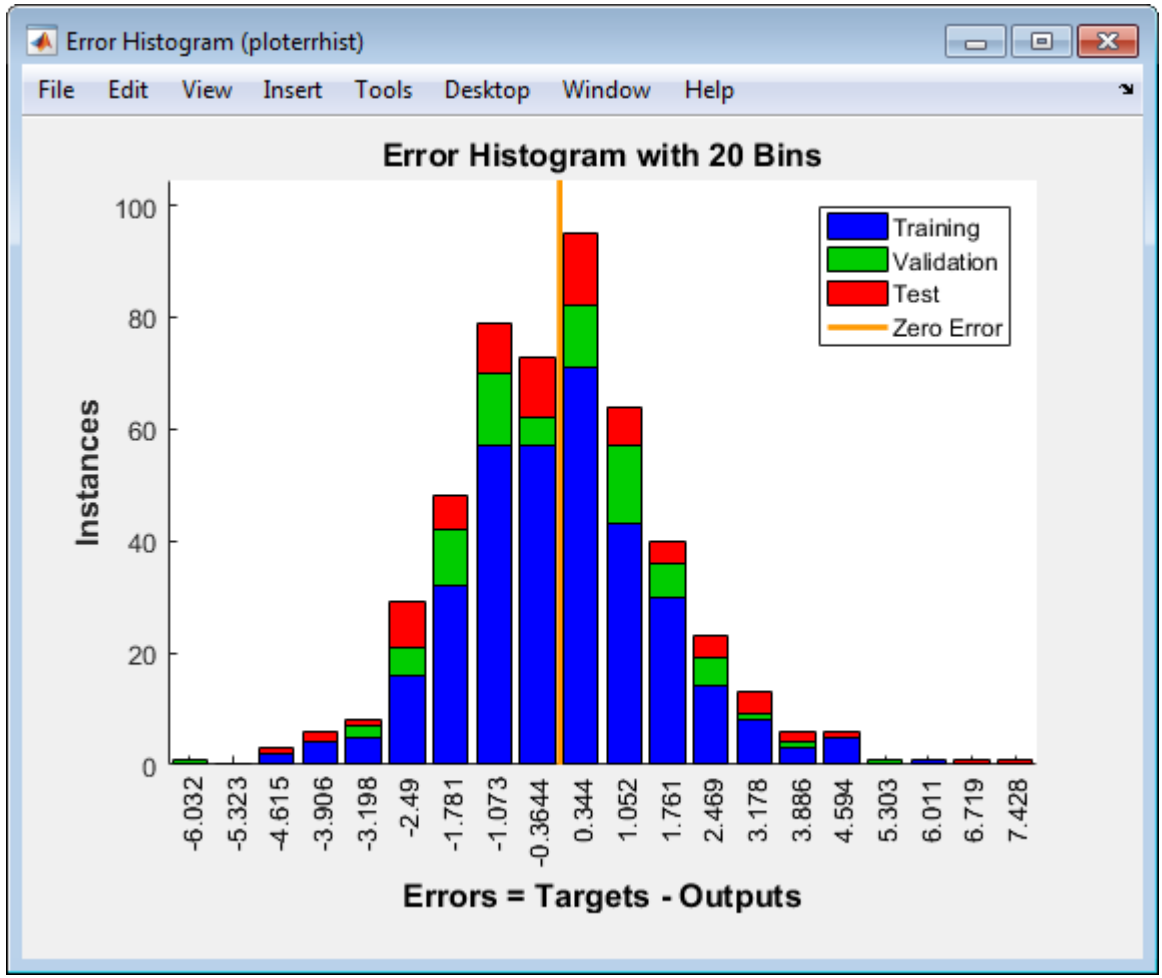

Neural Network Training (nntraintool)
Neural Network
Input
8
Hidden
W
b
10
Output
W
b
1
Output
1
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 19 iterations 1000
Time: 0:00:00
Performance: 908 2.34 0.00
Gradient: 3.07e+03 9.55 1.00e-07
Mu: 0.00100 0.00100 1.00e+10
Validation Checks: 0 6 6
Plots
Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Fit (plotfit)
Plot Interval: 1 epochs
Validation stop.
Stop Training
Cancel

10 Under **Plots**, click **Regression**. This is used to validate the network performance.

The following regression plots display the network outputs with respect to targets for training, validation, and test sets. For a perfect fit, the data should fall along a 45 degree line, where the network outputs are equal to the targets. For this problem, the fit is reasonably good for all data sets, with R values in each case of 0.93 or above. If even more accurate results were required, you could retrain the network by clicking **Retrain** in `nftool`. This will change the initial weights and biases of the network, and may produce an improved network after retraining. Other options are provided on the following pane.


Regression (plotregression)
File Edit View Insert Tools Desktop Window Help
Training: R=0.97327
Validation: R=0.96889
Test: R=0.94805
All: R=0.96877
Data
Fit
Y = T
Output ~= 0.94*Target + 29
Output ~= 0.97*Target + 16
Output ~= 0.91*Target + 44
Target

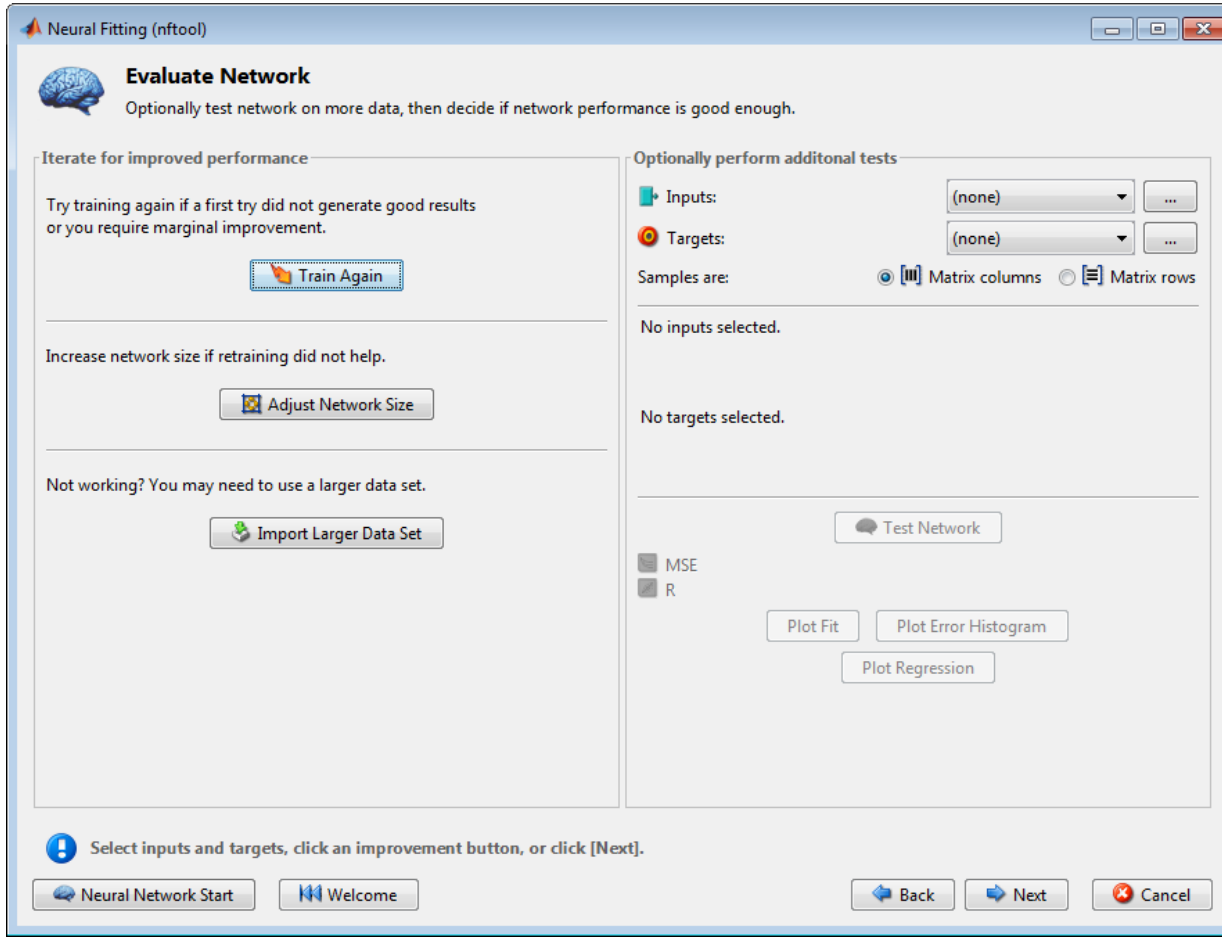
1 View the error histogram to obtain additional verification of network performance. Under the **Plots** pane, click **Error Histogram**.



The blue bars represent training data, the green bars represent validation data, and the red bars represent testing data. The histogram can give you an indication of outliers, which are data points where the fit is significantly worse than the majority of data. In this case, you can see that while most errors fall between -5 and 5, there is a training point with an error of 17 and validation points with errors of 12 and 13. These outliers are also visible on the testing regression plot. The first corresponds to

the point with a target of 50 and output near 33. It is a good idea to check the outliers to determine if the data is bad, or if those data points are different than the rest of the data set. If the outliers are valid data points, but are unlike the rest of the data, then the network is extrapolating for these points. You should collect more data that looks like the outlier points, and retrain the network.

2 Click **Next** in the Neural Network Fitting App to evaluate the network.



At this point, you can test the network against new data.

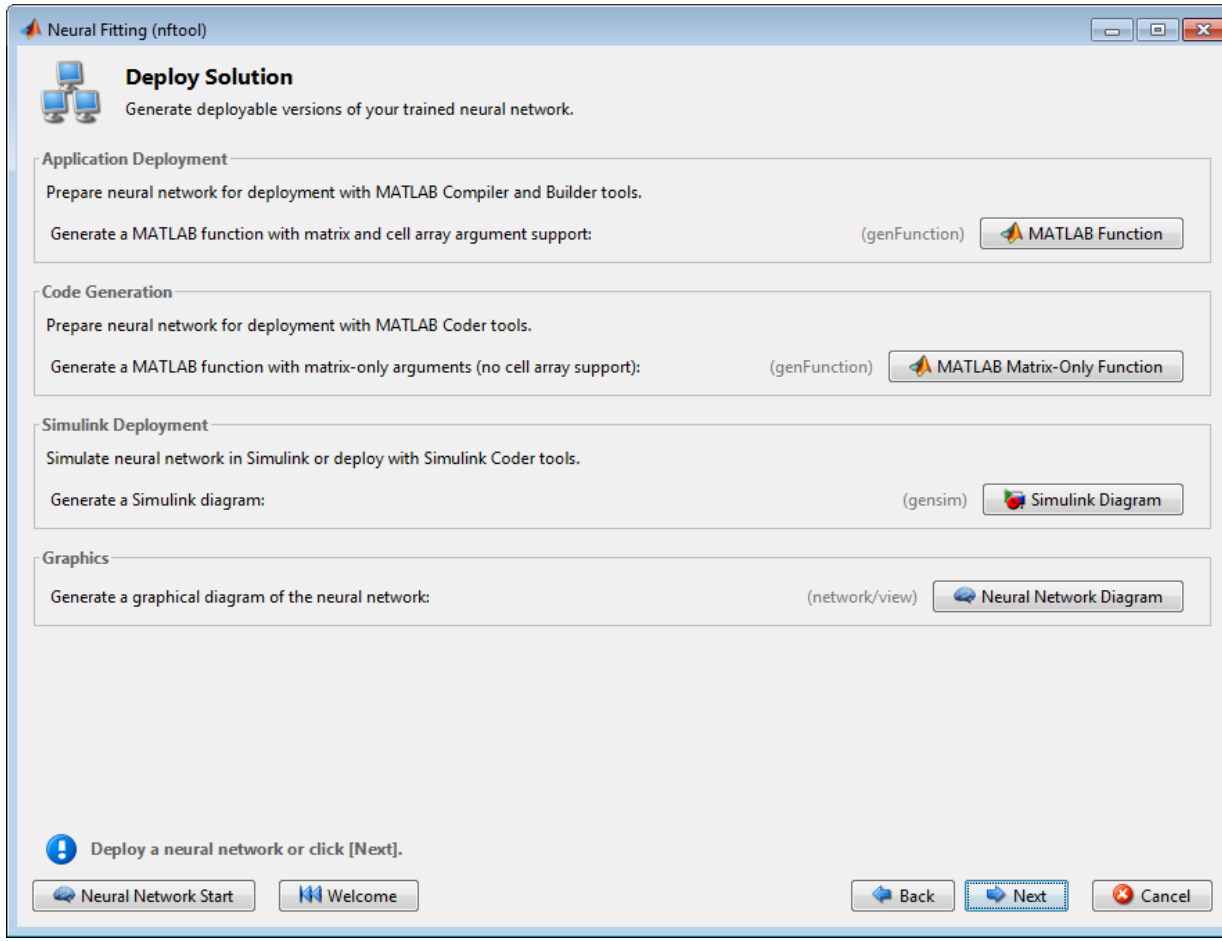
If you are dissatisfied with the network's performance on the original or new data, you can do one of the following:

- Train it again.
- Increase the number of neurons.
- Get a larger training data set.

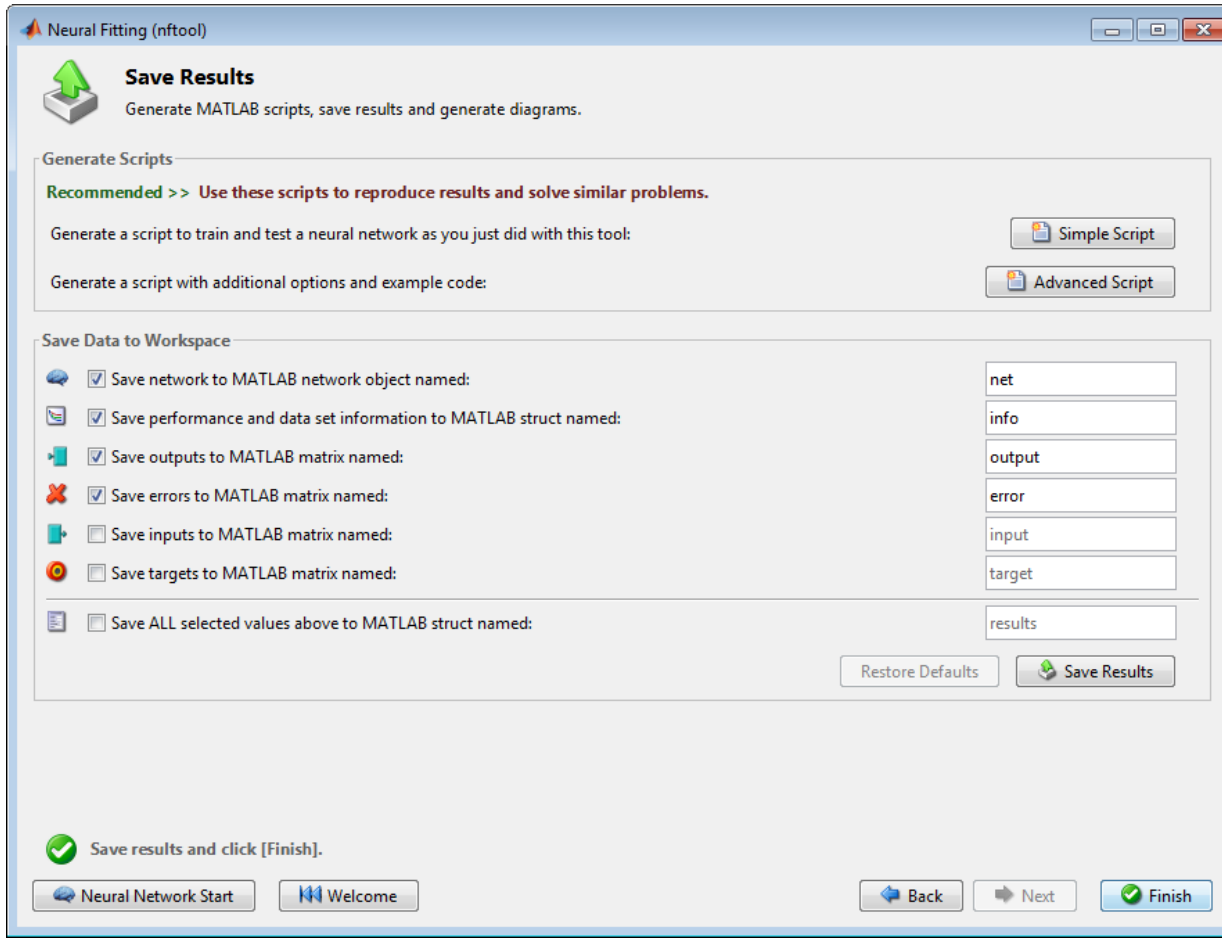
If the performance on the training set is good, but the test set performance is significantly worse, which could indicate overfitting, then reducing the number of neurons can improve your results. If training performance is poor, then you may want to increase the number of neurons.

3 If you are satisfied with the network performance, click **Next**.

4 Use this panel to generate a MATLAB function or Simulink® diagram for simulating your neural network. You can use the generated code or diagram to better understand how your neural network computes outputs from inputs, or deploy the network with MATLAB Compiler™ tools and other MATLAB code generation tools.



5 Use the buttons on this screen to generate scripts or to save your results.



- You can click **Simple Script** or **Advanced Script** to create MATLAB code that can be used to reproduce all of the previous steps from the command line. Creating MATLAB code can be helpful if you want to learn how to use the command-line functionality of the toolbox to customize the training process. In “Using Command-Line Functions” on page 1-51, you will investigate the generated scripts in more detail.
- You can also have the network saved as `net` in the workspace. You can perform additional tests on it or put it to work on new inputs.

6 When you have created the MATLAB code and saved your results, click **Finish**.

Using Command-Line Functions

The easiest way to learn how to use the command-line functionality of the toolbox is to generate scripts from the GUIs, and then modify them to customize the network training. As an example, look at the simple script that was created at step 14 of the previous section.

```
% Solve an Input-Output Fitting problem with a Neural Network
% Script generated by NFTOOL
%
% This script assumes these variables are defined:
%
%   houseInputs - input data.
%   houseTargets - target data.

inputs = houseInputs;
targets = houseTargets;

% Create a Fitting Network
hiddenLayerSize = 10;
net = fitnet(hiddenLayerSize);

% Set up Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;

% Train the Network
[net,tr] = train(net,inputs,targets);

% Test the Network
outputs = net(inputs);
errors = gsubtract(outputs,targets);
performance = perform(net,targets,outputs)

% View the Network
view(net)

% Plots
% Uncomment these lines to enable various plots.
% figure, plotperform(tr)
% figure, plottrainstate(tr)
% figure, plotfit(targets,outputs)
```

```
% figure, plotregression(targets,outputs)
% figure, ploterrhist(errors)
```

You can save the script, and then run it from the command line to reproduce the results of the previous GUI session. You can also edit the script to customize the training process. In this case, follow each step in the script.

1. The script assumes that the input vectors and target vectors are already loaded into the workspace. If the data are not loaded, you can load them as follows:

   ```
   load bodyfat_dataset
   inputs = bodyfatInputs;
   targets = bodyfatTargets;
   ```

   This data set is one of the sample data sets that is part of the toolbox (see “Sample Data Sets for Shallow Neural Networks” on page 1-137). You can see a list of all available data sets by entering the command `help nndatasets`. The `load` command also allows you to load the variables from any of these data sets using your own variable names. For example, the command

   ```
   [inputs,targets] = bodyfat_dataset;
   ```

   will load the body fat inputs into the array `inputs` and the body fat targets into the array `targets`.

2. Create a network. The default network for function fitting (or regression) problems, `fitnet`, is a feedforward network with the default tan-sigmoid transfer function in the hidden layer and linear transfer function in the output layer. You assigned ten neurons (somewhat arbitrary) to the one hidden layer in the previous section. The network has one output neuron, because there is only one target value associated with each input vector.

   ```
   hiddenLayerSize = 10;
   net = fitnet(hiddenLayerSize);
   ```

   Note More neurons require more computation, and they have a tendency to overfit the data when the number is set too high, but they allow the network to solve more complicated problems. More layers require more computation, but their use might result in the network solving complex problems more efficiently. To use more than one hidden layer, enter the hidden layer sizes as elements of an array in the `fitnet` command.

3. Set up the division of data.

```
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
```

With these settings, the input vectors and target vectors will be randomly divided, with 70% used for training, 15% for validation and 15% for testing. (See “Dividing the Data” for more discussion of the data division process.)

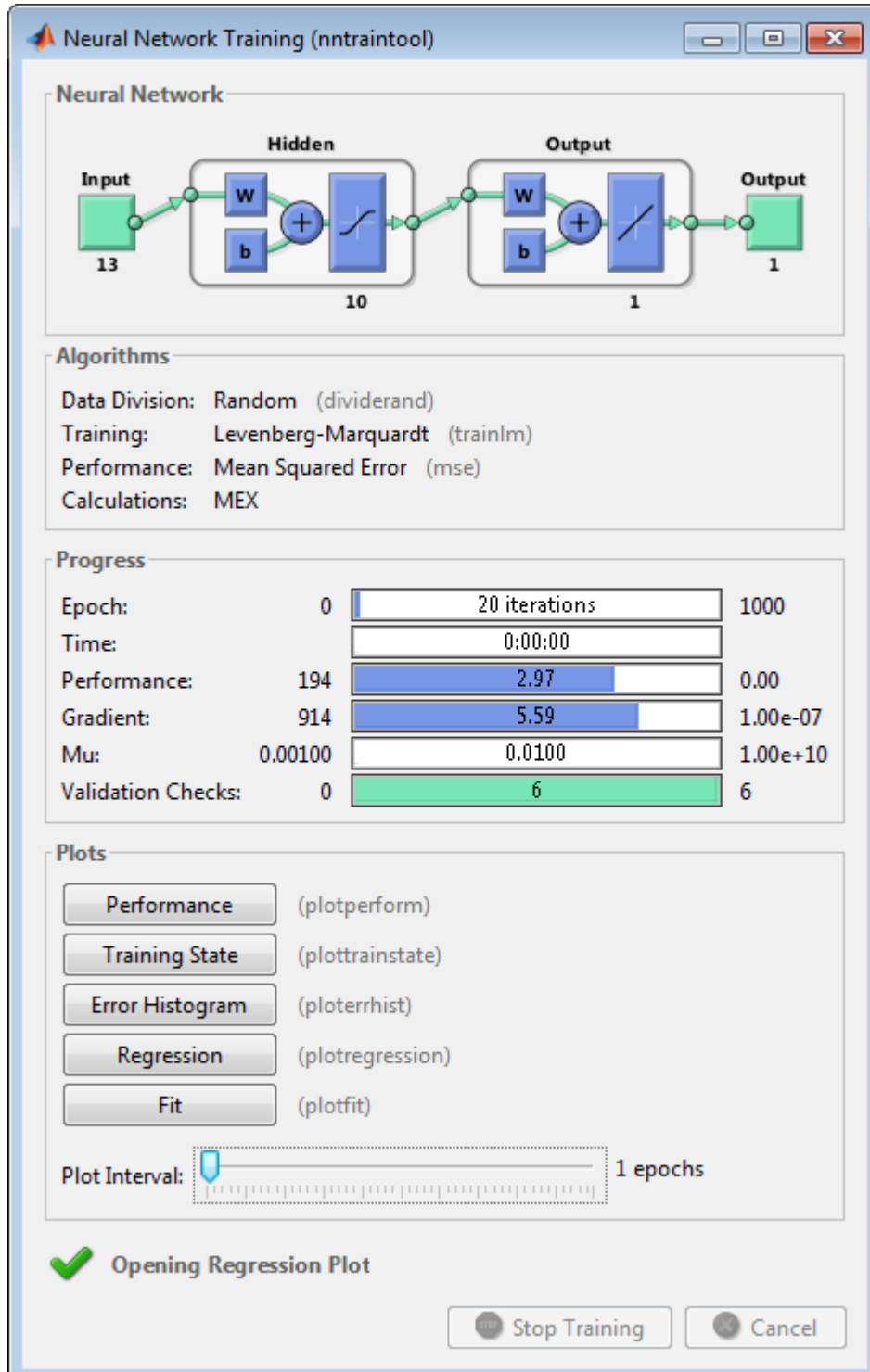
4 Train the network. The network uses the default Levenberg-Marquardt algorithm (`trainlm`) for training. For problems in which Levenberg-Marquardt does not produce as accurate results as desired, or for large data problems, consider setting the network training function to Bayesian Regularization (`trainbr`) or Scaled Conjugate Gradient (`trainscg`), respectively, with either

```
net.trainFcn = 'trainbr';
net.trainFcn = 'trainscg';
```

To train the network, enter:

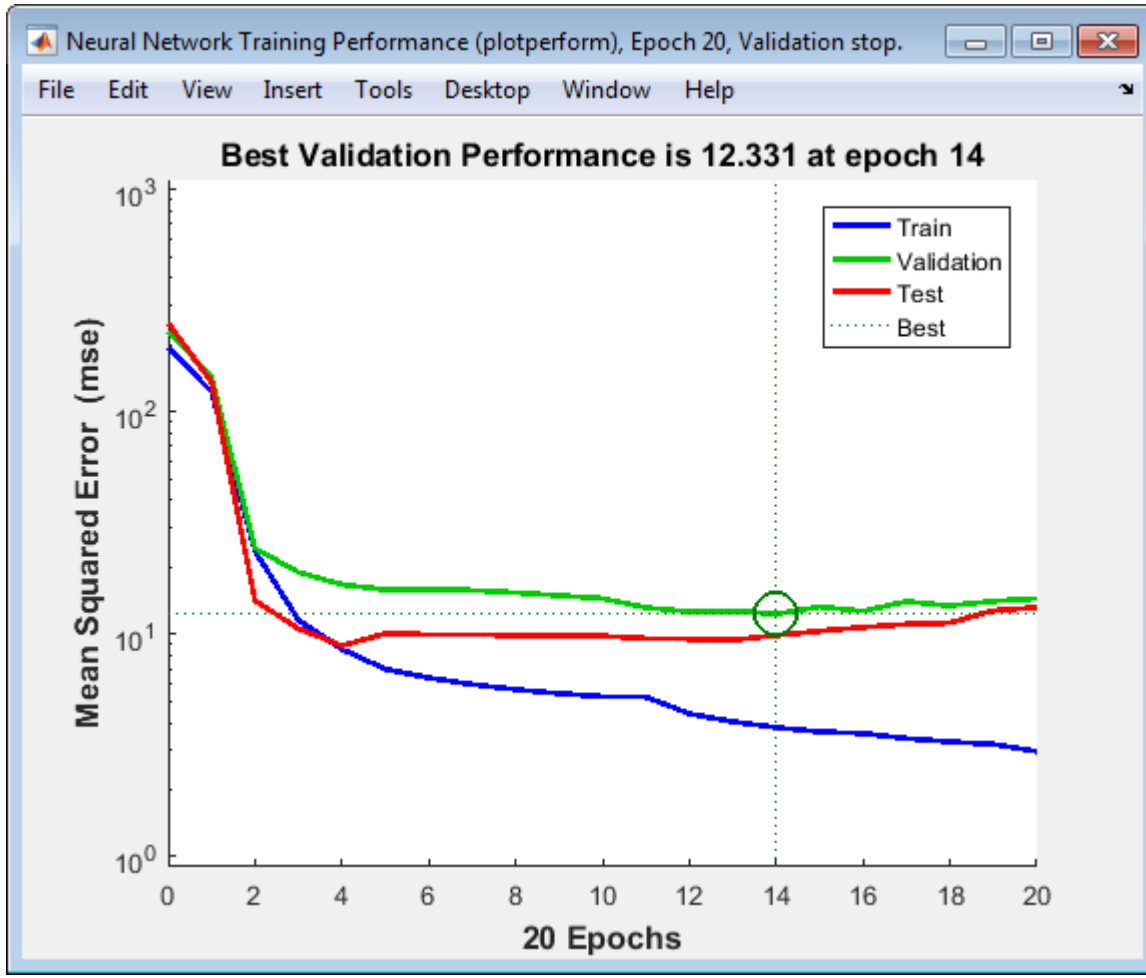
```
[net, tr] = train(net, inputs, targets);
```

During training, the following training window opens. This window displays training progress and allows you to interrupt training at any point by clicking **Stop Training**.


Neural Network Training (nntraintool)
Neural Network
Input
13
Hidden
W
b
10
Output
W
b
1
Output
1
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 20 iterations 1000
Time: 0:00:00
Performance: 194 2.97 0.00
Gradient: 914 5.59 1.00e-07
Mu: 0.00100 0.0100 1.00e+10
Validation Checks: 0 6 6
Plots
Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Fit (plotfit)
Plot Interval: 1 epochs
Opening Regression Plot
Stop Training
Cancel

This training stopped when the validation error increased for six iterations, which occurred at iteration 20. If you click **Performance** in the training window, a plot of the training errors, validation errors, and test errors appears, as shown in the following figure. In this example, the result is reasonable because of the following considerations:

- The final mean-square error is small.
- The test set error and the validation set error have similar characteristics.
- No significant overfitting has occurred by iteration 14 (where the best validation performance occurs).



5 Test the network. After the network has been trained, you can use it to compute the network outputs. The following code calculates the network outputs, errors and overall performance.

```
outputs = net(inputs);
errors = gsubtract(targets, outputs);
performance = perform(net, targets, outputs)


performance =
```

```
   19.3193
```

It is also possible to calculate the network performance only on the test set, by using the testing indices, which are located in the training record. (See “Analyze Shallow Neural Network Performance After Training” for a full description of the training record.)

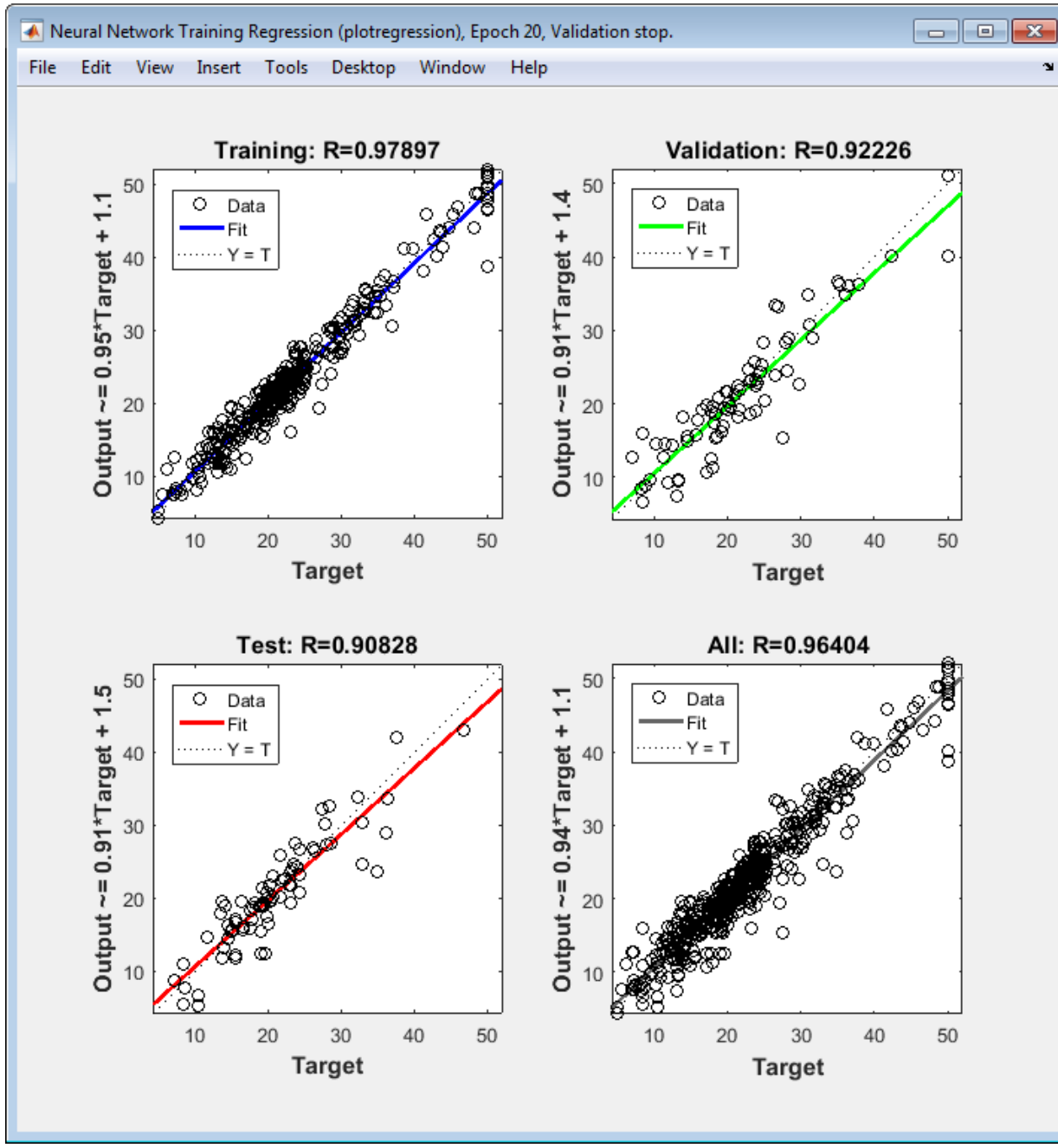
```
tInd = tr.testInd;
tstOutputs = net(inputs(:, tInd));
tstPerform = perform(net, targets(tInd), tstOutputs)
```

```
tstPerform =

   53.7680
```

6 Perform some analysis of the network response. If you click **Regression** in the training window, you can perform a linear regression between the network outputs and the corresponding targets.

The following figure shows the results.


Neural Network Training Regression (plotregression), Epoch 20, Validation stop.
File Edit View Insert Tools Desktop Window Help
Training: R=0.97897
Output ~= 0.95*Target + 1.1
Data
Fit
Y = T
Target
Validation: R=0.92226
Output ~= 0.91*Target + 1.4
Test: R=0.90828
Output ~= 0.91*Target + 1.5
All: R=0.96404
Output ~= 0.94*Target + 1.1

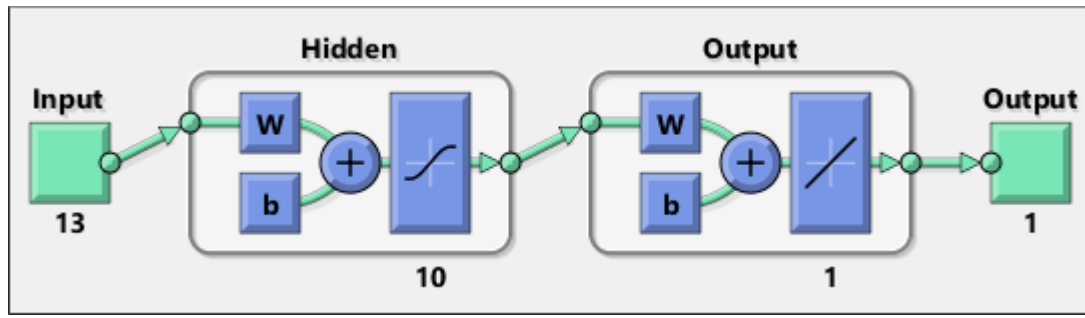
The output tracks the targets very well for training, testing, and validation, and the R-value is over 0.96 for the total response. If even more accurate results were required, you could try any of these approaches:

- Reset the initial network weights and biases to new values with `init` and train again (see “Initializing Weights” (init)).
- Increase the number of hidden neurons.
- Increase the number of training vectors.
- Increase the number of input values, if more relevant information is available.
- Try a different training algorithm (see “Training Algorithms”).

In this case, the network response is satisfactory, and you can now put the network to use on new inputs.

7 View the network diagram.

```
view(net)
```



To get more experience in command-line operations, try some of these tasks:

- During training, open a plot window (such as the regression plot), and watch it animate.
- Plot from the command line with functions such as `plotfit`, `plotregression`, `plottrainstate` and `plotperform`. (For more information on using these functions, see their reference pages.)

Also, see the advanced script for more options, when training from the command line.

Each time a neural network is trained, can result in a different solution due to different initial weight and bias values and different divisions of data into training, validation, and

test sets. As a result, different neural networks trained on the same problem can give different outputs for the same input. To ensure that a neural network of good accuracy has been found, retrain several times.

There are several other techniques for improving upon initial solutions if higher accuracy is desired. For more information, see “Improve Shallow Neural Network Generalization and Avoid Overfitting”.

Classify Patterns with a Shallow Neural Network

In addition to function fitting, neural networks are also good at recognizing patterns.

For example, suppose you want to classify a tumor as benign or malignant, based on uniformity of cell size, clump thickness, mitosis, etc. You have 699 example cases for which you have 9 items of data and the correct classification as benign or malignant.

As with function fitting, there are two ways to solve this problem:

- Use the `nprtool` GUI, as described in “Using the Neural Network Pattern Recognition App” on page 1-62.
- Use a command-line solution, as described in “Using Command-Line Functions” on page 1-79.

It is generally best to start with the GUI, and then to use the GUI to automatically generate command-line scripts. Before using either method, the first step is to define the problem by selecting a data set. The next section describes the data format.

Defining a Problem

To define a pattern recognition problem, arrange a set of Q input vectors as columns in a matrix. Then arrange another set of Q target vectors so that they indicate the classes to which the input vectors are assigned (see “Data Structures” for a detailed description of data formatting for static and time series data).

When there are only two classes; you set each scalar target value to either 0 or 1, indicating which class the corresponding input belongs to. For instance, you can define the two-class exclusive-or classification problem as follows:

```
inputs  = [0 1 0 1; 0 0 1 1];
targets = [1 0 0 1; 0 1 1 0];
```

When inputs are to be classified into *N* different classes, the target vectors have N elements. For each target vector, one element is 1 and the others are 0. For example, the following lines show how to define a classification problem that divides the corners of a 5-by-5-by-5 cube into three classes:

- The origin (the first input vector) in one class
- The corner farthest from the origin (the last input vector) in a second class

- All other points in a third class

```
inputs  = [0 0 0 0 5 5 5 5; 0 0 5 5 0 0 5 5; 0 5 0 5 0 5 0 5];
targets = [1 0 0 0 0 0 0 0; 0 1 1 1 1 1 1 0; 0 0 0 0 0 0 0 1];
```

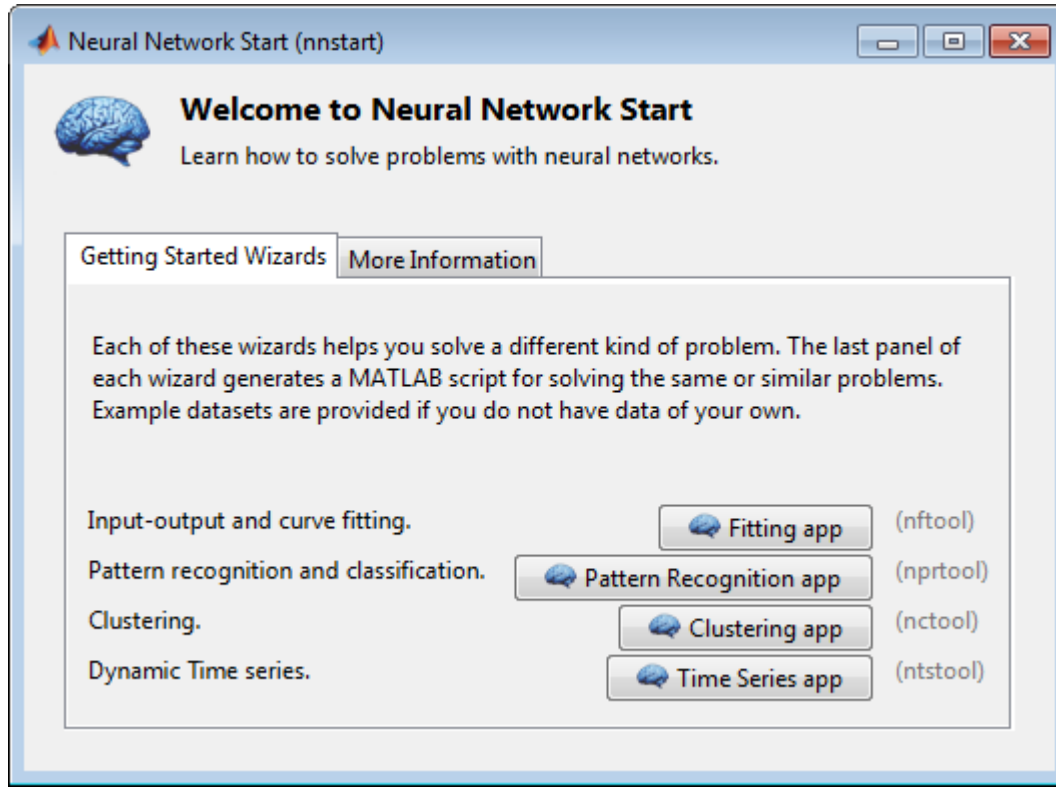
Classification problems involving only two classes can be represented using either format. The targets can consist of either scalar 1/0 elements or two-element vectors, with one element being 1 and the other element being 0.

The next section shows how to train a network to recognize patterns, using the neural network pattern recognition app, `nprtool`. This example uses the cancer data set provided with the toolbox. This data set consists of 699 nine-element input vectors and two-element target vectors. There are two elements in each target vector, because there are two categories (benign or malignant) associated with each input vector.

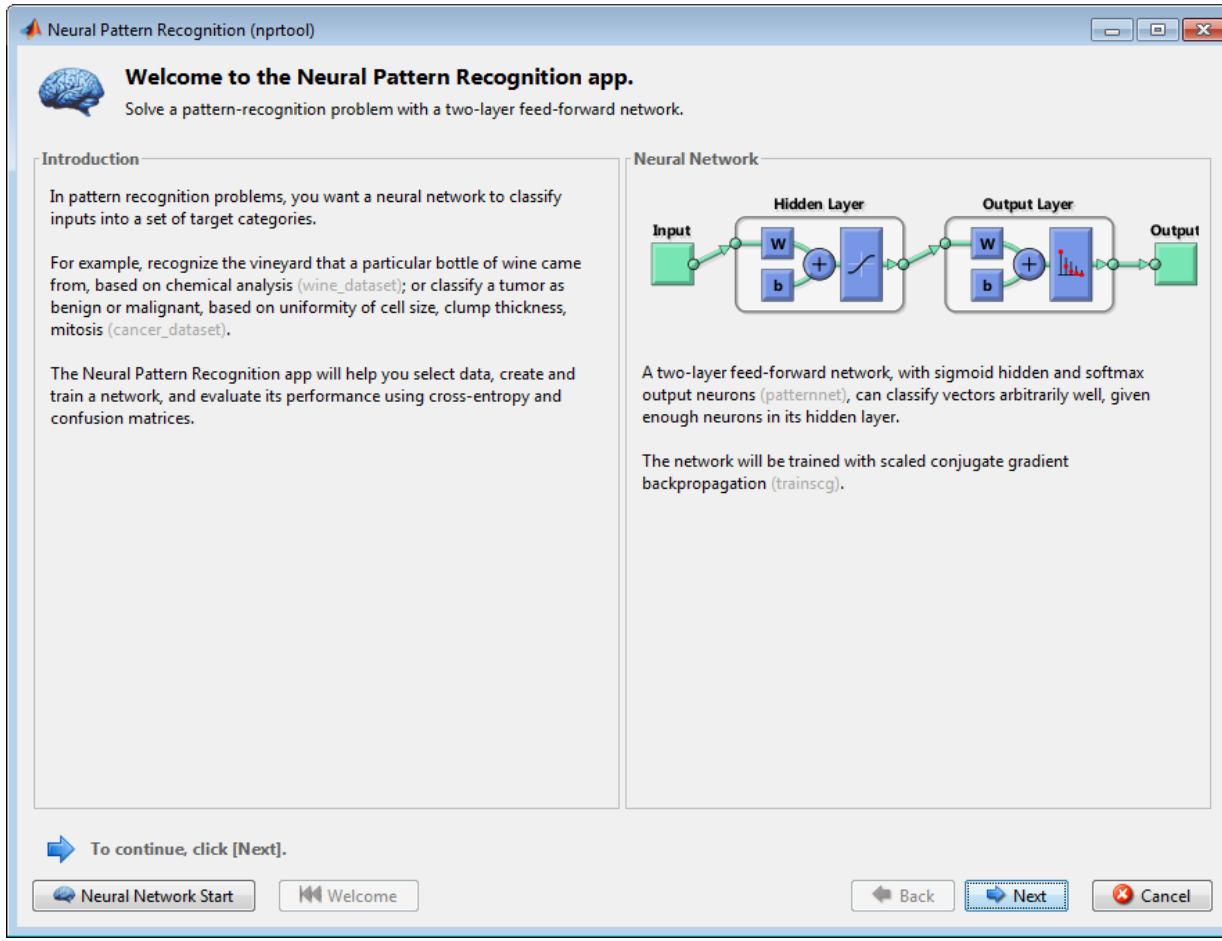
Using the Neural Network Pattern Recognition App

1 If needed, open the Neural Network Start GUI with this command:

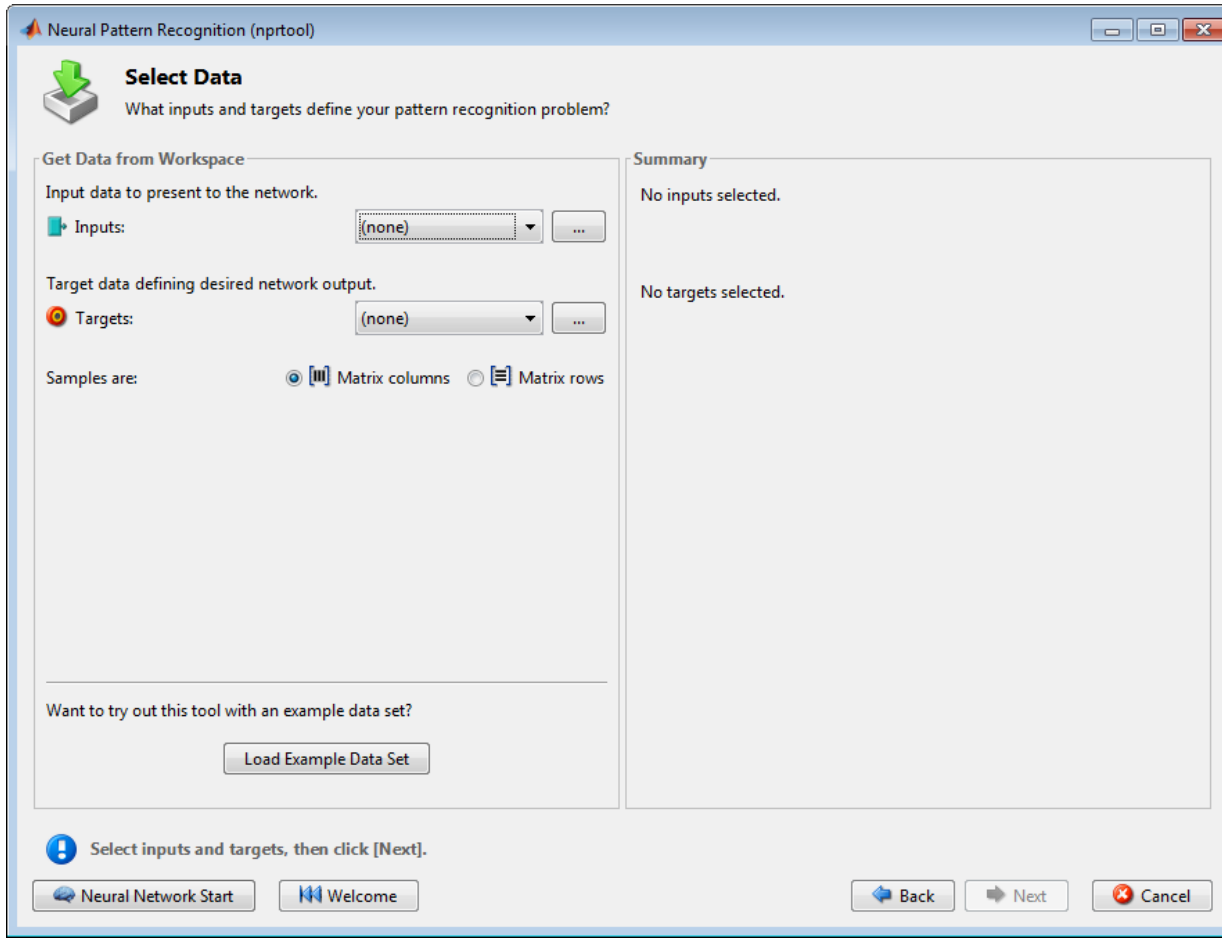
```
nnstart
```



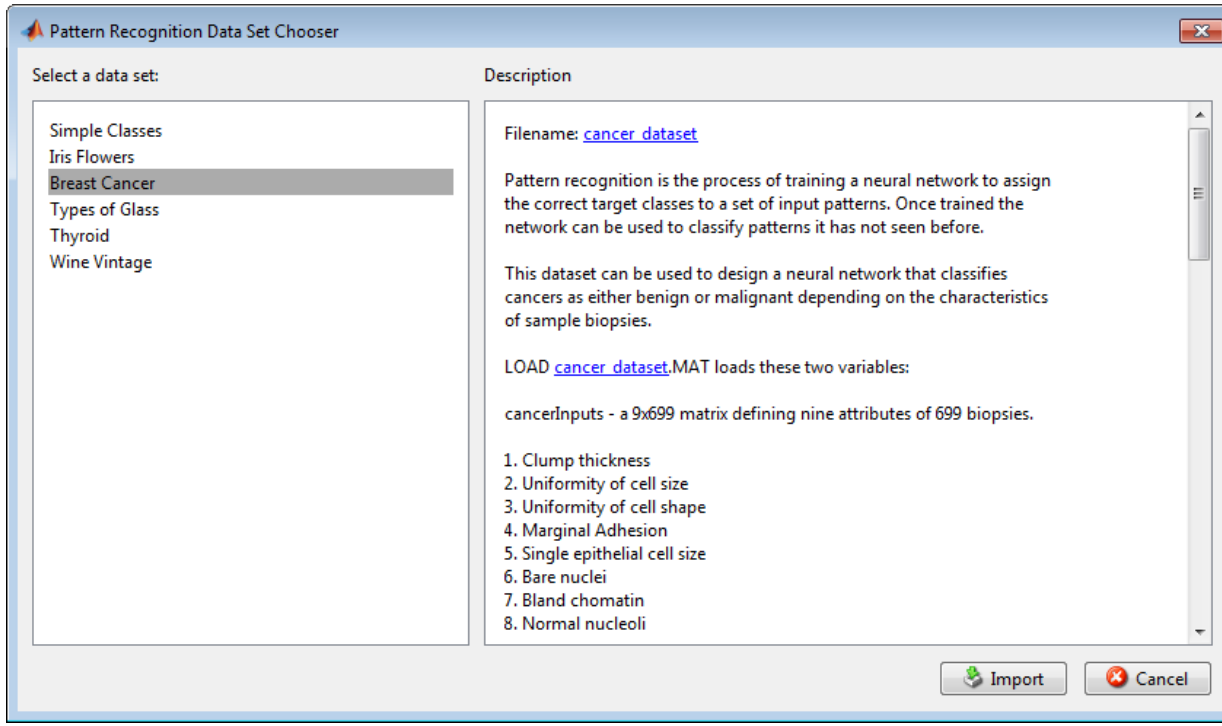
2 Click **Pattern Recognition app** to open the Neural Network Pattern Recognition app. (You can also use the command `nprtool`.)



3 Click **Next** to proceed. The Select Data window opens.

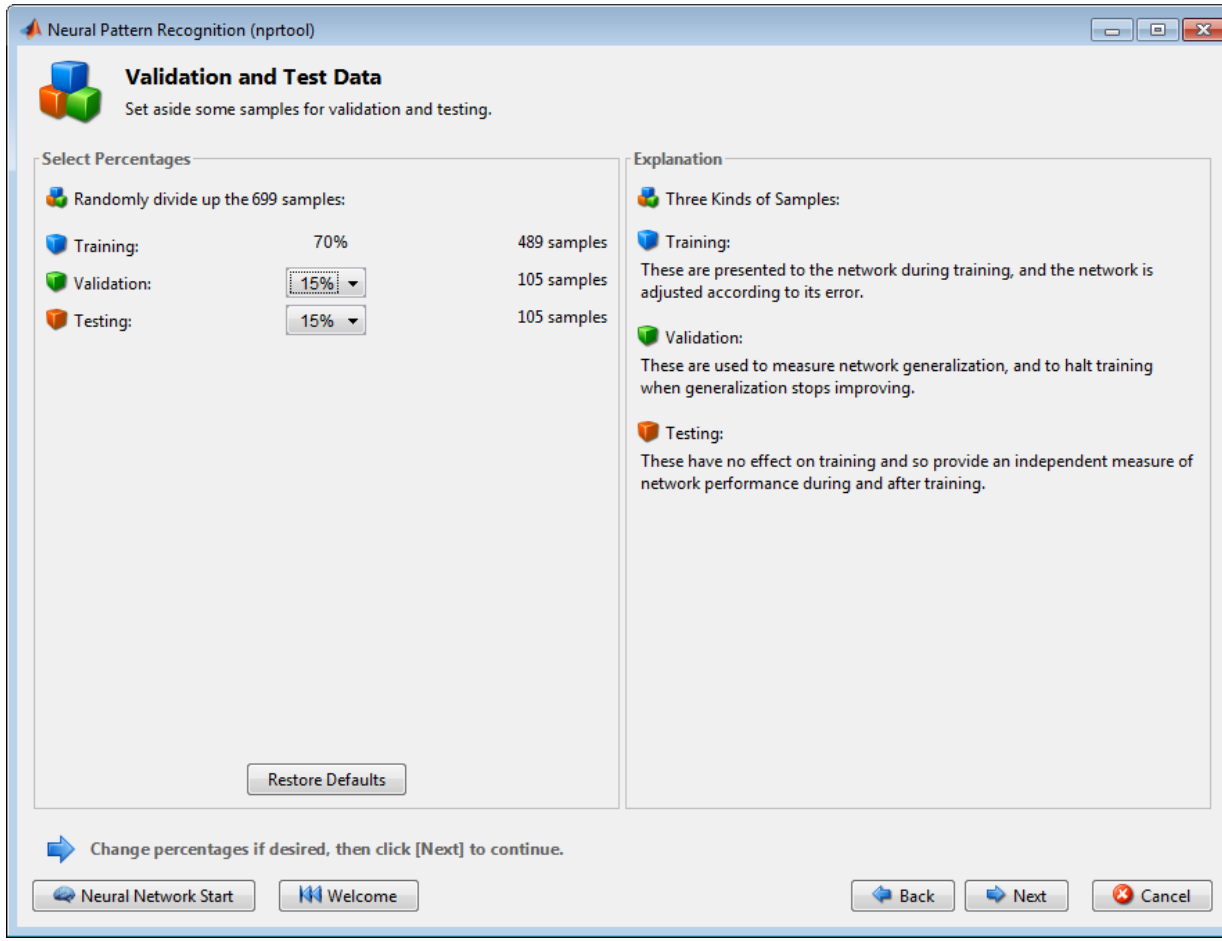


4 Click **Load Example Data Set**. The Pattern Recognition Data Set Chooser window opens.



5 Select **Breast Cancer** and click **Import**. You return to the Select Data window.

6 Click **Next** to continue to the Validation and Test Data window.



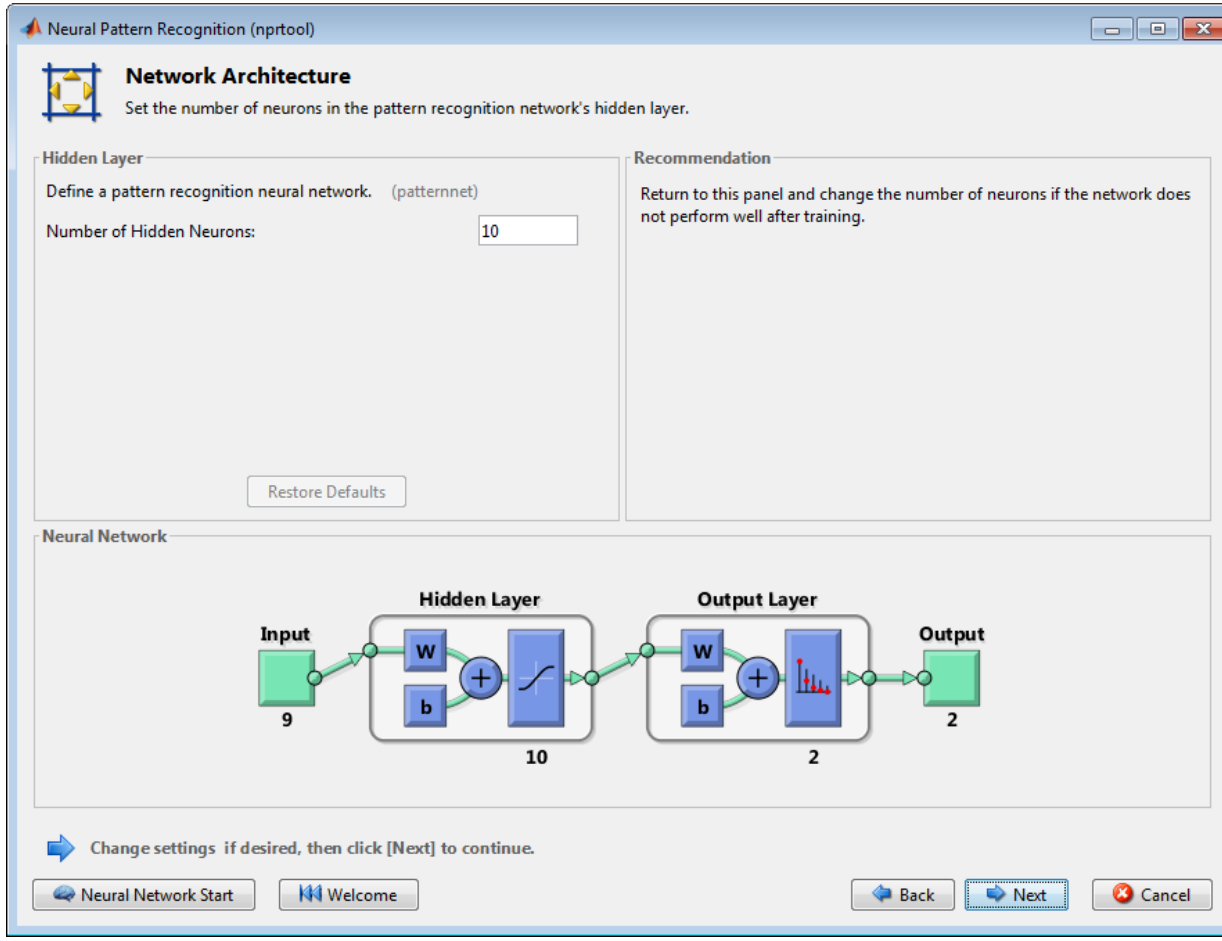
Validation and test data sets are each set to 15% of the original data. With these settings, the input vectors and target vectors will be randomly divided into three sets as follows:

- 70% are used for training.
- 15% are used to validate that the network is generalizing and to stop training before overfitting.
- The last 15% are used as a completely independent test of network generalization.

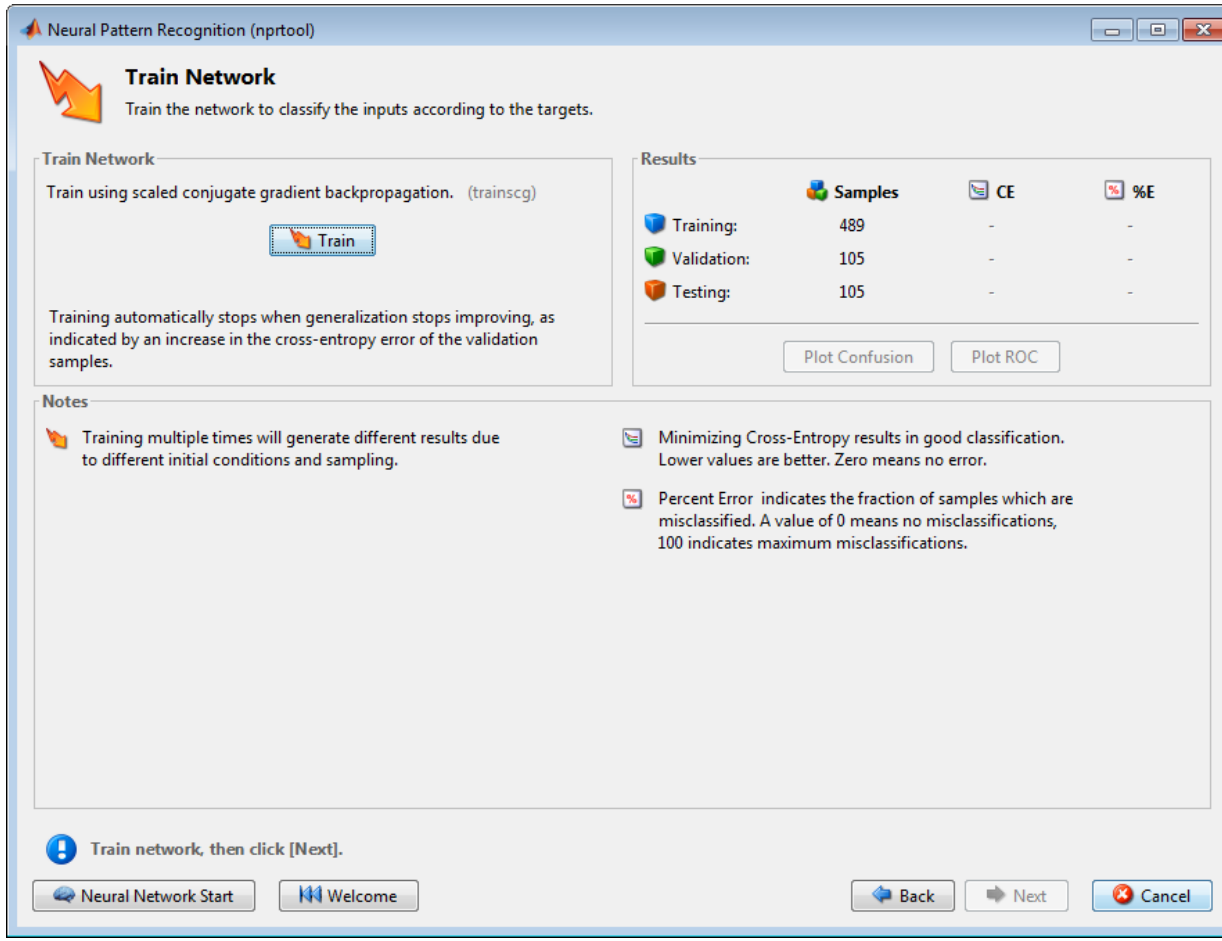
(See “Dividing the Data” for more discussion of the data division process.)

7 Click **Next**.

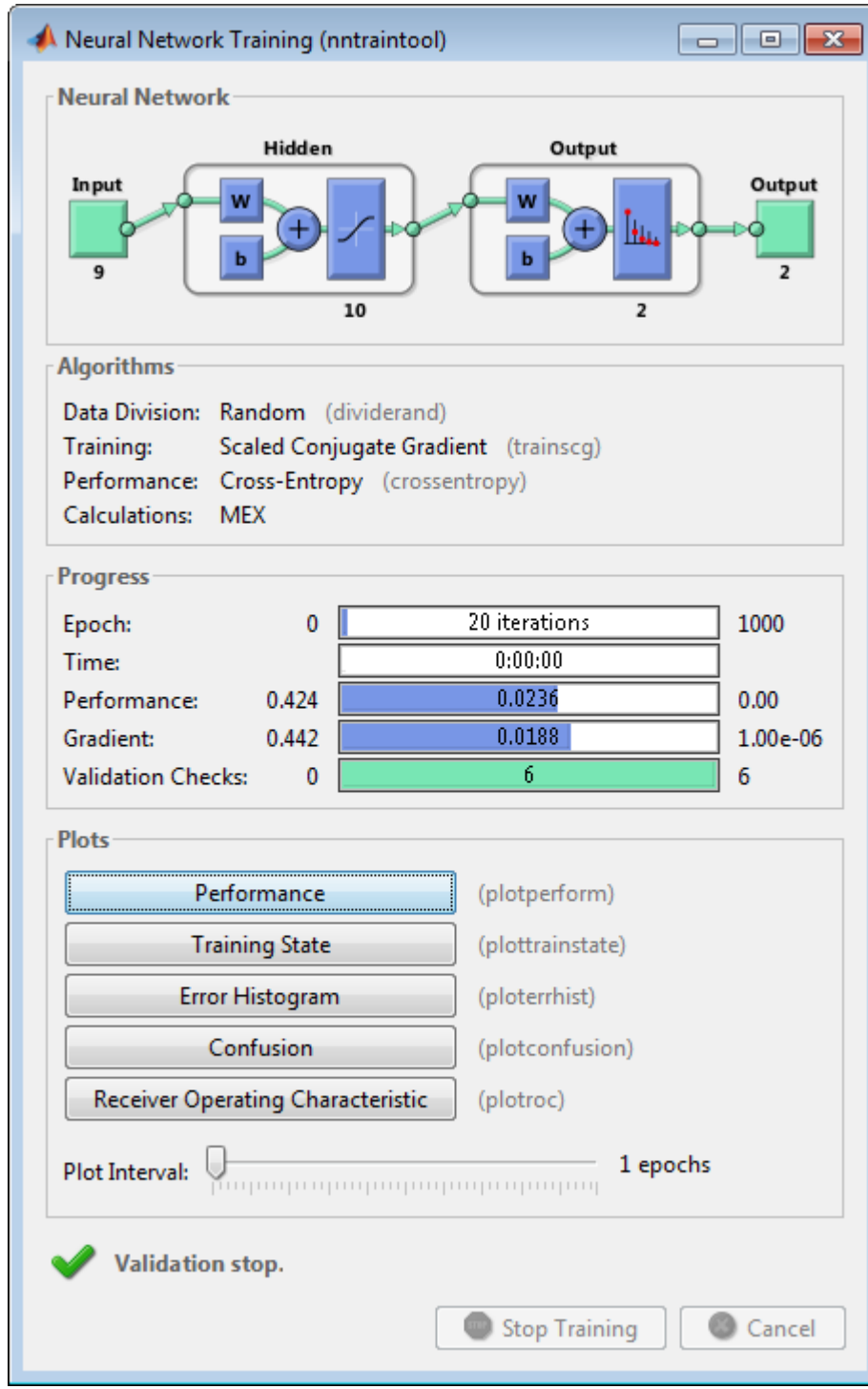
The standard network that is used for pattern recognition is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer. The default number of hidden neurons is set to 10. You might want to come back and increase this number if the network does not perform as well as you expect. The number of output neurons is set to 2, which is equal to the number of elements in the target vector (the number of categories).



8 Click **Next**.



9 Click **Train**.


Neural Network Training (nntraintool)
Neural Network
Input
9
Hidden
W
b
10
Output
W
b
2
Output
2
Algorithms
Data Division: Random (dividerand)
Training: Scaled Conjugate Gradient (trainscg)
Performance: Cross-Entropy (crossentropy)
Calculations: MEX
Progress
Epoch: 0 20 iterations 1000
Time: 0:00:00
Performance: 0.424 0.0236 0.00
Gradient: 0.442 0.0188 1.00e-06
Validation Checks: 0 6 6
Plots
Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Confusion (plotconfusion)
Receiver Operating Characteristic (plotroc)
Plot Interval: 1 epochs
Validation stop.
Stop Training
Cancel

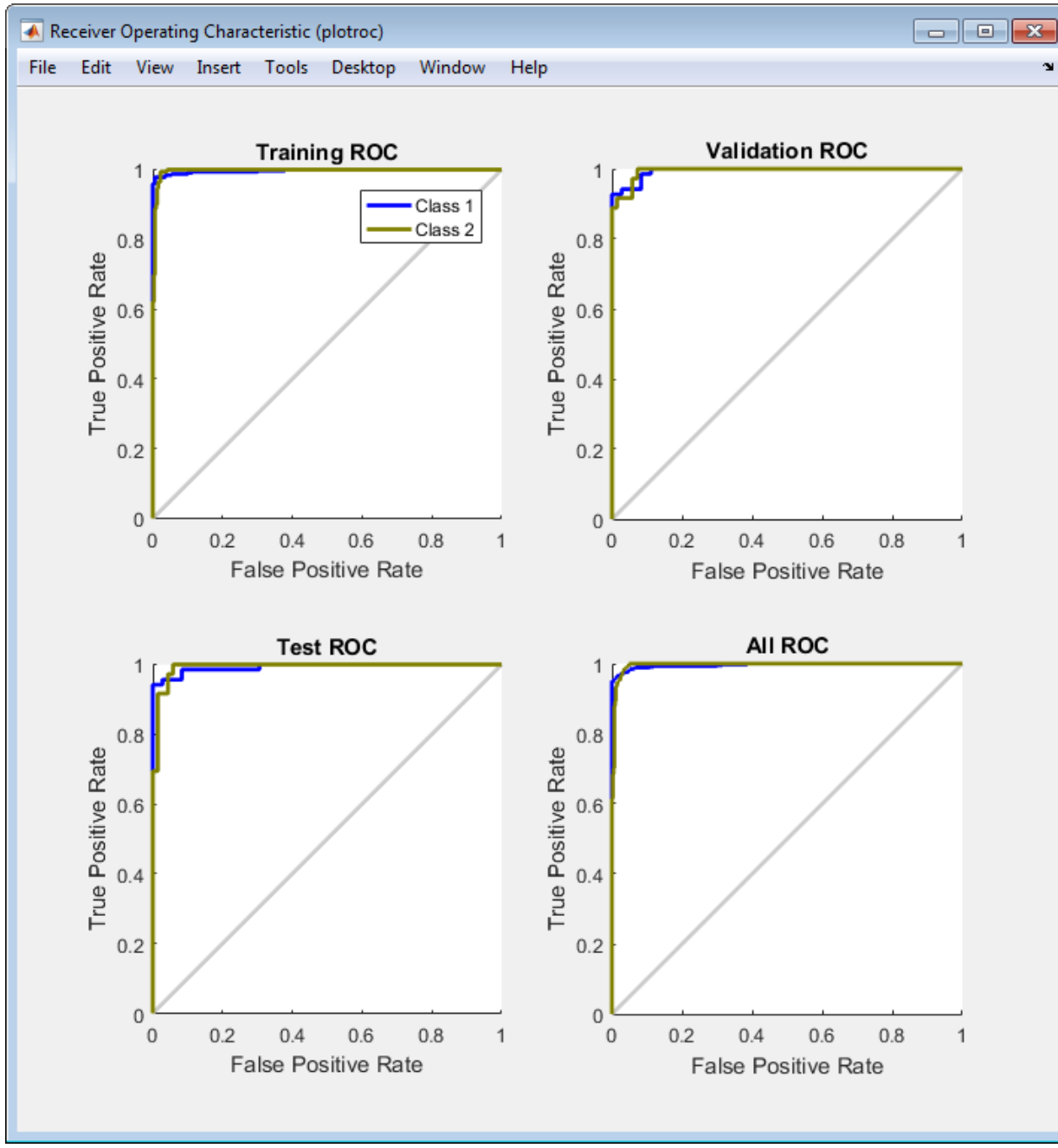
The training continues for 55 iterations.

10 Under the **Plots** pane, click **Confusion** in the Neural Network Pattern Recognition App.

The next figure shows the confusion matrices for training, testing, and validation, and the three kinds of data combined. The network outputs are very accurate, as you can see by the high numbers of correct responses in the green squares and the low numbers of incorrect responses in the red squares. The lower right blue squares illustrate the overall accuracies.

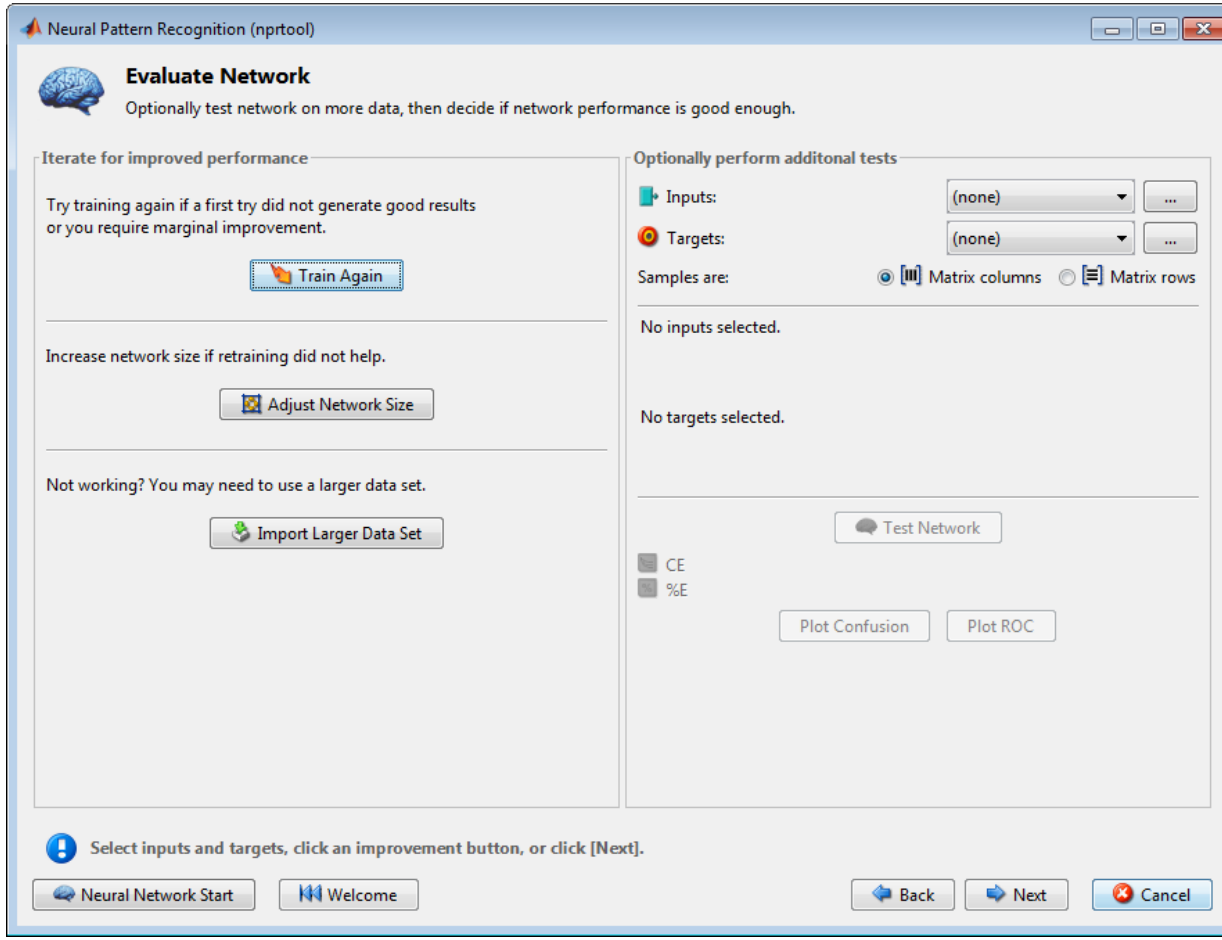
Confusion (plotconfusion)
File Edit View Insert Tools Desktop Window Help
Training Confusion Matrix
312 63.8%
1 0.2%
99.7% 0.3%
8 1.6%
168 34.4%
95.5% 4.5%
97.5% 2.5%
99.4% 0.6%
98.2% 1.8%
Validation Confusion Matrix
66 62.9%
3 2.9%
95.7% 4.3%
3 2.9%
33 31.4%
91.7% 8.3%
95.7% 4.3%
91.7% 8.3%
94.3% 5.7%
Test Confusion Matrix
66 62.9%
3 2.9%
95.7% 4.3%
3 2.9%
33 31.4%
91.7% 8.3%
95.7% 4.3%
91.7% 8.3%
94.3% 5.7%
All Confusion Matrix
444 63.5%
7 1.0%
98.4% 1.6%
14 2.0%
234 33.5%
94.4% 5.6%
96.9% 3.1%
97.1% 2.9%
97.0% 3.0%
Output Class
Target Class
1
2

11 Plot the Receiver Operating Characteristic (ROC) curve. Under the **Plots** pane, click **Receiver Operating Characteristic** in the Neural Network Pattern Recognition App.


Receiver Operating Characteristic (plotroc)
File Edit View Insert Tools Desktop Window Help
Training ROC
Class 1
Class 2
Validation ROC
Test ROC
All ROC
True Positive Rate
False Positive Rate
0 0.2 0.4 0.6 0.8 1

The colored lines in each axis represent the ROC curves. The *ROC curve* is a plot of the true positive rate (sensitivity) versus the false positive rate (1 - specificity) as the threshold is varied. A perfect test would show points in the upper-left corner, with 100% sensitivity and 100% specificity. For this problem, the network performs very well.

12 In the Neural Network Pattern Recognition App, click **Next** to evaluate the network.

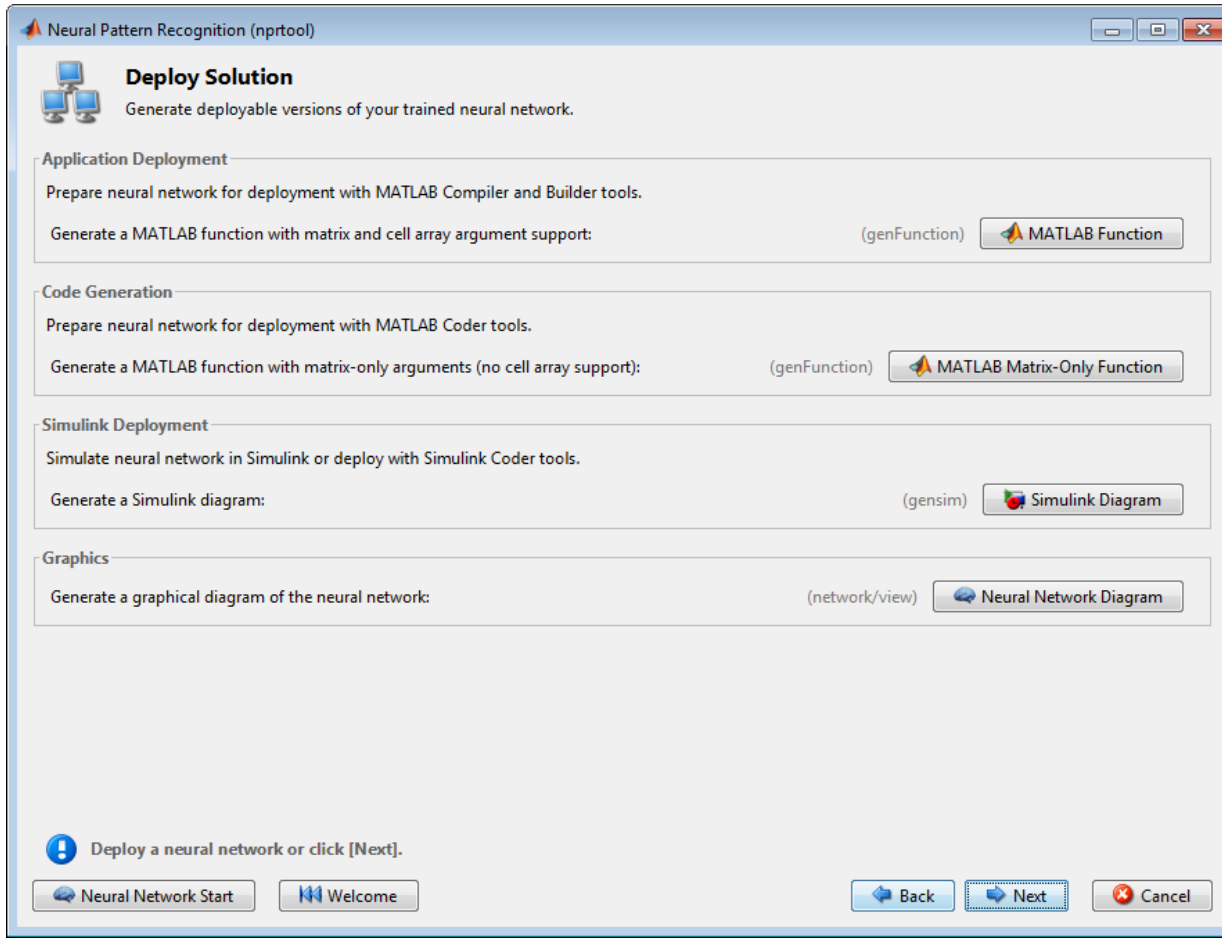


At this point, you can test the network against new data.

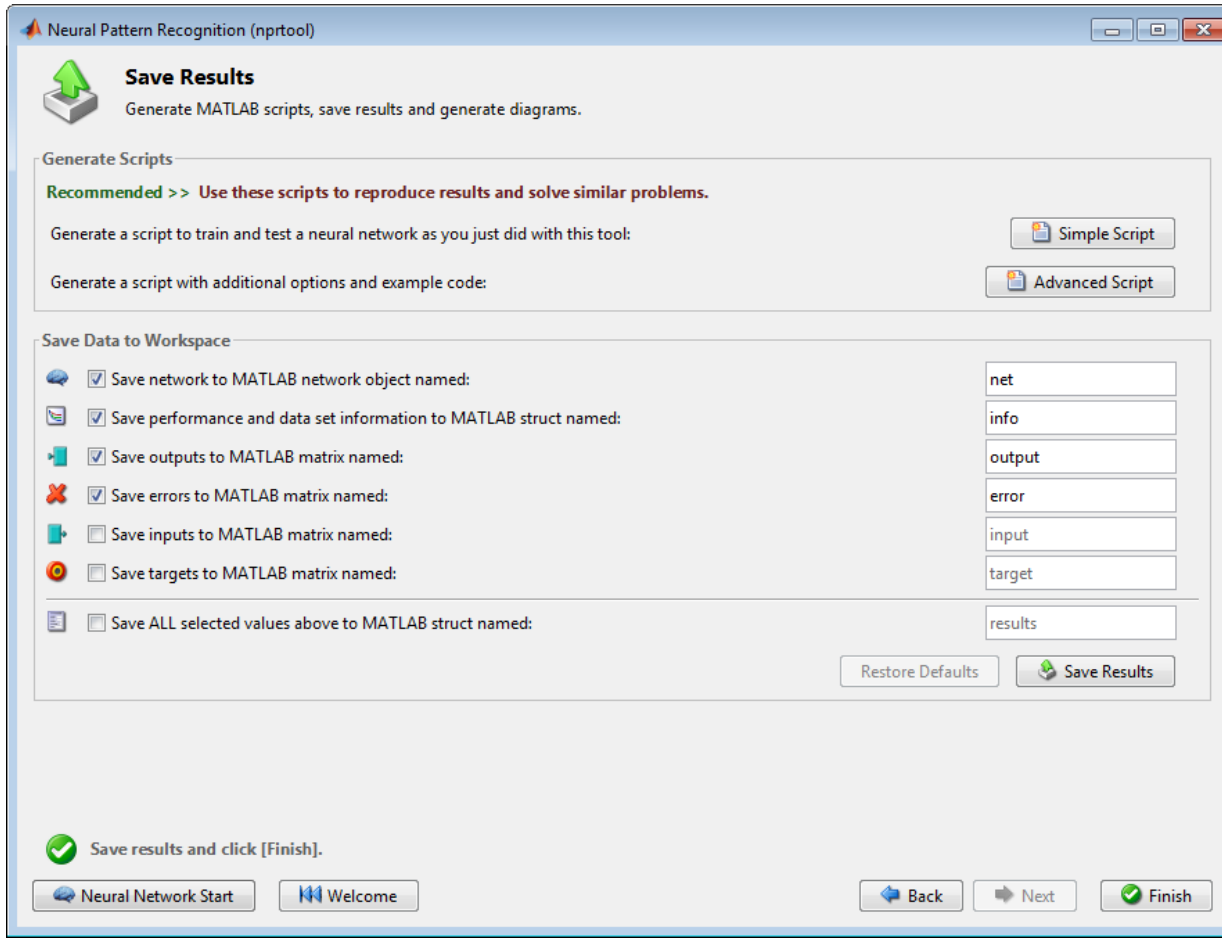
If you are dissatisfied with the network's performance on the original or new data, you can train it again, increase the number of neurons, or perhaps get a larger training data set. If the performance on the training set is good, but the test set performance is significantly worse, which could indicate overfitting, then reducing the number of neurons can improve your results.

13 When you are satisfied with the network performance, click **Next**.

Use this panel to generate a MATLAB function or Simulink diagram for simulating your neural network. You can use the generated code or diagram to better understand how your neural network computes outputs from inputs or deploy the network with MATLAB Compiler tools and other MATLAB code generation tools.



14 Click **Next**. Use the buttons on this screen to save your results.



- You can click **Simple Script** or **Advanced Script** to create MATLAB code that can be used to reproduce all of the previous steps from the command line. Creating MATLAB code can be helpful if you want to learn how to use the command-line functionality of the toolbox to customize the training process. In “Using Command-Line Functions” on page 1-79, you will investigate the generated scripts in more detail.
- You can also save the network as `net` in the workspace. You can perform additional tests on it or put it to work on new inputs.

15 When you have saved your results, click **Finish**.

Using Command-Line Functions

The easiest way to learn how to use the command-line functionality of the toolbox is to generate scripts from the GUIs, and then modify them to customize the network training. For example, look at the simple script that was created at step 14 of the previous section.

```
% Solve a Pattern Recognition Problem with a Neural Network
% Script generated by NPRTOOL
%
% This script assumes these variables are defined:
%
%   cancerInputs - input data.
%   cancerTargets - target data.

inputs = cancerInputs;
targets = cancerTargets;

% Create a Pattern Recognition Network
hiddenLayerSize = 10;
net = patternnet(hiddenLayerSize);


% Set up Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;


% Train the Network
[net,tr] = train(net,inputs,targets);

% Test the Network
outputs = net(inputs);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)

% View the Network
view(net)

% Plots
% Uncomment these lines to enable various plots.
% figure, plotperform(tr)
% figure, plottrainstate(tr)
```

```
% figure, plotconfusion(targets,outputs)
% figure, ploterrhist(errors)
```

You can save the script, and then run it from the command line to reproduce the results of the previous GUI session. You can also edit the script to customize the training process. In this case, follow each step in the script.

1. The script assumes that the input vectors and target vectors are already loaded into the workspace. If the data are not loaded, you can load them as follows:

   ```
   [inputs,targets] = cancer_dataset;
   ```

2. Create the network. The default network for function fitting (or regression) problems, `patternnet`, is a feedforward network with the default tan-sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer. You assigned ten neurons (somewhat arbitrary) to the one hidden layer in the previous section.

   - The network has two output neurons, because there are two target values (categories) associated with each input vector.
   - Each output neuron represents a category.
   - When an input vector of the appropriate category is applied to the network, the corresponding neuron should produce a 1, and the other neurons should output a 0.

   To create the network, enter these commands:

   ```
   hiddenLayerSize = 10;
   net = patternnet(hiddenLayerSize);
   ```

   Note The choice of network architecture for pattern recognition problems follows similar guidelines to function fitting problems. More neurons require more computation, and they have a tendency to overfit the data when the number is set too high, but they allow the network to solve more complicated problems. More layers require more computation, but their use might result in the network solving complex problems more efficiently. To use more than one hidden layer, enter the hidden layer sizes as elements of an array in the `patternnet` command.

3. Set up the division of data.

   ```
   net.divideParam.trainRatio = 70/100;
   net.divideParam.valRatio   = 15/100;
   net.divideParam.testRatio  = 15/100;
   ```

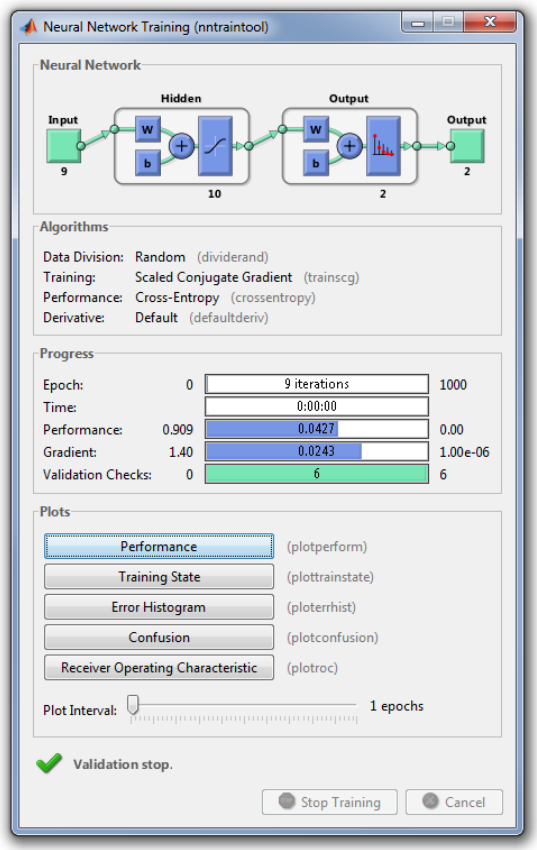
With these settings, the input vectors and target vectors will be randomly divided, with 70% used for training, 15% for validation and 15% for testing.

(See “Dividing the Data” for more discussion of the data division process.)

4 Train the network. The pattern recognition network uses the default Scaled Conjugate Gradient (`trainscg`) algorithm for training. To train the network, enter this command:

```
[net,tr] = train(net,inputs,targets);
```

During training, as in function fitting, the training window opens. This window displays training progress. To interrupt training at any point, click **Stop Training**.



This training stopped when the validation error increased for six iterations, which occurred at iteration 24.

5 Test the network. After the network has been trained, you can use it to compute the network outputs. The following code calculates the network outputs, errors and overall performance.

```
outputs = net(inputs);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)
```

```
performance =

    0.0419
```

It is also possible to calculate the network performance only on the test set, by using the testing indices, which are located in the training record.

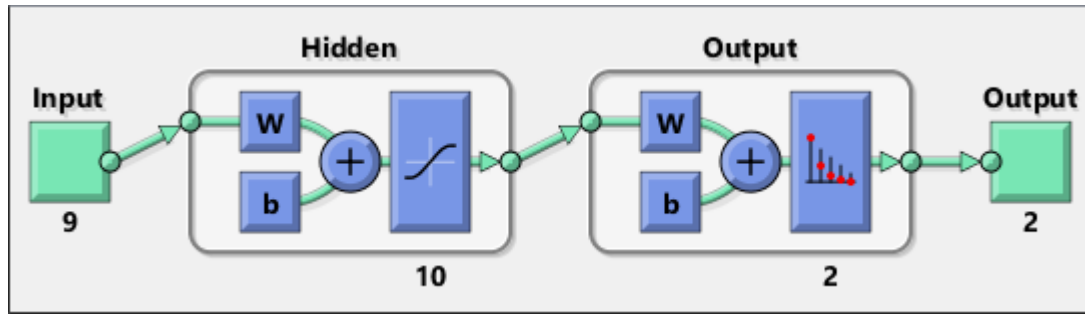
```
tInd = tr.testInd;
tstOutputs = net(inputs(:,tInd));
tstPerform = perform(net,targets(:,tInd),tstOutputs)
```

```
tstPerform =

    0.0263
```

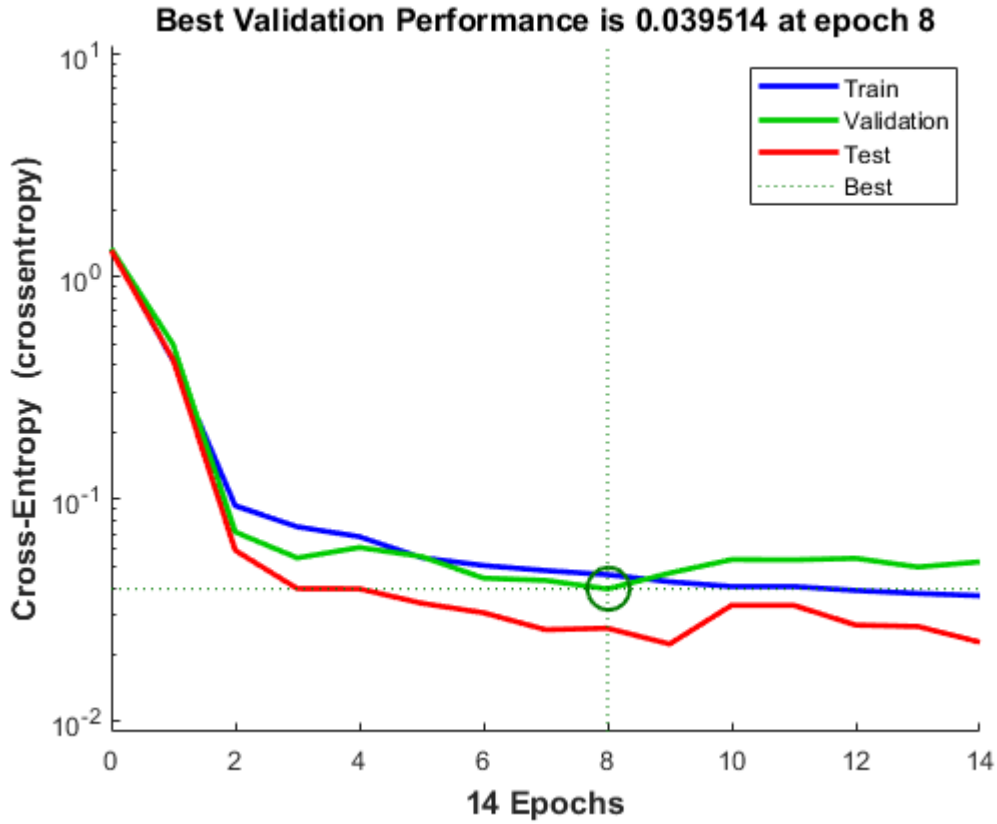
6 View the network diagram.

```
view(net)
```



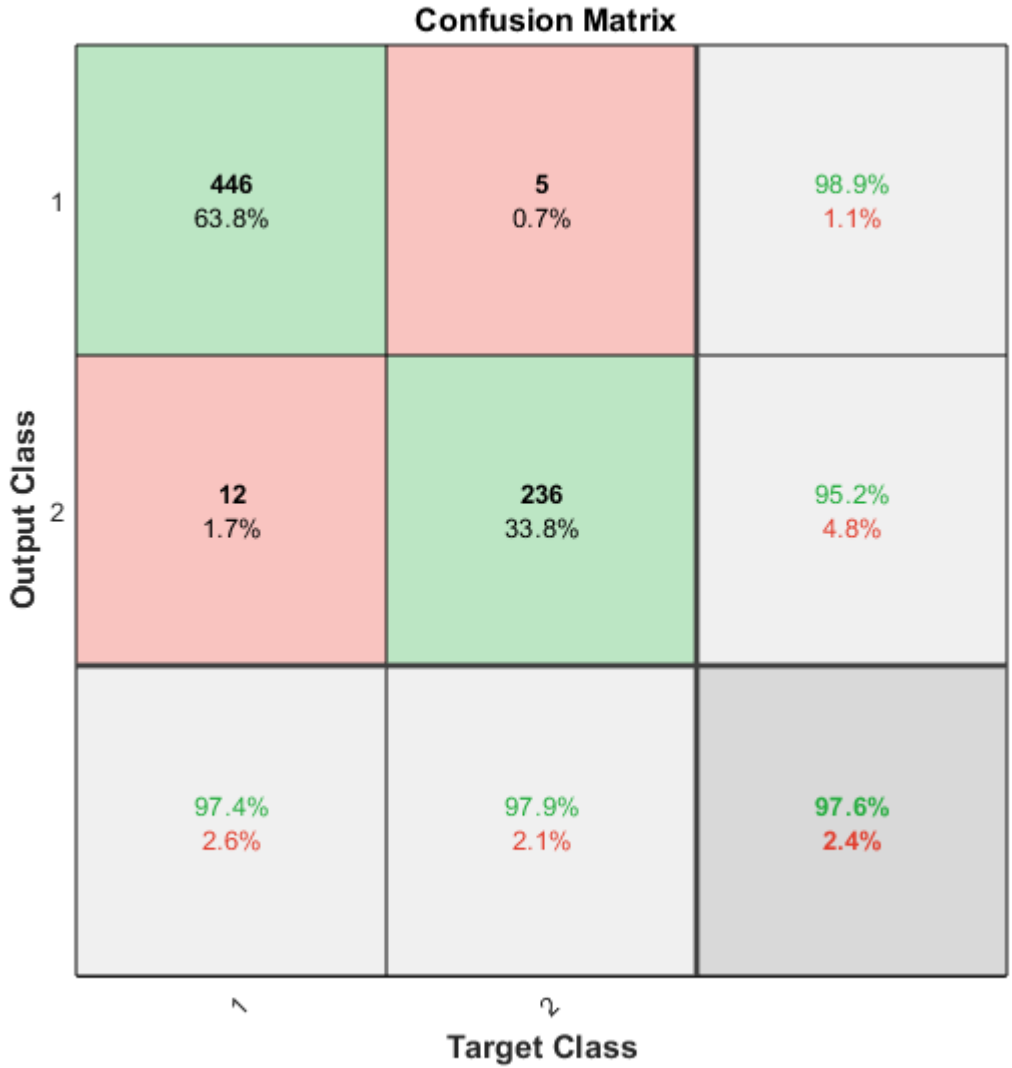
7 Plot the training, validation, and test performance.

```
figure, plotperform(tr)
```



8 Use the `plotconfusion` function to plot the confusion matrix. It shows the various types of errors that occurred for the final trained network.

```
figure, plotconfusion(targets,outputs)
```


Confusion Matrix
446
63.8%
5
0.7%
98.9%
1.1%
12
1.7%
236
33.8%
95.2%
4.8%
97.4%
2.6%
97.9%
2.1%
97.6%
2.4%
Output Class
1
2
1
2
Target Class

The diagonal cells show the number of cases that were correctly classified, and the off-diagonal cells show the misclassified cases. The blue cell in the bottom right shows the total percent of correctly classified cases (in green) and the total percent of misclassified cases (in red). The results show very good recognition. If you needed even more accurate results, you could try any of the following approaches:

- Reset the initial network weights and biases to new values with `init` and train again.
- Increase the number of hidden neurons.
- Increase the number of training vectors.
- Increase the number of input values, if more relevant information is available.
- Try a different training algorithm (see “Training Algorithms”).

In this case, the network response is satisfactory, and you can now put the network to use on new inputs.

To get more experience in command-line operations, here are some tasks you can try:

- During training, open a plot window (such as the confusion plot), and watch it animate.
- Plot from the command line with functions such as `plotroc` and `plottrainstate`.

Also, see the advanced script for more options, when training from the command line.

Each time a neural network is trained, can result in a different solution due to different initial weight and bias values and different divisions of data into training, validation, and test sets. As a result, different neural networks trained on the same problem can give different outputs for the same input. To ensure that a neural network of good accuracy has been found, retrain several times.

There are several other techniques for improving upon initial solutions if higher accuracy is desired. For more information, see “Improve Shallow Neural Network Generalization and Avoid Overfitting”.

Cluster Data with a Self-Organizing Map

Clustering data is another excellent application for neural networks. This process involves grouping data by similarity. For example, you might perform:

- Market segmentation by grouping people according to their buying patterns
- Data mining by partitioning data into related subsets
- Bioinformatic analysis by grouping genes with related expression patterns

Suppose that you want to cluster flower types according to petal length, petal width, sepal length, and sepal width. You have 150 example cases for which you have these four measurements.

As with function fitting and pattern recognition, there are two ways to solve this problem:

- Use the `nctool` GUI, as described in “Using the Neural Network Clustering App” on page 1-86.
- Use a command-line solution, as described in “Using Command-Line Functions” on page 1-99.

Defining a Problem

To define a clustering problem, simply arrange Q input vectors to be clustered as columns in an input matrix (see “Data Structures” for a detailed description of data formatting for static and time series data). For instance, you might want to cluster this set of 10 two-element vectors:

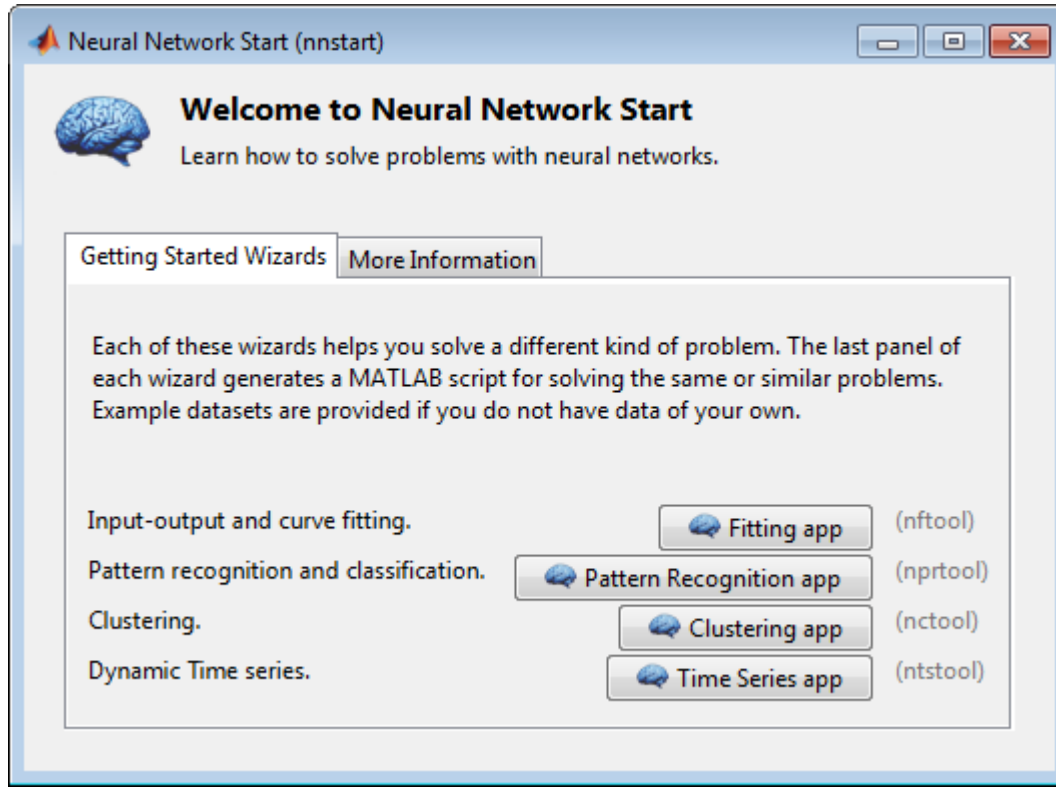
```
inputs = [7 0 6 2 6 5 6 1 0 1; 6 2 5 0 7 5 5 1 2 2]
```

The next section shows how to train a network using the `nctool` GUI.

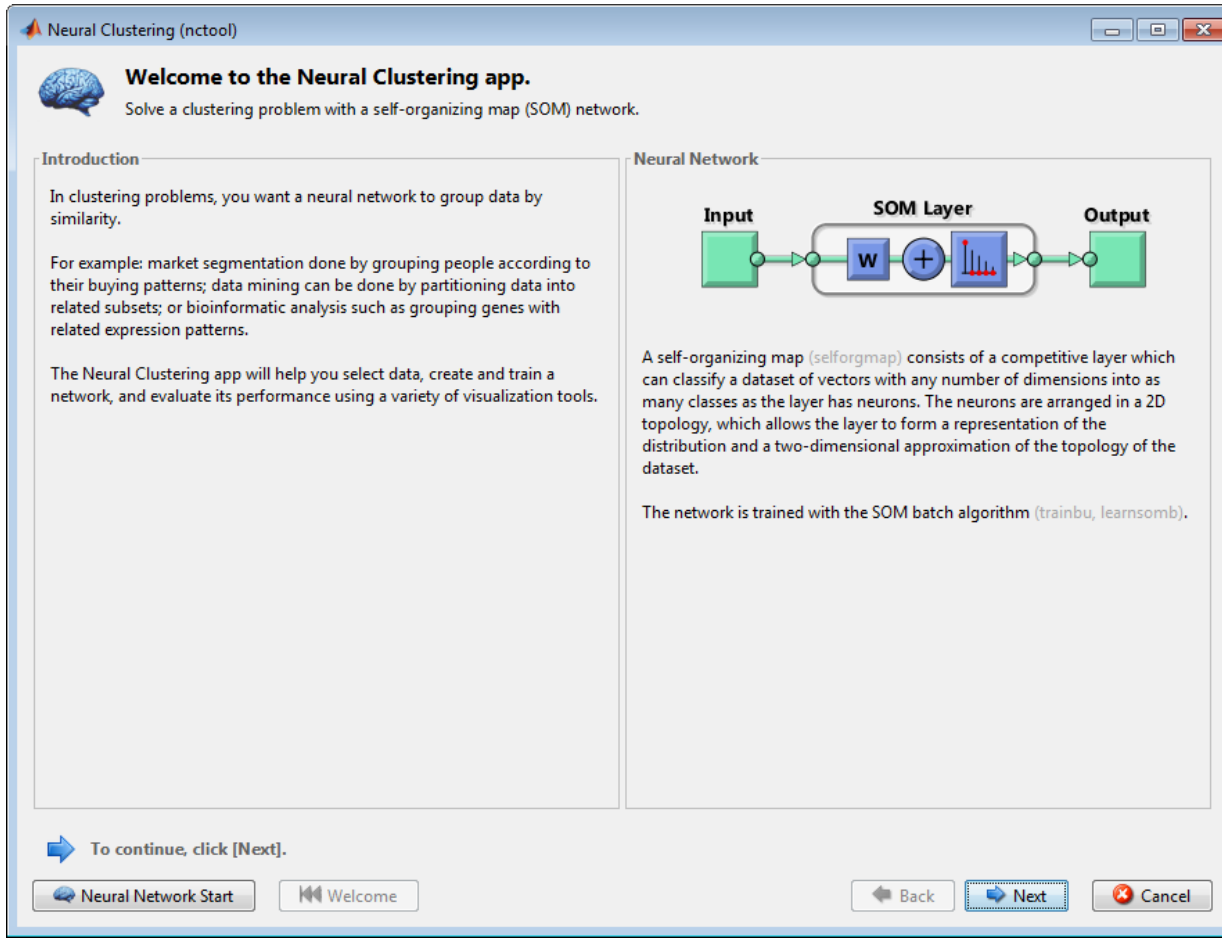
Using the Neural Network Clustering App

1 If needed, open the Neural Network Start GUI with this command:

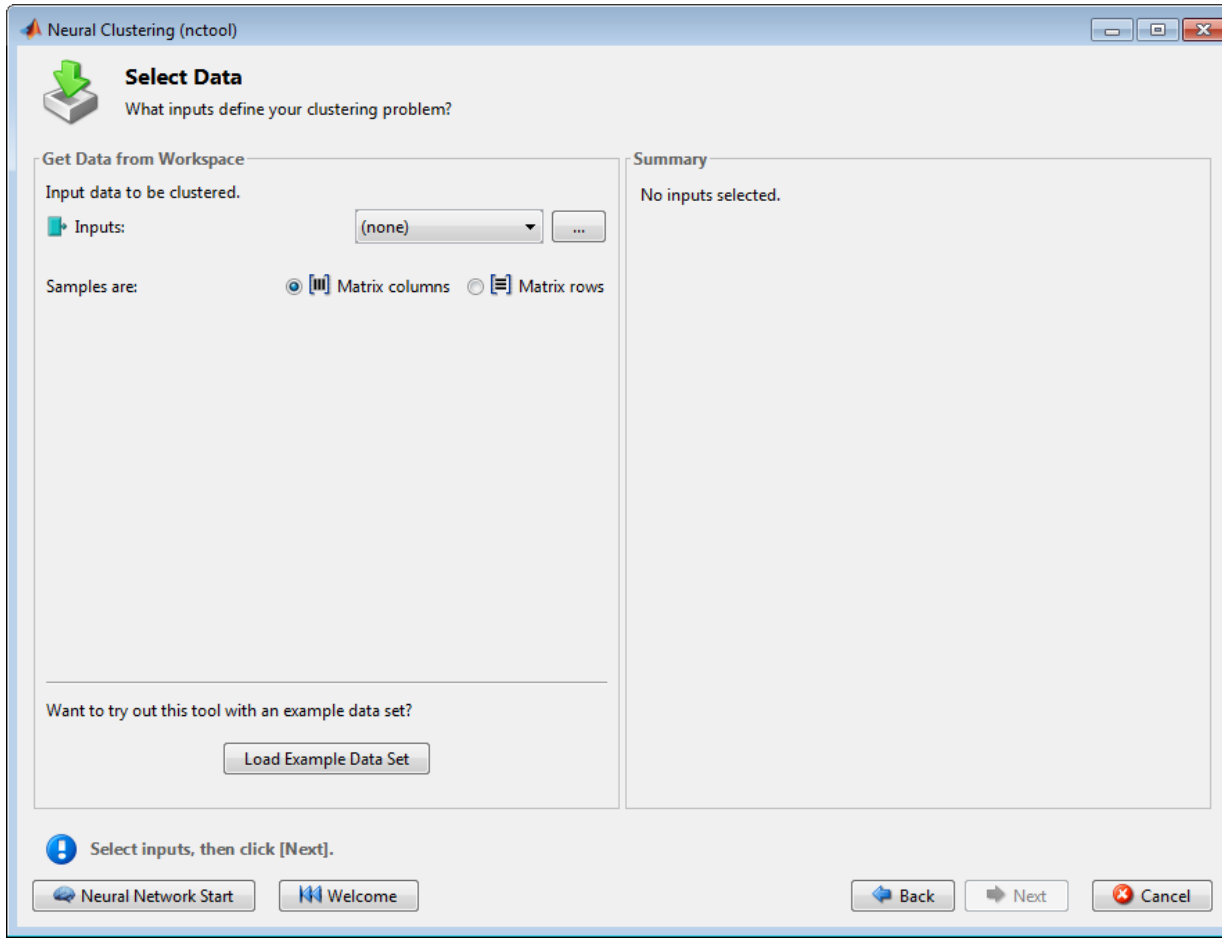
```
nnstart
```



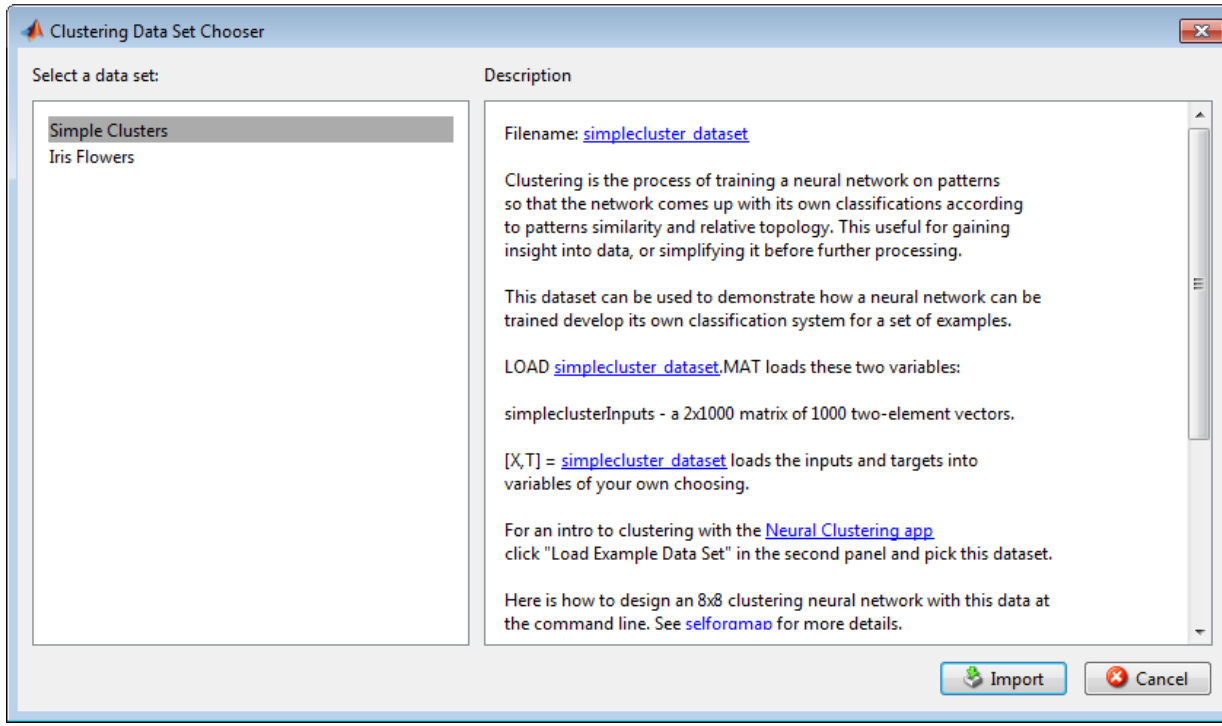
2 Click **Clustering app** to open the Neural Network Clustering App. (You can also use the command `nctool`.)



3 Click **Next**. The Select Data window appears.



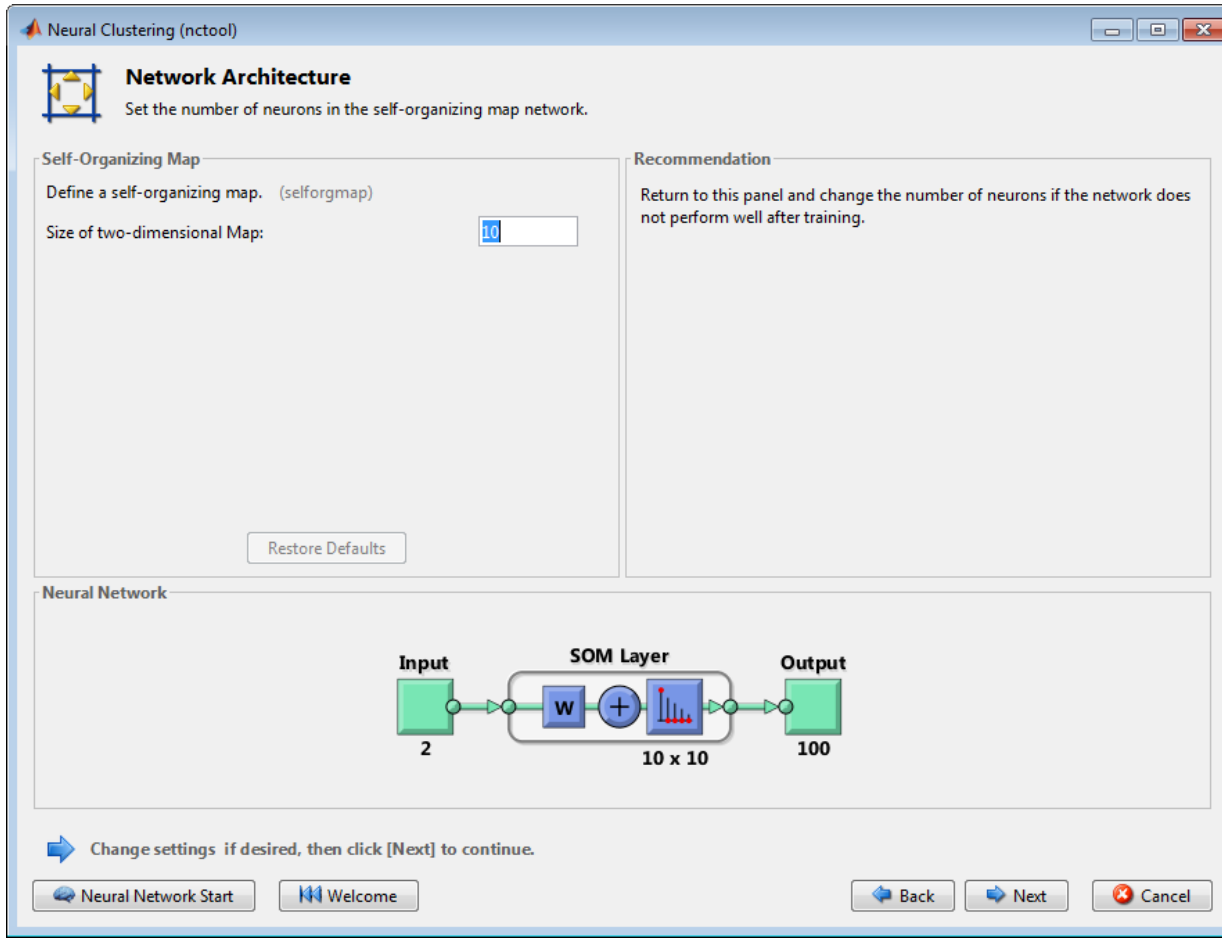
4 Click **Load Example Data Set**. The Clustering Data Set Chooser window appears.



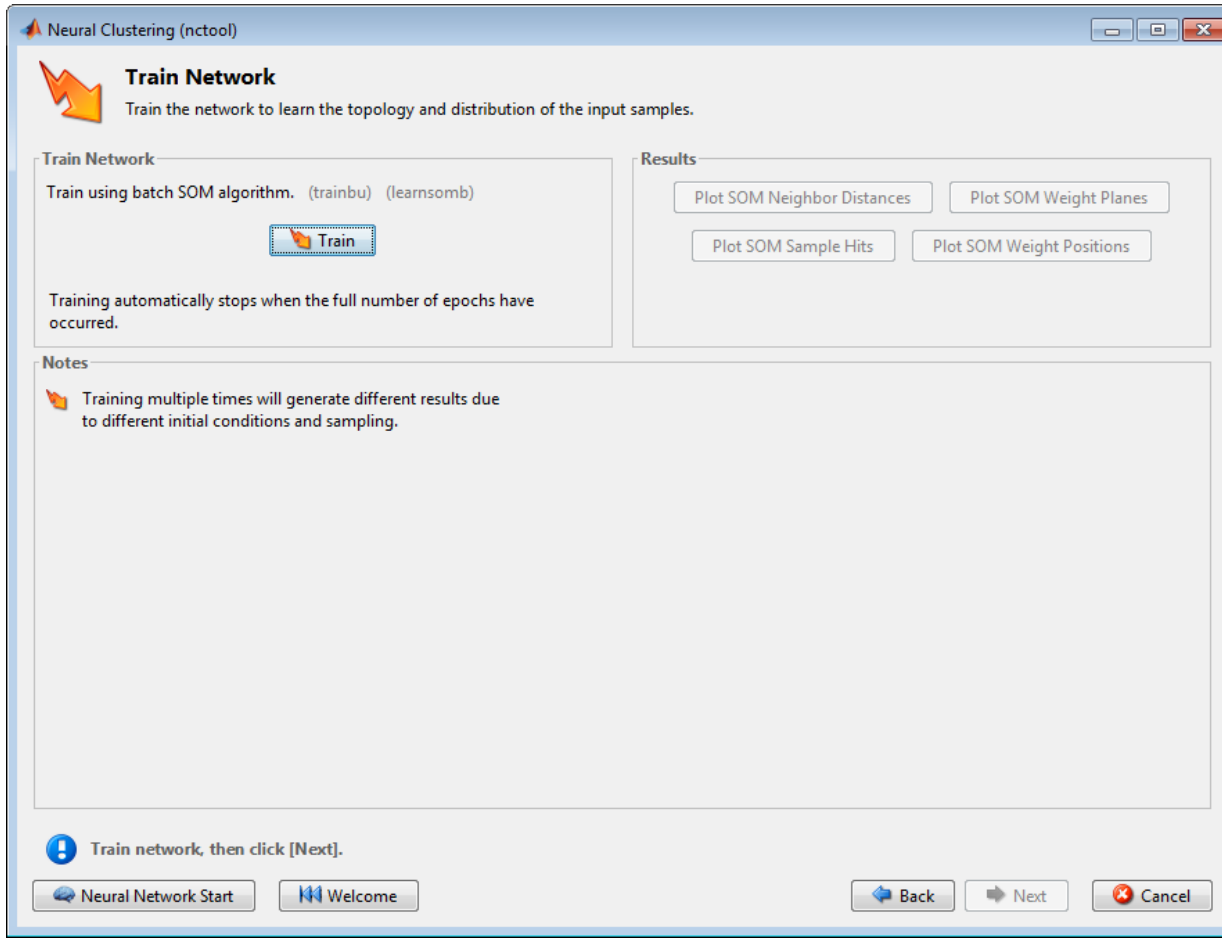
5 In this window, select **Simple Clusters**, and click **Import**. You return to the Select Data window.

6 Click **Next** to continue to the Network Size window, shown in the following figure.

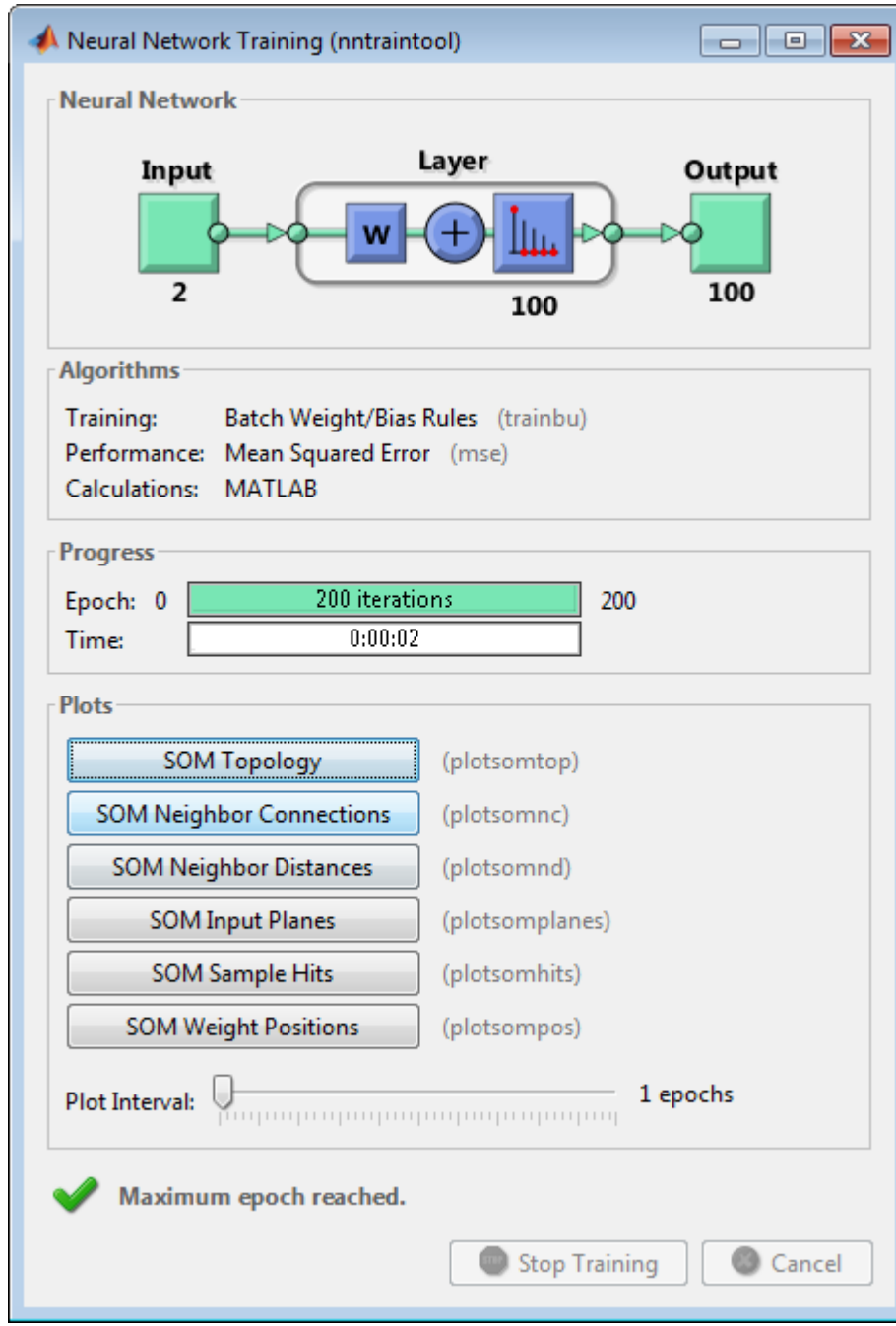
For clustering problems, the self-organizing feature map (SOM) is the most commonly used network, because after the network has been trained, there are many visualization tools that can be used to analyze the resulting clusters. This network has one layer, with neurons organized in a grid. (For more information on the SOM, see “Self-Organizing Feature Maps”.) When creating the network, you specify the numbers of rows and columns in the grid. Here, the number of rows and columns is set to `10`. The total number of neurons is 100. You can change this number in another run if you want.



7 Click **Next**. The Train Network window appears.

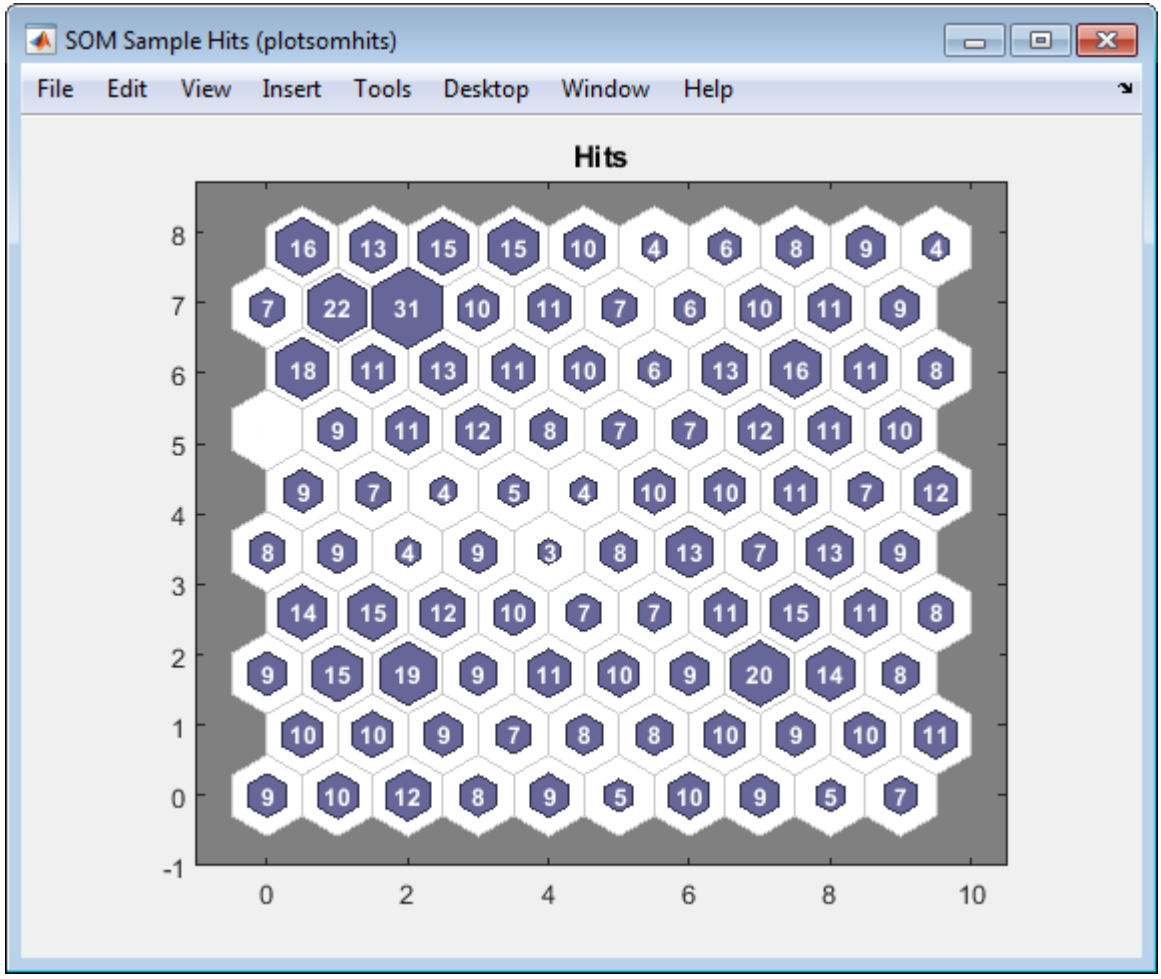


8 Click **Train**.


Neural Network Training (nntraintool)
Neural Network
Input
Layer
Output
W
2
100
100
Algorithms
Training: Batch Weight/Bias Rules (trainbu)
Performance: Mean Squared Error (mse)
Calculations: MATLAB
Progress
Epoch: 0
200 iterations
200
Time:
0:00:02
Plots
SOM Topology (plotsomtop)
SOM Neighbor Connections (plotsomnc)
SOM Neighbor Distances (plotsomnd)
SOM Input Planes (plotsomplanes)
SOM Sample Hits (plotsomhits)
SOM Weight Positions (plotsompos)
Plot Interval:
1 epochs
Maximum epoch reached.
Stop Training
Cancel

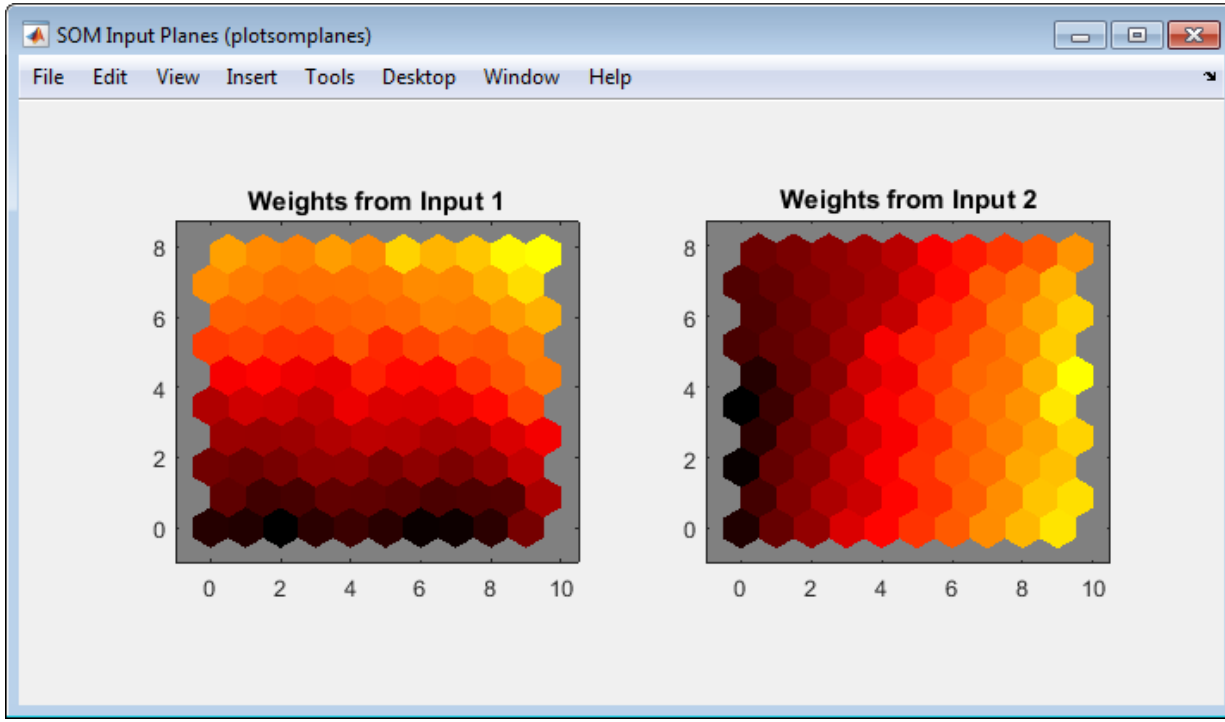
The training runs for the maximum number of epochs, which is 200.

9 For SOM training, the weight vector associated with each neuron moves to become the center of a cluster of input vectors. In addition, neurons that are adjacent to each other in the topology should also move close to each other in the input space, therefore it is possible to visualize a high-dimensional inputs space in the two dimensions of the network topology. Investigate some of the visualization tools for the SOM. Under the **Plots** pane, click **SOM Sample Hits**.



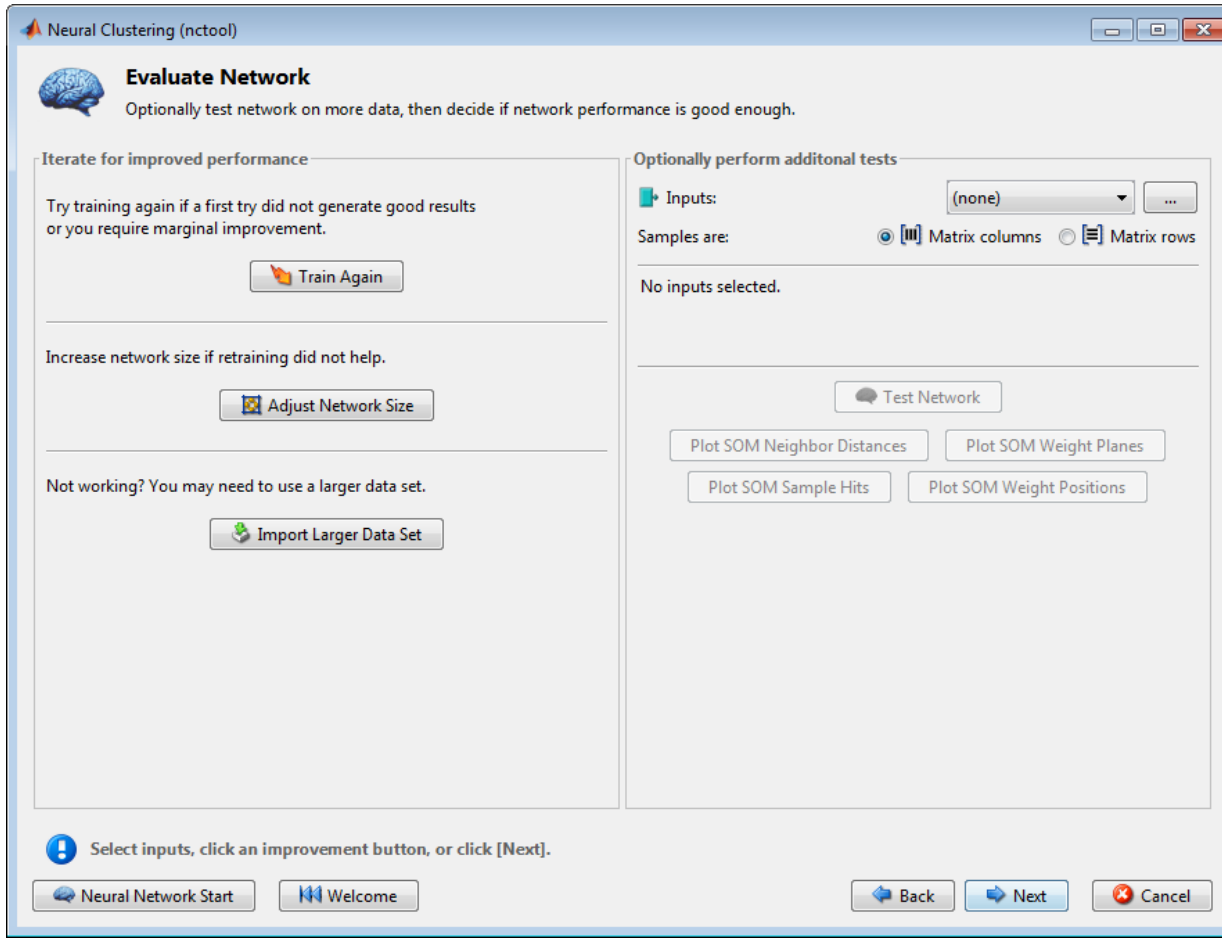
The default topology of the SOM is hexagonal. This figure shows the neuron locations in the topology, and indicates how many of the training data are associated with each of the neurons (cluster centers). The topology is a 10-by-10 grid, so there are 100 neurons. The maximum number of hits associated with any neuron is 31. Thus, there are 31 input vectors in that cluster.

10 You can also visualize the SOM by displaying weight planes (also referred to as *component planes*). Click **SOM Weight Planes** in the Neural Network Clustering App.



This figure shows a weight plane for each element of the input vector (two, in this case). They are visualizations of the weights that connect each input to each of the neurons. (Darker colors represent larger weights.) If the connection patterns of two inputs were very similar, you can assume that the inputs are highly correlated. In this case, input 1 has connections that are very different than those of input 2.

11 In the Neural Network Clustering App, click **Next** to evaluate the network.



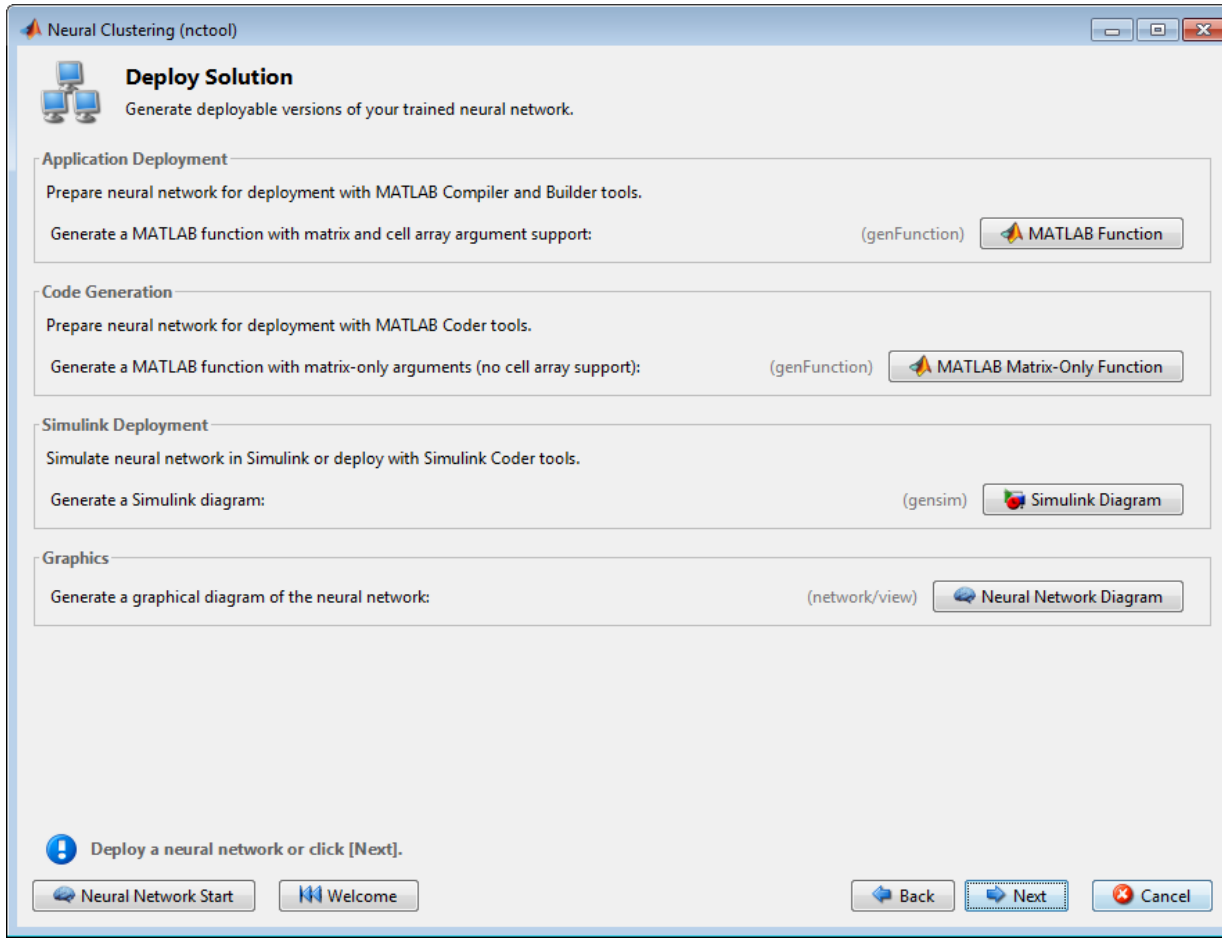
At this point you can test the network against new data.

If you are dissatisfied with the network's performance on the original or new data, you can increase the number of neurons, or perhaps get a larger training data set.

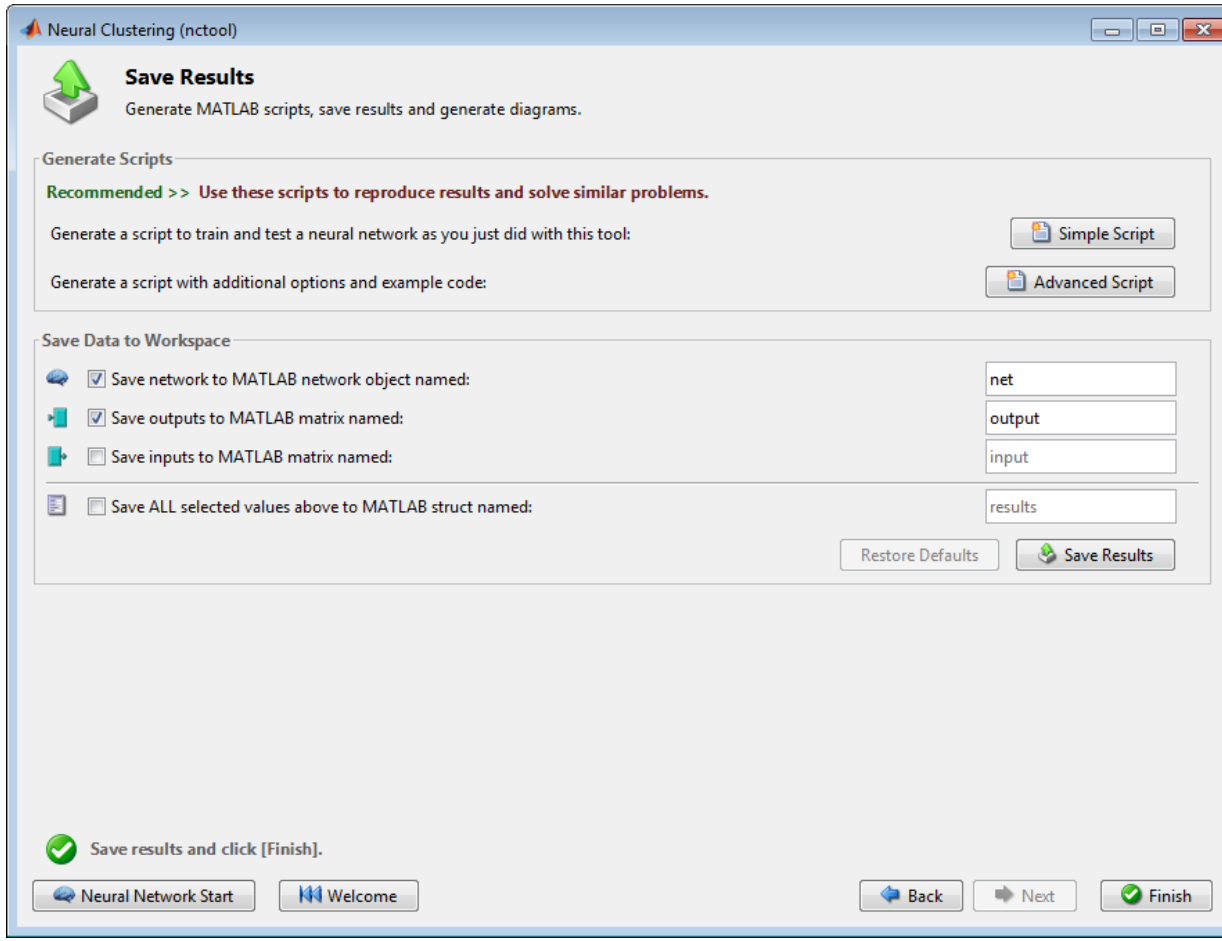
12 When you are satisfied with the network performance, click **Next**.

13 Use this panel to generate a MATLAB function or Simulink diagram for simulating your neural network. You can use the generated code or diagram to better understand how your neural network computes outputs from inputs or deploy the

network with MATLAB Compiler tools and other MATLAB and Simulink code generation tools.



14 Use the buttons on this screen to save your results.



- You can click **Simple Script** or **Advanced Script** to create MATLAB code that can be used to reproduce all of the previous steps from the command line. Creating MATLAB code can be helpful if you want to learn how to use the command-line functionality of the toolbox to customize the training process. In “Using Command-Line Functions” on page 1-99, you will investigate the generated scripts in more detail.
- You can also save the network as `net` in the workspace. You can perform additional tests on it or put it to work on new inputs.

15 When you have generated scripts and saved your results, click **Finish**.

Using Command-Line Functions

The easiest way to learn how to use the command-line functionality of the toolbox is to generate scripts from the GUIs, and then modify them to customize the network training. As an example, look at the simple script that was created in step 14 of the previous section.

```
% Solve a Clustering Problem with a Self-Organizing Map
% Script generated by NCTOOL
%
% This script assumes these variables are defined:
%
%   simpleclusterInputs - input data.

inputs = simpleclusterInputs;

% Create a Self-Organizing Map
dimension1 = 10;
dimension2 = 10;
net = selforgmap([dimension1 dimension2]);

% Train the Network
[net,tr] = train(net,inputs);

% Test the Network
outputs = net(inputs);

% View the Network
view(net)

% Plots
% Uncomment these lines to enable various plots.
% figure, plotsomtop(net)
% figure, plotsomnc(net)
% figure, plotsomnd(net)
% figure, plotsomplanes(net)
% figure, plotsomhits(net,inputs)
% figure, plotsompos(net,inputs)
```

You can save the script, and then run it from the command line to reproduce the results of the previous GUI session. You can also edit the script to customize the training process. In this case, let's follow each of the steps in the script.

1. The script assumes that the input vectors are already loaded into the workspace. To show the command-line operations, you can use a different data set than you used for the GUI operation. Use the flower data set as an example. The iris data set consists of 150 four-element input vectors.

   ```
   load iris_dataset
   inputs = irisInputs;
   ```

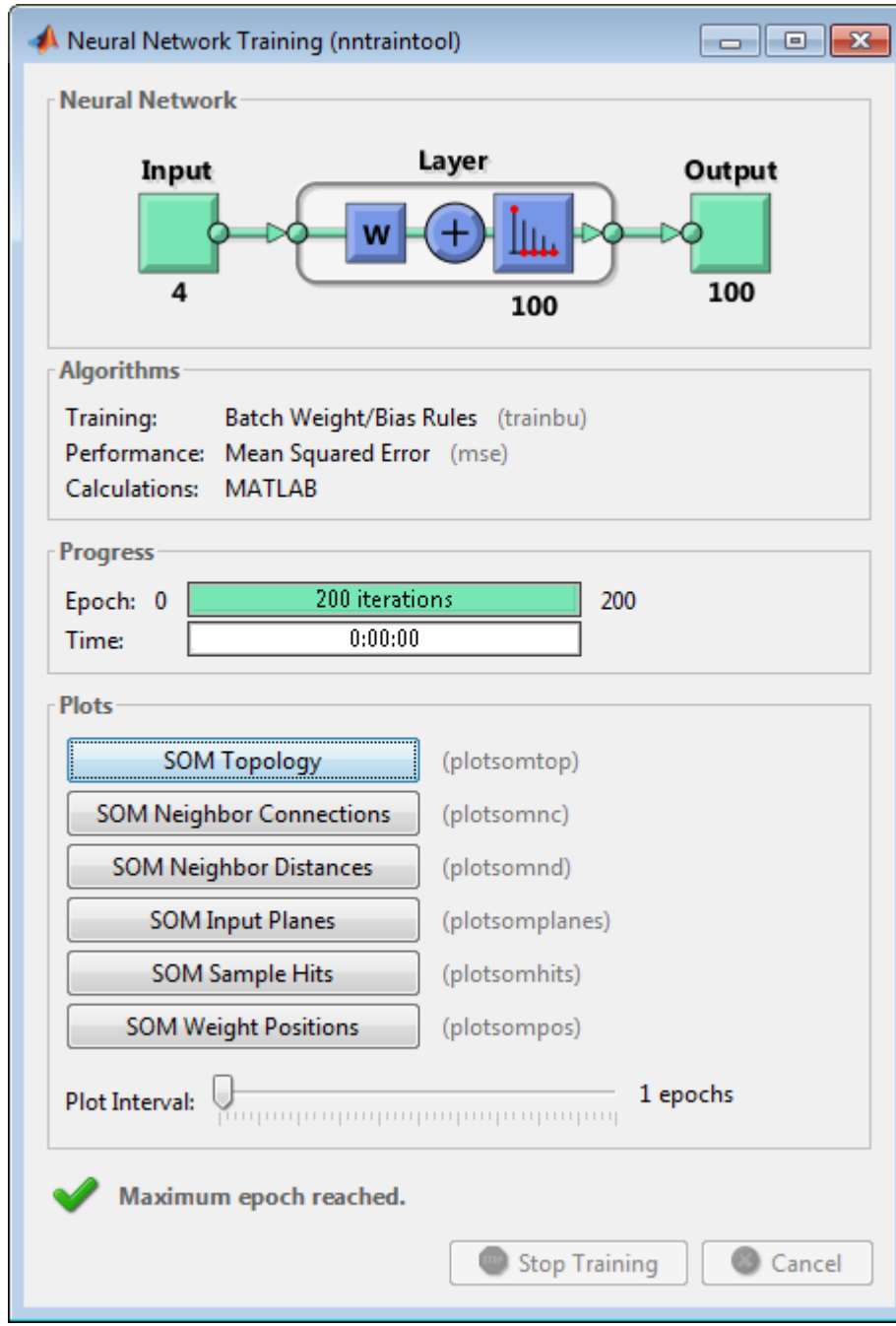
2. Create a network. For this example, you use a self-organizing map (SOM). This network has one layer, with the neurons organized in a grid. (For more information, see “Self-Organizing Feature Maps”.) When creating the network with `selforgmap`, you specify the number of rows and columns in the grid:

   ```
   dimension1 = 10;
   dimension2 = 10;
   net = selforgmap([dimension1 dimension2]);
   ```

3. Train the network. The SOM network uses the default batch SOM algorithm for training.

   ```
   [net,tr] = train(net,inputs);
   ```

4. During training, the training window opens and displays the training progress. To interrupt training at any point, click **Stop Training**.

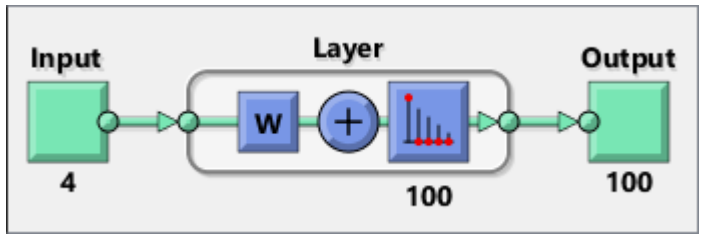

Neural Network Training (nntraintool)
Neural Network
Input
Layer
Output
4
100
100
Algorithms
Training: Batch Weight/Bias Rules (trainbu)
Performance: Mean Squared Error (mse)
Calculations: MATLAB
Progress
Epoch: 0
200 iterations
200
Time:
0:00:00
Plots
SOM Topology (plotsomtop)
SOM Neighbor Connections (plotsomnc)
SOM Neighbor Distances (plotsomnd)
SOM Input Planes (plotsomplanes)
SOM Sample Hits (plotsomhits)
SOM Weight Positions (plotsompos)
Plot Interval:
1 epochs
Maximum epoch reached.
Stop Training
Cancel

5 Test the network. After the network has been trained, you can use it to compute the network outputs.

```
outputs = net(inputs);
```

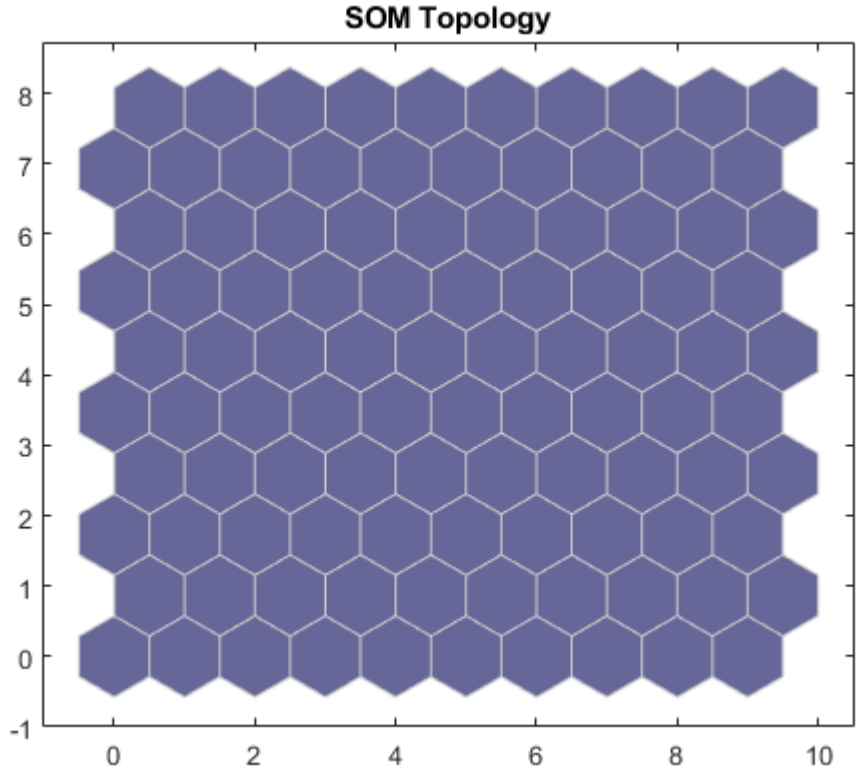
6 View the network diagram.

```
view(net)
```



7 For SOM training, the weight vector associated with each neuron moves to become the center of a cluster of input vectors. In addition, neurons that are adjacent to each other in the topology should also move close to each other in the input space, therefore it is possible to visualize a high-dimensional inputs space in the two dimensions of the network topology. The default SOM topology is hexagonal; to view it, enter the following commands.

```
figure, plotsomtop(net)
```



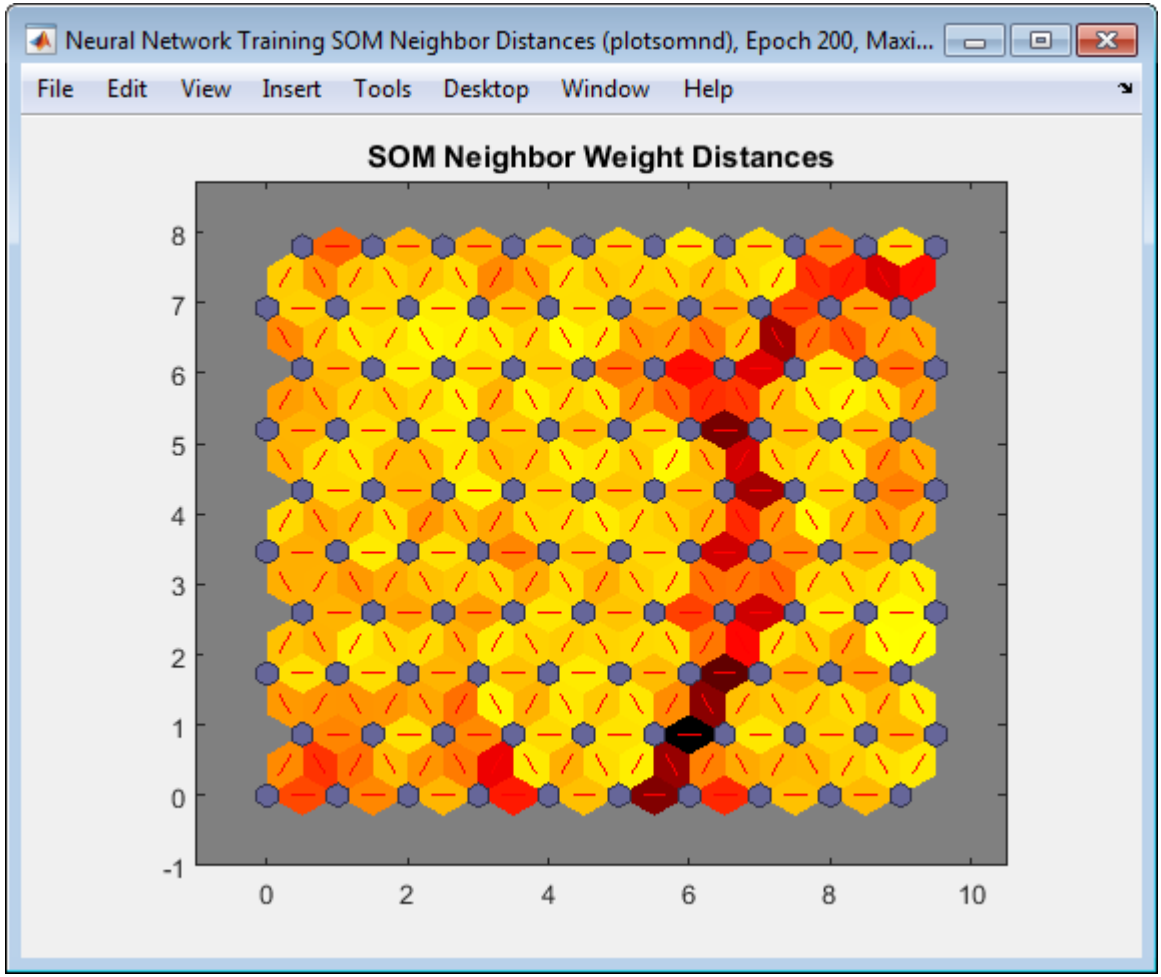
In this figure, each of the hexagons represents a neuron. The grid is 10-by-10, so there are a total of 100 neurons in this network. There are four elements in each input vector, so the input space is four-dimensional. The weight vectors (cluster centers) fall within this space.

Because this SOM has a two-dimensional topology, you can visualize in two dimensions the relationships among the four-dimensional cluster centers. One visualization tool for the SOM is the *weight distance matrix* (also called the *U-matrix*).

8 To view the U-matrix, click **SOM Neighbor Distances** in the training window.

In this figure, the blue hexagons represent the neurons. The red lines connect neighboring neurons. The colors in the regions containing the red lines indicate the distances between neurons. The darker colors represent larger distances, and the

lighter colors represent smaller distances. A band of dark segments crosses from the lower-center region to the upper-right region. The SOM network appears to have clustered the flowers into two distinct groups.



To get more experience in command-line operations, try some of these tasks:

- During training, open a plot window (such as the SOM weight position plot) and watch it animate.

- Plot from the command line with functions such as `plotsomhits`, `plotsomnc`, `plotsomnd`, `plotsomplanes`, `plotsompos`, and `plotsomtop`. (For more information on using these functions, see their reference pages.)

Also, see the advanced script for more options, when training from the command line.

Shallow Neural Network Time-Series Prediction and Modeling

Dynamic neural networks are good at time-series prediction. To see examples of using NARX networks being applied in open-loop form, closed-loop form and open/closed-loop multistep prediction see “Multistep Neural Network Prediction”.

Tip For deep learning with time series data, see instead “Sequence Classification Using Deep Learning”.

Suppose, for instance, that you have data from a pH neutralization process. You want to design a network that can predict the pH of a solution in a tank from past values of the pH and past values of the acid and base flow rate into the tank. You have a total of 2001 time steps for which you have those series.

You can solve this problem in two ways:

- Use a graphical user interface, `ntstool`, as described in “Using the Neural Network Time Series App” on page 1-107.
- Use command-line functions, as described in “Using Command-Line Functions” on page 1-123.

It is generally best to start with the GUI, and then to use the GUI to automatically generate command-line scripts. Before using either method, the first step is to define the problem by selecting a data set. Each GUI has access to many sample data sets that you can use to experiment with the toolbox. If you have a specific problem that you want to solve, you can load your own data into the workspace. The next section describes the data format.

Defining a Problem

To define a time series problem for the toolbox, arrange a set of TS input vectors as columns in a cell array. Then, arrange another set of TS target vectors (the correct output vectors for each of the input vectors) into a second cell array (see “Data Structures” for a detailed description of data formatting for static and time series data). However, there are cases in which you only need to have a target data set. For example, you can define the following time series problem, in which you want to use previous values of a series to predict the next value:

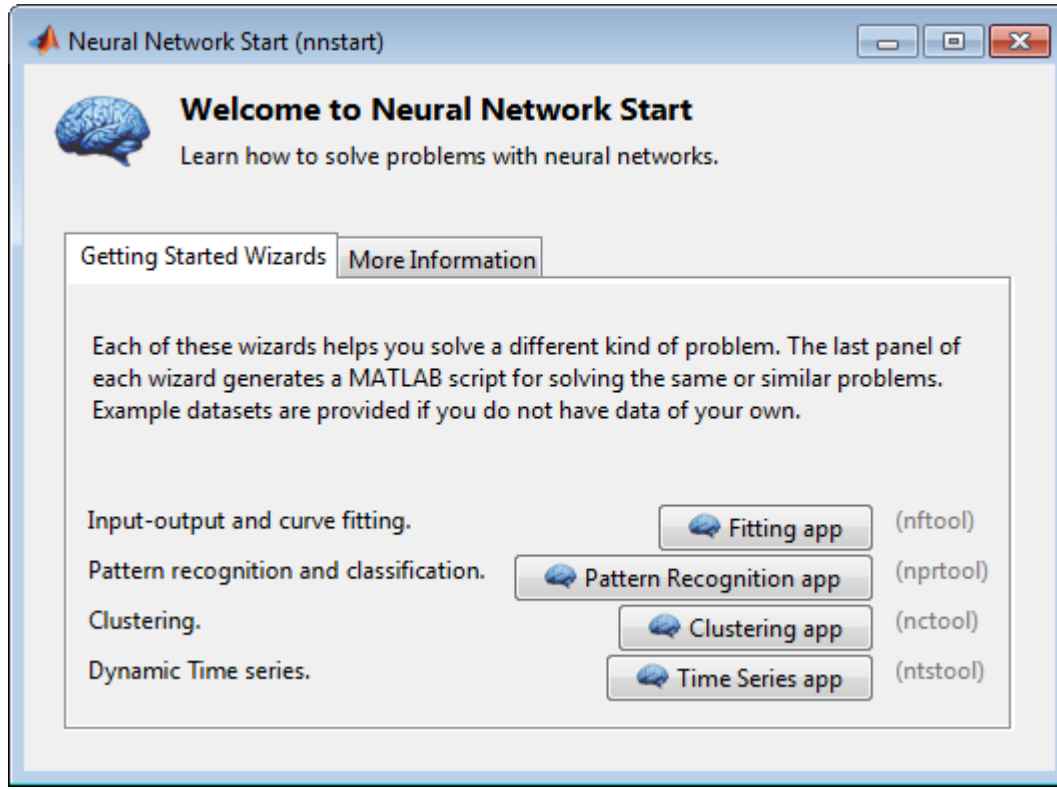
```
targets = {1 2 3 4 5};
```

The next section shows how to train a network to fit a time series data set, using the neural network time series app, `ntstool`. This example uses the pH neutralization data set provided with the toolbox.

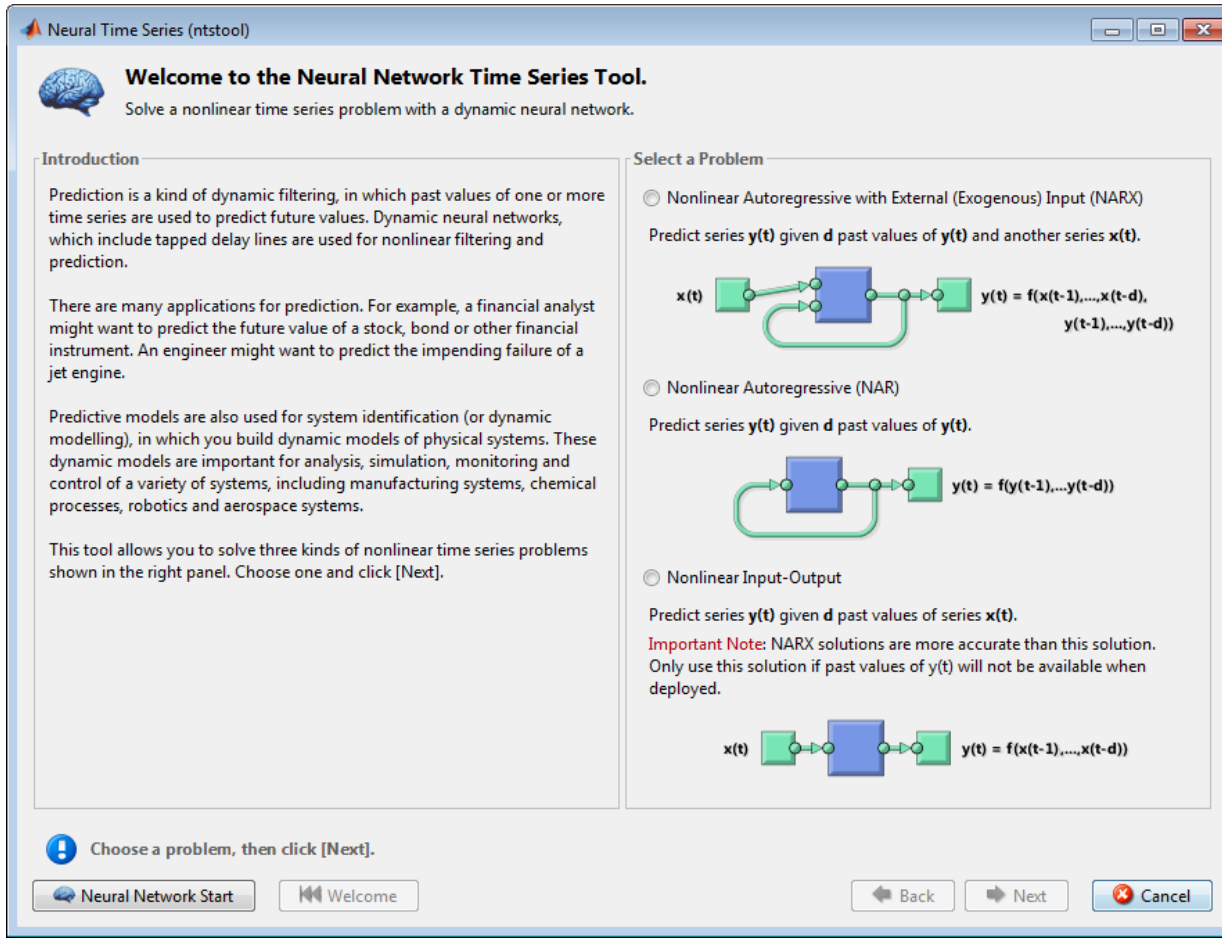
Using the Neural Network Time Series App

1 If needed, open the Neural Network Start GUI with this command:

```
nnstart
```



2 Click **Time Series App** to open the Neural Network Time Series App. (You can also use the command `ntstool`.)



Notice that this opening pane is different than the opening panes for the other GUIs. This is because `ntstool` can be used to solve three different kinds of time series problems.

- In the first type of time series problem, you would like to predict future values of a time series $y(t)$ from past values of that time series and past values of a second time series $x(t)$. This form of prediction is called nonlinear autoregressive with exogenous (external) input, or NARX (see “NARX Network” (narxnet, closeloop)), and can be written as follows:

$$y(t) = f(y(t-1), \ldots, y(t-d), x(t-1), \ldots, (t-d))$$

This model could be used to predict future values of a stock or bond, based on such economic variables as unemployment rates, GDP, etc. It could also be used for system identification, in which models are developed to represent dynamic systems, such as chemical processes, manufacturing systems, robotics, aerospace vehicles, etc.

- In the second type of time series problem, there is only one series involved. The future values of a time series $y(t)$ are predicted only from past values of that series. This form of prediction is called nonlinear autoregressive, or NAR, and can be written as follows:

  $$y(t) = f(y(t-1), \ldots, y(t-d))$$

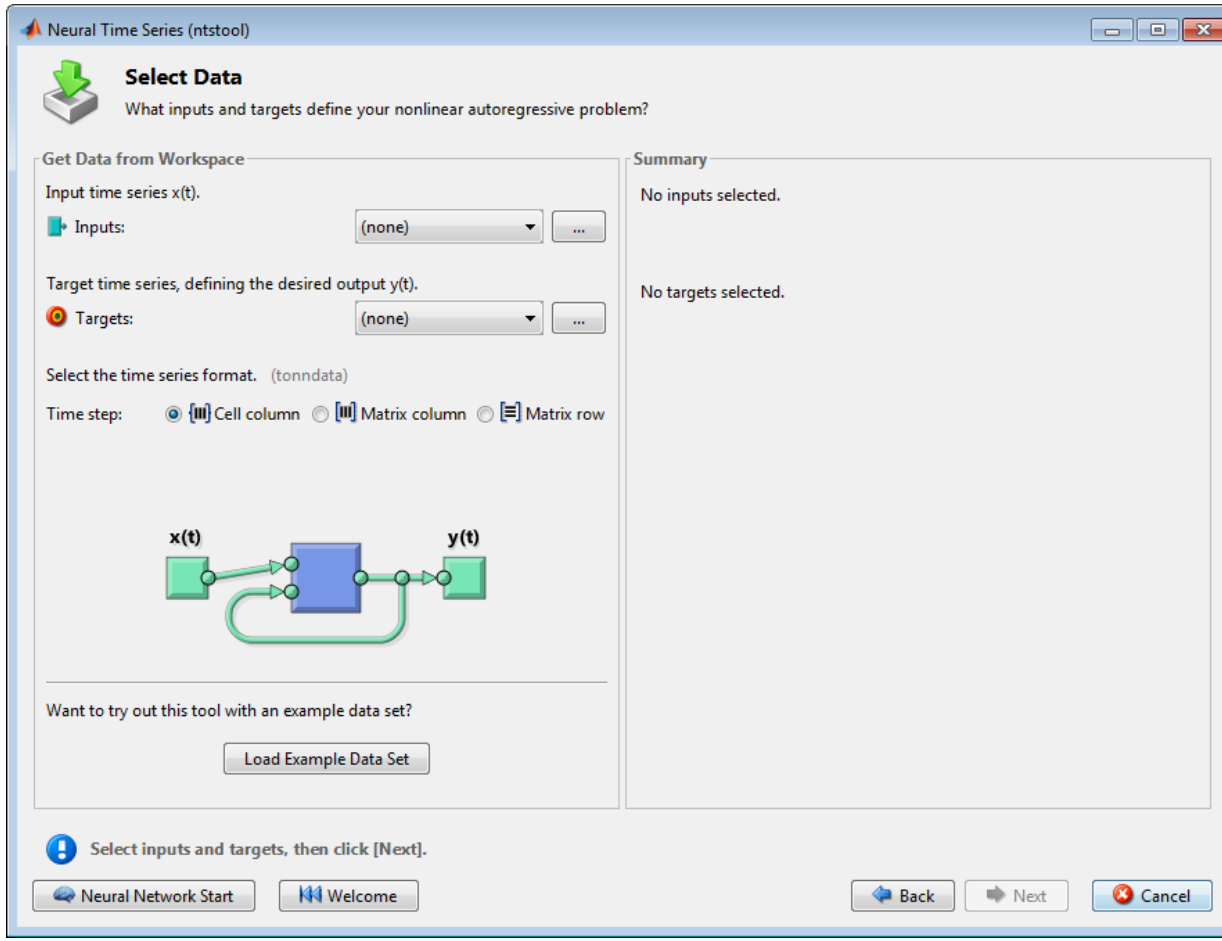
  This model could also be used to predict financial instruments, but without the use of a companion series.

- The third time series problem is similar to the first type, in that two series are involved, an input series $x(t)$ and an output/target series $y(t)$. Here you want to predict values of $y(t)$ from previous values of $x(t)$, but without knowledge of previous values of $y(t)$. This input/output model can be written as follows:

  $$y(t) = f(x(t-1), \ldots, x(t-d))$$

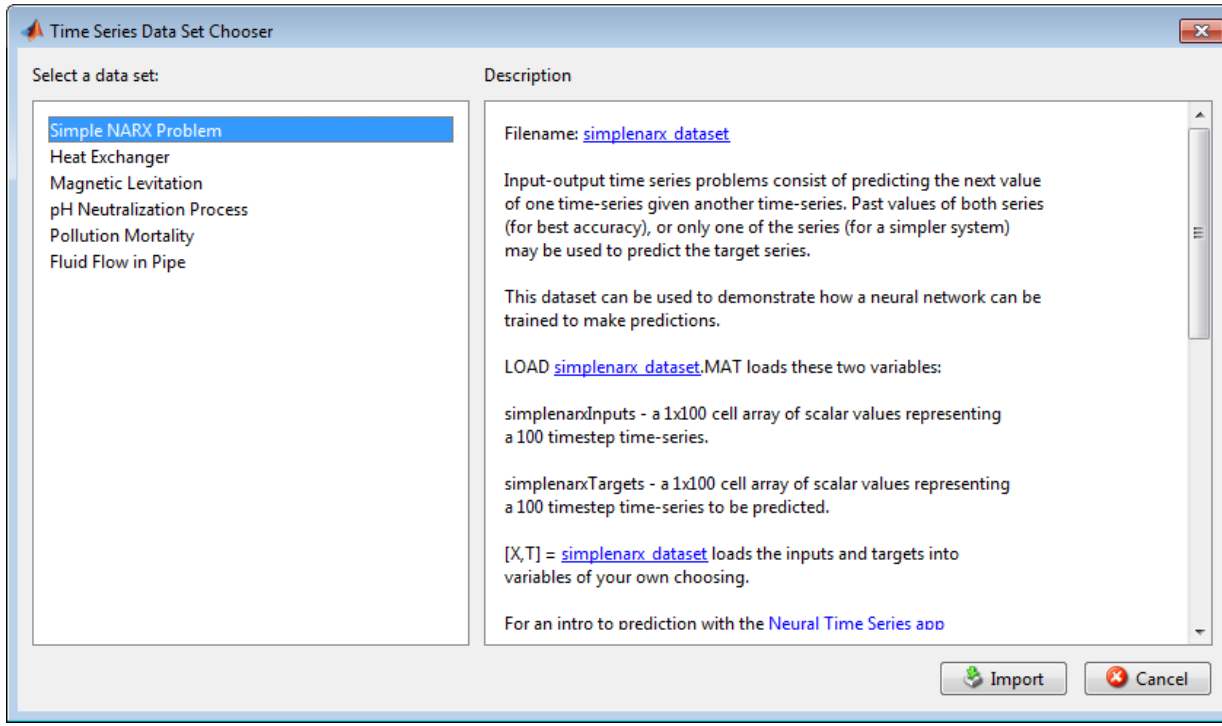
  The NARX model will provide better predictions than this input-output model, because it uses the additional information contained in the previous values of $y(t)$. However, there may be some applications in which the previous values of $y(t)$ would not be available. Those are the only cases where you would want to use the input-output model instead of the NARX model.

3 For this example, select the NARX model and click **Next** to proceed.



4 Click **Load Example Data Set** in the Select Data window. The Time Series Data Set Chooser window opens.

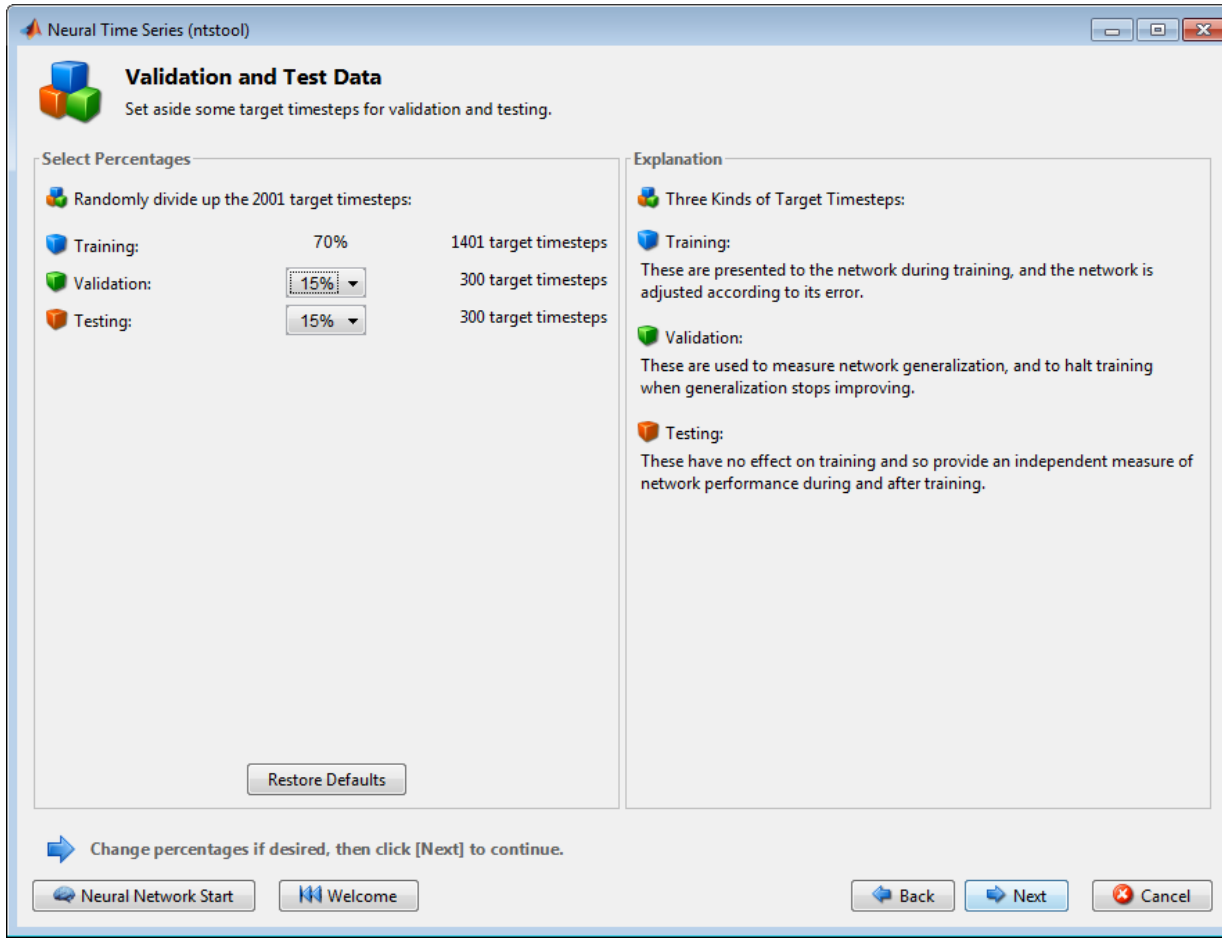
Note Use the **Inputs** and **Targets** options in the Select Data window when you need to load data from the MATLAB workspace.



5 Select **pH Neutralization Process**, and click **Import**. This returns you to the Select Data window.

6 Click **Next** to open the Validation and Test Data window, shown in the following figure.

The validation and test data sets are each set to 15% of the original data.



With these settings, the input vectors and target vectors will be randomly divided into three sets as follows:

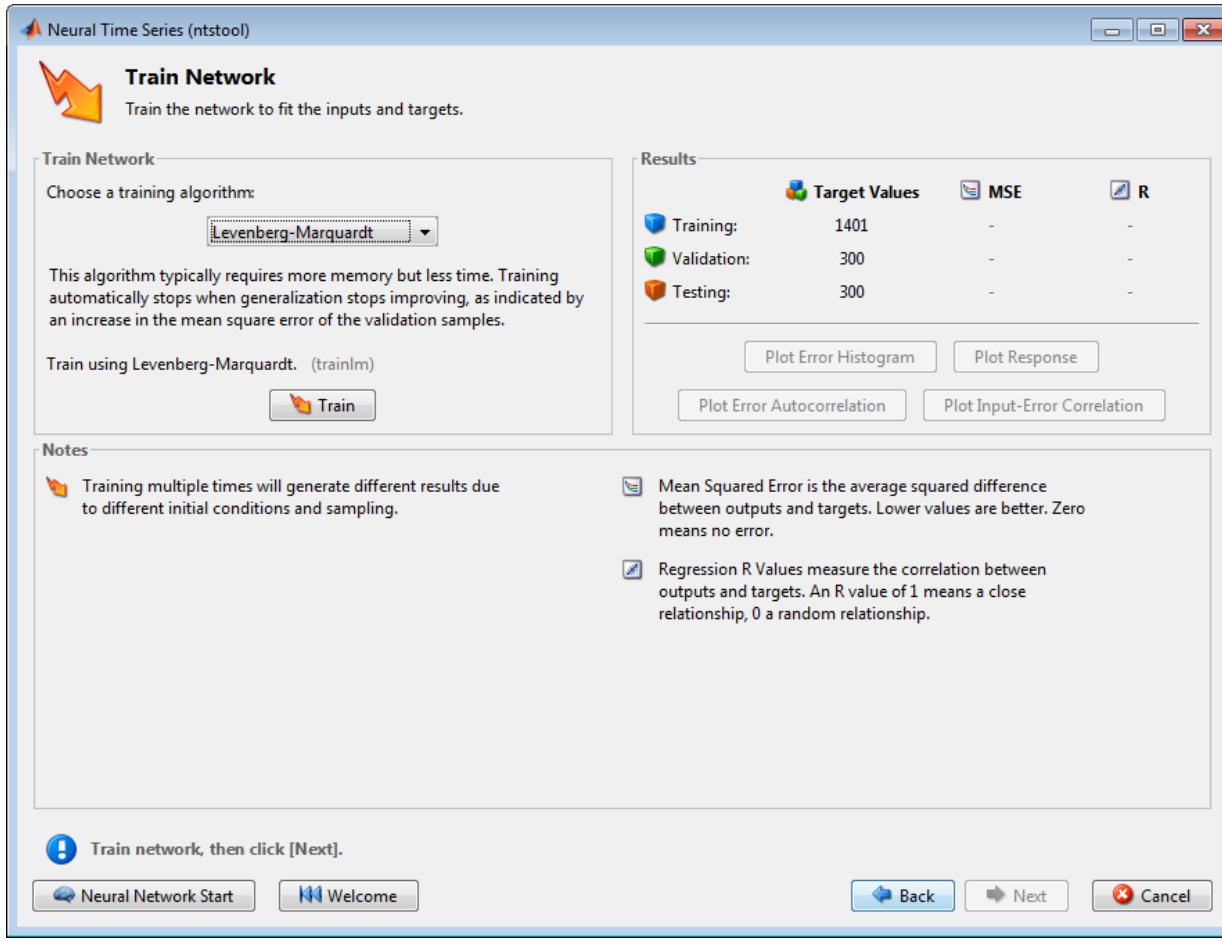
- 70% will be used for training.
- 15% will be used to validate that the network is generalizing and to stop training before overfitting.
- The last 15% will be used as a completely independent test of network generalization.

(See “Dividing the Data” for more discussion of the data division process.)

Because the true output is available during the training of the network, you can use the open-loop architecture shown above, in which the true output is used instead of feeding back the estimated output. This has two advantages. The first is that the input to the feedforward network is more accurate. The second is that the resulting network has a purely feedforward architecture, and therefore a more efficient algorithm can be used for training. This network is discussed in more detail in “NARX Network” (narxnet, closeloop).

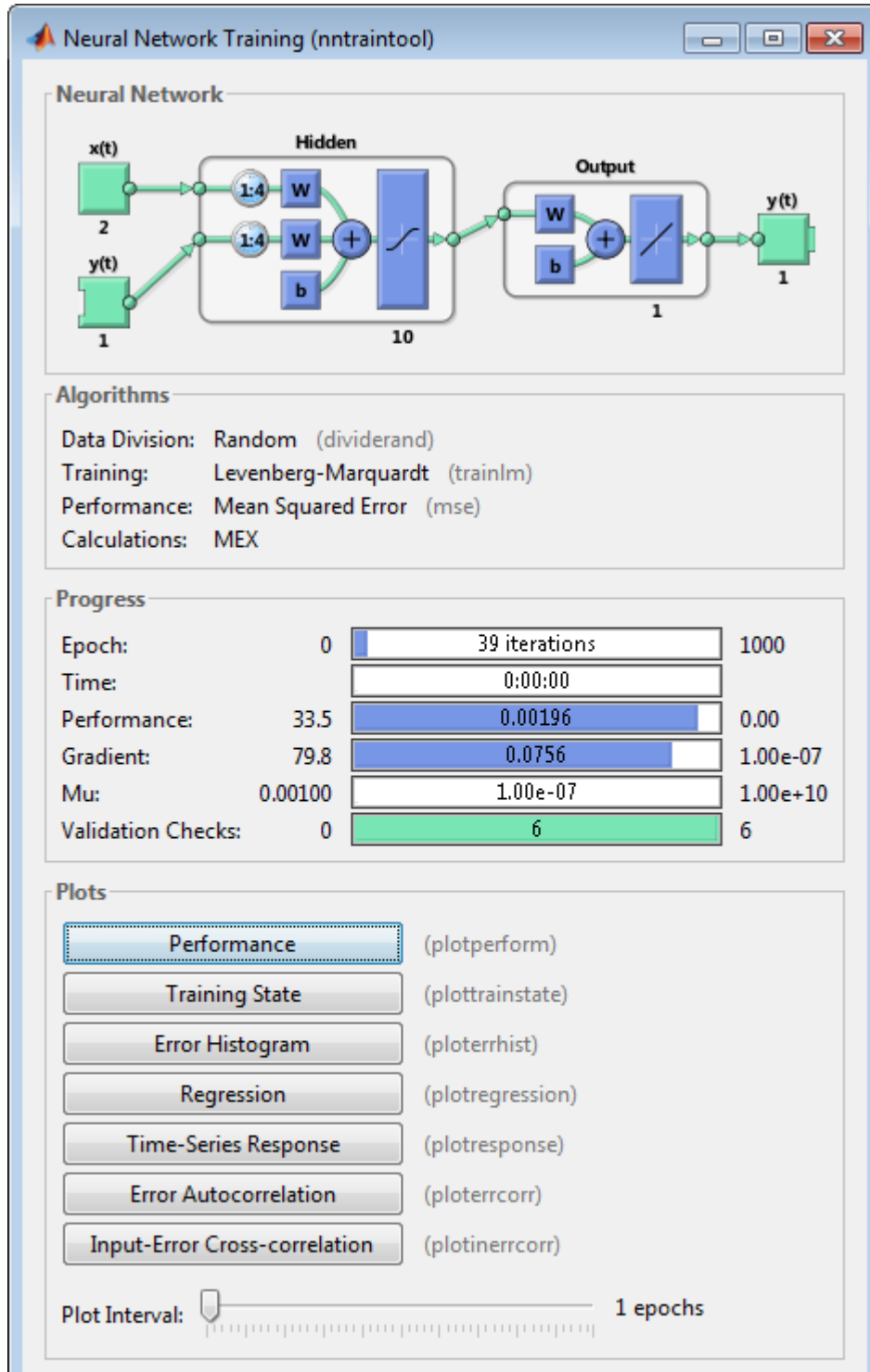
The default number of hidden neurons is set to 10. The default number of delays is 2. Change this value to 4. You might want to adjust these numbers if the network training performance is poor.

8 Click **Next**.



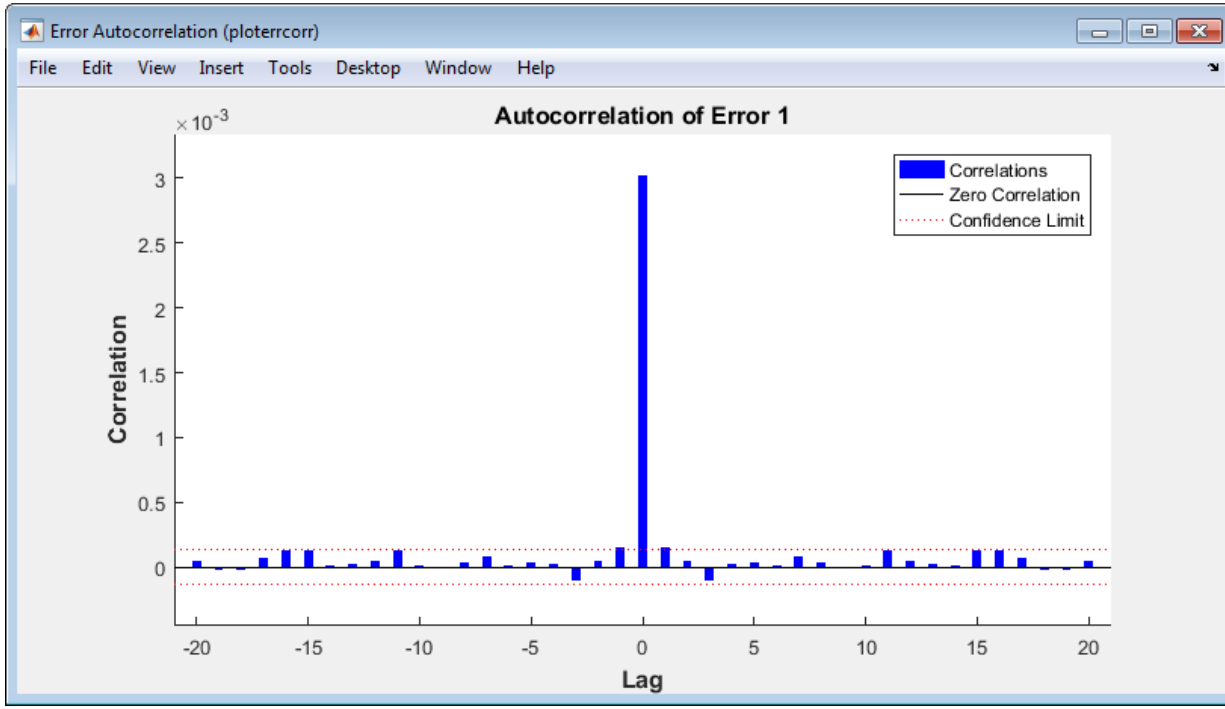
9 Select a training algorithm, then click **Train**. Levenberg-Marquardt (`trainlm`) is recommended for most problems, but for some noisy and small problems Bayesian Regularization (`trainbr`) can take longer but obtain a better solution. For large problems, however, Scaled Conjugate Gradient (`trainscg`) is recommended as it uses gradient calculations which are more memory efficient than the Jacobian calculations the other two algorithms use. This example uses the default Levenberg-Marquardt.

The training continued until the validation error failed to decrease for six iterations (validation stop).

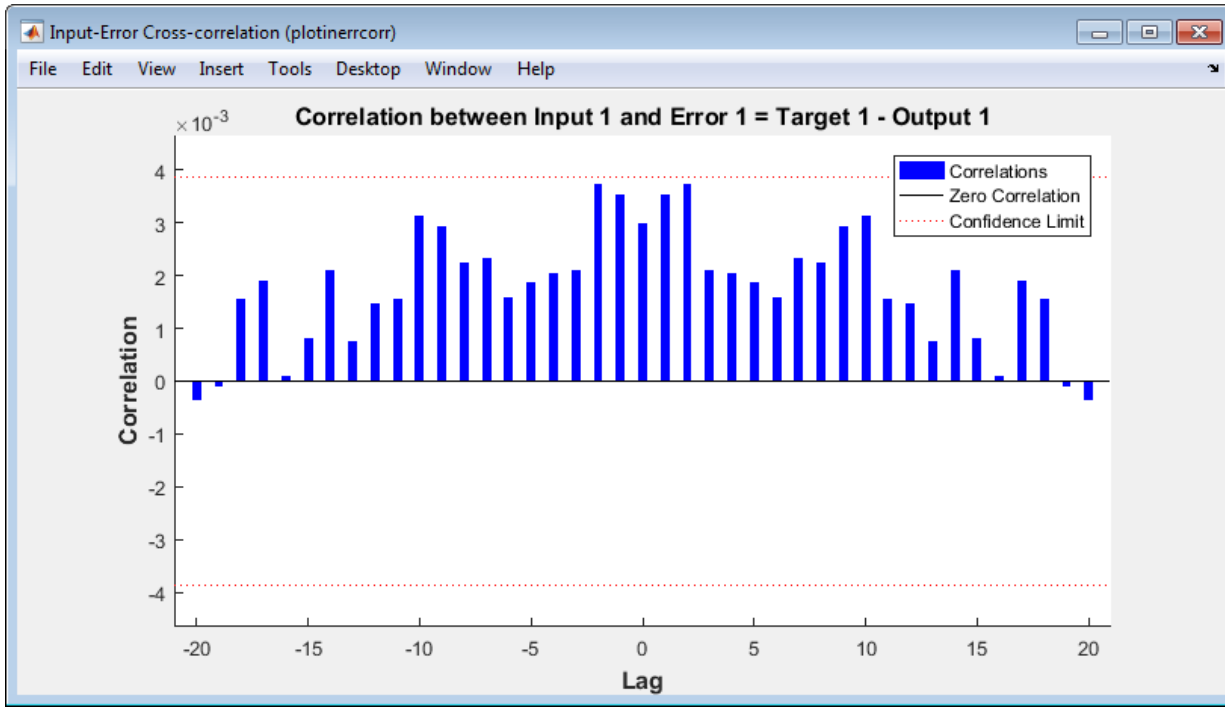

Neural Network Training (nntraintool)
Neural Network
x(t)
2
y(t)
1
Hidden
1:4
W
1:4
W
b
10
Output
W
b
1
y(t)
1
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 39 iterations 1000
Time: 0:00:00
Performance: 33.5 0.00196 0.00
Gradient: 79.8 0.0756 1.00e-07
Mu: 0.00100 1.00e-07 1.00e+10
Validation Checks: 0 6 6
Plots
Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Time-Series Response (plotresponse)
Error Autocorrelation (ploterrcorr)
Input-Error Cross-correlation (plotinerrcorr)
Plot Interval: 1 epochs

10 Under **Plots**, click **Error Autocorrelation**. This is used to validate the network performance.

The following plot displays the error autocorrelation function. It describes how the prediction errors are related in time. For a perfect prediction model, there should only be one nonzero value of the autocorrelation function, and it should occur at zero lag. (This is the mean square error.) This would mean that the prediction errors were completely uncorrelated with each other (white noise). If there was significant correlation in the prediction errors, then it should be possible to improve the prediction - perhaps by increasing the number of delays in the tapped delay lines. In this case, the correlations, except for the one at zero lag, fall approximately within the 95% confidence limits around zero, so the model seems to be adequate. If even more accurate results were required, you could retrain the network by clicking **Retrain** in `ntstool`. This will change the initial weights and biases of the network, and may produce an improved network after retraining.

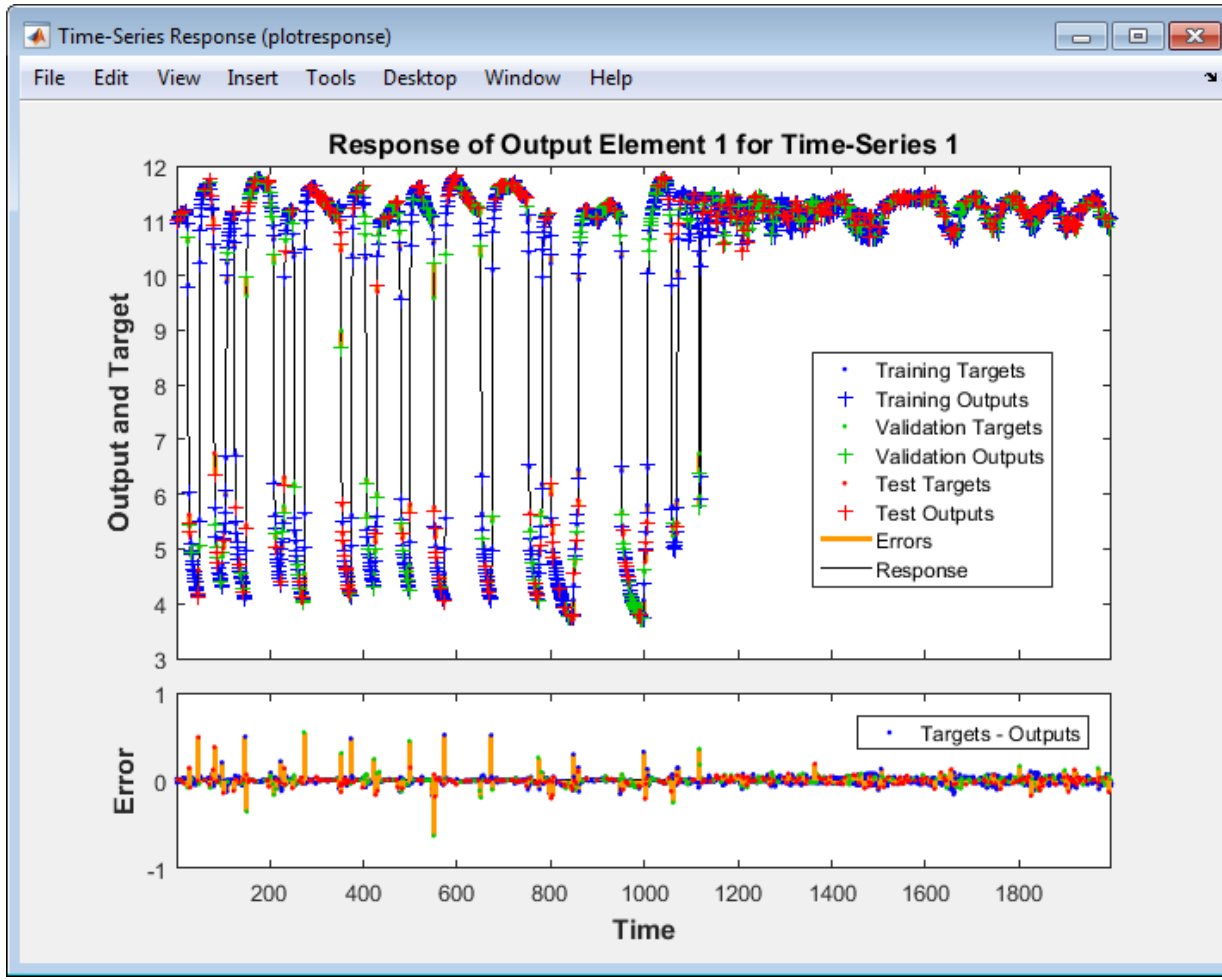


11 View the input-error cross-correlation function to obtain additional verification of network performance. Under the **Plots** pane, click **Input-Error Cross-correlation**.

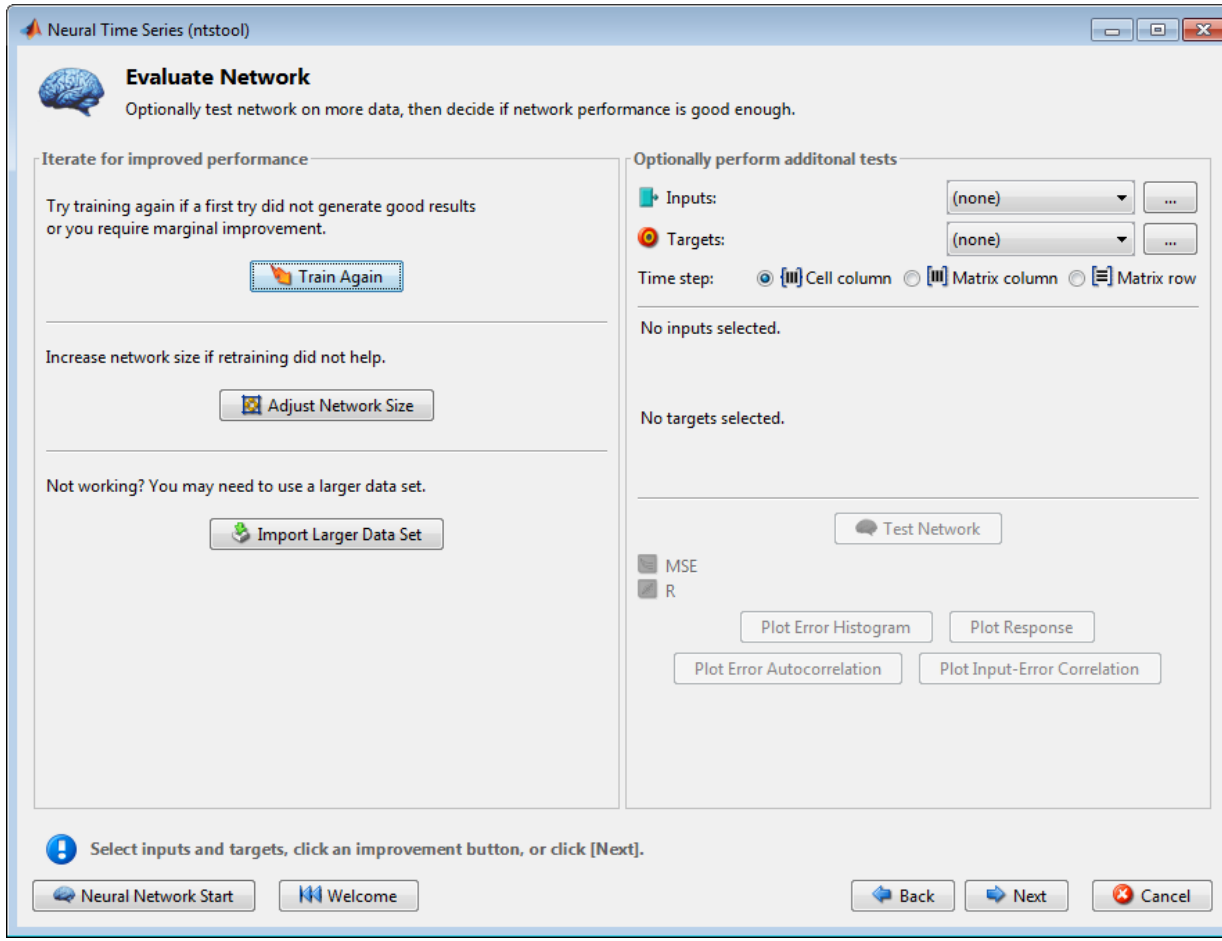


This input-error cross-correlation function illustrates how the errors are correlated with the input sequence $x(t)$. For a perfect prediction model, all of the correlations should be zero. If the input is correlated with the error, then it should be possible to improve the prediction, perhaps by increasing the number of delays in the tapped delay lines. In this case, all of the correlations fall within the confidence bounds around zero.

12 Under **Plots**, click **Time Series Response**. This displays the inputs, targets and errors versus time. It also indicates which time points were selected for training, testing and validation.



13 Click **Next** in the Neural Network Time Series App to evaluate the network.



At this point, you can test the network against new data.

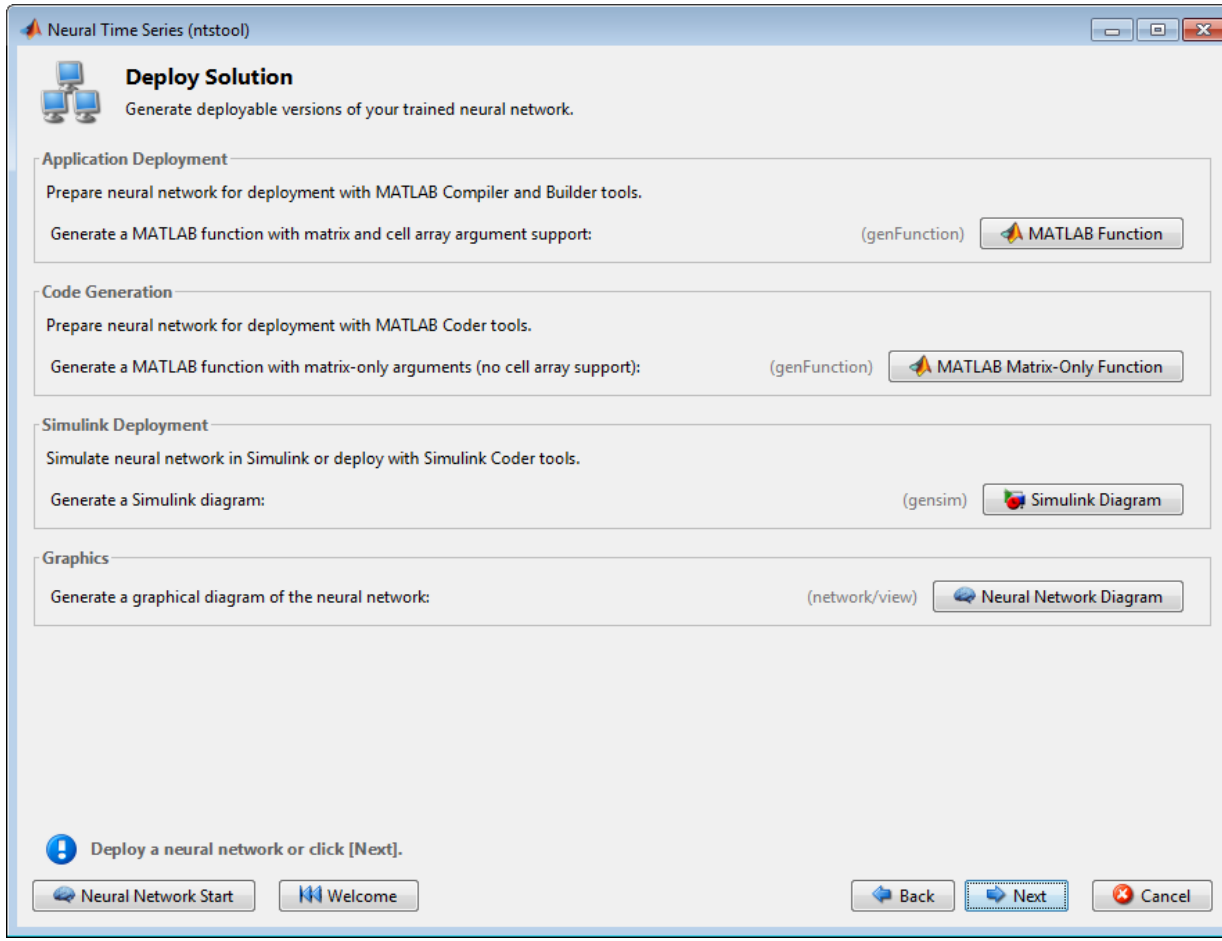
If you are dissatisfied with the network's performance on the original or new data, you can do any of the following:

- Train it again.
- Increase the number of neurons and/or the number of delays.
- Get a larger training data set.

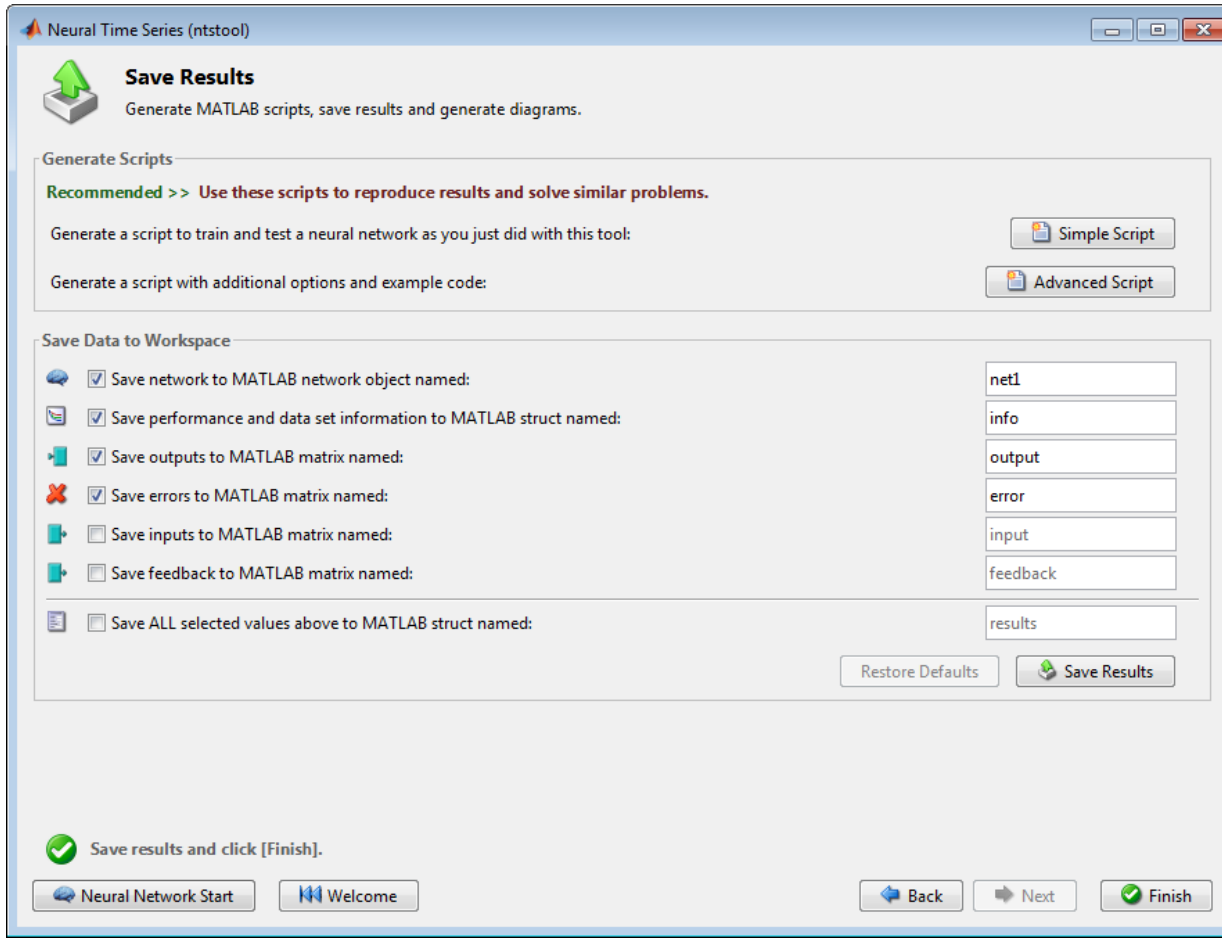
If the performance on the training set is good, but the test set performance is significantly worse, which could indicate overfitting, then reducing the number of neurons can improve your results.

14 If you are satisfied with the network performance, click **Next**.

15 Use this panel to generate a MATLAB function or Simulink diagram for simulating your neural network. You can use the generated code or diagram to better understand how your neural network computes outputs from inputs, or deploy the network with MATLAB Compiler tools and other MATLAB and Simulink code generation tools.



16 Use the buttons on this screen to generate scripts or to save your results.



- You can click **Simple Script** or **Advanced Script** to create MATLAB code that can be used to reproduce all of the previous steps from the command line. Creating MATLAB code can be helpful if you want to learn how to use the command-line functionality of the toolbox to customize the training process. In “Using Command-Line Functions” on page 1-123, you will investigate the generated scripts in more detail.
- You can also have the network saved as `net` in the workspace. You can perform additional tests on it or put it to work on new inputs.

17 After creating MATLAB code and saving your results, click **Finish**.

Using Command-Line Functions

The easiest way to learn how to use the command-line functionality of the toolbox is to generate scripts from the GUIs, and then modify them to customize the network training. As an example, look at the simple script that was created at step 15 of the previous section.

```
% Solve an Autoregression Problem with External
% Input with a NARX Neural Network
% Script generated by NTSTOOL
%
% This script assumes the variables on the right of
% these equalities are defined:
%
%   phInputs - input time series.
%   phTargets - feedback time series.

inputSeries = phInputs;
targetSeries = phTargets;

% Create a Nonlinear Autoregressive Network with External Input
inputDelays = 1:4;
feedbackDelays = 1:4;
hiddenLayerSize = 10;
net = narxnet(inputDelays,feedbackDelays,hiddenLayerSize);

% Prepare the Data for Training and Simulation
% The function PREPARETS prepares time series data
% for a particular network, shifting time by the minimum
% amount to fill input states and layer states.
% Using PREPARETS allows you to keep your original
% time series data unchanged, while easily customizing it
% for networks with differing numbers of delays, with
% open loop or closed loop feedback modes.
[inputs,inputStates,layerStates,targets] = ...
    preparets(net,inputSeries,{},targetSeries);

% Set up Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
```

```
% Train the Network
[net,tr] = train(net,inputs,targets,inputStates,layerStates);

% Test the Network
outputs = net(inputs,inputStates,layerStates);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)

% View the Network
view(net)

% Plots
% Uncomment these lines to enable various plots.
% figure, plotperform(tr)
% figure, plottrainstate(tr)
% figure, plotregression(targets,outputs)
% figure, plotresponse(targets,outputs)
% figure, ploterrcorr(errors)
% figure, plotinerrcorr(inputs,errors)

% Closed Loop Network
% Use this network to do multi-step prediction.
% The function CLOSELOOP replaces the feedback input with a direct
% connection from the output layer.
netc = closeloop(net);
netc.name = [net.name ' - Closed Loop'];
view(netc)
[xc,xic,aic,tc] = preparets(netc,inputSeries,{},targetSeries);
yc = netc(xc,xic,aic);
closedLoopPerformance = perform(netc,tc,yc)

% Early Prediction Network
% For some applications it helps to get the prediction a
% timestep early.
% The original network returns predicted y(t+1) at the same
% time it is given y(t+1).
% For some applications such as decision making, it would
% help to have predicted y(t+1) once y(t) is available, but
% before the actual y(t+1) occurs.
% The network can be made to return its output a timestep early
% by removing one delay so that its minimal tap delay is now
% 0 instead of 1.  The new network returns the same outputs as
% the original network, but outputs are shifted left one timestep.
```

```
nets = removedelay(net);
nets.name = [net.name ' - Predict One Step Ahead'];
view(nets)
[xs,xis,ais,ts] = preparets(nets,inputSeries,{},targetSeries);
ys = nets(xs,xis,ais);
earlyPredictPerformance = perform(nets,ts,ys)
```

You can save the script, and then run it from the command line to reproduce the results of the previous GUI session. You can also edit the script to customize the training process. In this case, follow each of the steps in the script.

1. The script assumes that the input vectors and target vectors are already loaded into the workspace. If the data are not loaded, you can load them as follows:

   ```
   load ph_dataset
   inputSeries = phInputs;
   targetSeries = phTargets;
   ```

2. Create a network. The NARX network, `narxnet`, is a feedforward network with the default tan-sigmoid transfer function in the hidden layer and linear transfer function in the output layer. This network has two inputs. One is an external input, and the other is a feedback connection from the network output. (After the network has been trained, this feedback connection can be closed, as you will see at a later step.) For each of these inputs, there is a tapped delay line to store previous values. To assign the network architecture for a NARX network, you must select the delays associated with each tapped delay line, and also the number of hidden layer neurons. In the following steps, you assign the input delays and the feedback delays to range from 1 to 4 and the number of hidden neurons to be 10.

   ```
   inputDelays = 1:4;
   feedbackDelays = 1:4;
   hiddenLayerSize = 10;
   net = narxnet(inputDelays,feedbackDelays,hiddenLayerSize);
   ```

   Note Increasing the number of neurons and the number of delays requires more computation, and this has a tendency to overfit the data when the numbers are set too high, but it allows the network to solve more complicated problems. More layers require more computation, but their use might result in the network solving complex problems more efficiently. To use more than one hidden layer, enter the hidden layer sizes as elements of an array in the `fitnet` command.

3. Prepare the data for training. When training a network containing tapped delay lines, it is necessary to fill the delays with initial values of the inputs and outputs of the

network. There is a toolbox command that facilitates this process - `preparets`. This function has three input arguments: the network, the input sequence and the target sequence. The function returns the initial conditions that are needed to fill the tapped delay lines in the network, and modified input and target sequences, where the initial conditions have been removed. You can call the function as follows:

```
[inputs,inputStates,layerStates,targets] = ...
    preparets(net,inputSeries,{},targetSeries);
```

4 Set up the division of data.

```
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio   = 15/100;
net.divideParam.testRatio  = 15/100;
```

With these settings, the input vectors and target vectors will be randomly divided, with 70% used for training, 15% for validation and 15% for testing.

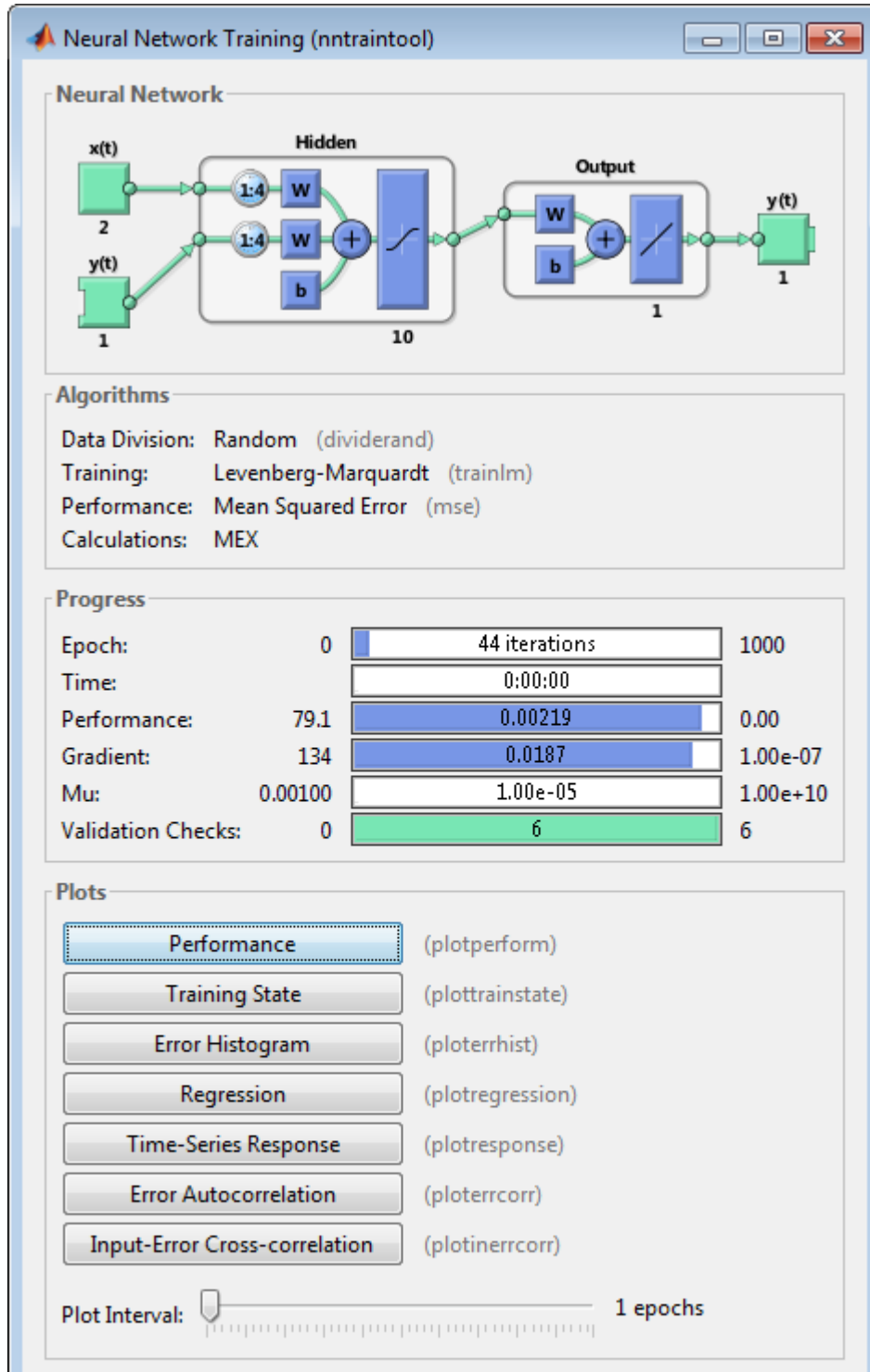
5 Train the network. The network uses the default Levenberg-Marquardt algorithm (`trainlm`) for training. For problems in which Levenberg-Marquardt does not produce as accurate results as desired, or for large data problems, consider setting the network training function to Bayesian Regularization (`trainbr`) or Scaled Conjugate Gradient (`trainscg`), respectively, with either

```
net.trainFcn = 'trainbr';
net.trainFcn = 'trainscg';
```

To train the network, enter:

```
[net,tr] = train(net,inputs,targets,inputStates,layerStates);
```

During training, the following training window opens. This window displays training progress and allows you to interrupt training at any point by clicking **Stop Training**.


Neural Network Training (nntraintool)
Neural Network
x(t)
2
y(t)
1
Hidden
1:4
W
1:4
W
b
10
Output
W
b
1
y(t)
1
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 44 iterations 1000
Time: 0:00:00
Performance: 79.1 0.00219 0.00
Gradient: 134 0.0187 1.00e-07
Mu: 0.00100 1.00e-05 1.00e+10
Validation Checks: 0 6 6
Plots
Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Time-Series Response (plotresponse)
Error Autocorrelation (ploterrcorr)
Input-Error Cross-correlation (plotinerrcorr)
Plot Interval: 1 epochs

This training stopped when the validation error increased for six iterations, which occurred at iteration 44.

6 Test the network. After the network has been trained, you can use it to compute the network outputs. The following code calculates the network outputs, errors and overall performance. Note that to simulate a network with tapped delay lines, you need to assign the initial values for these delayed signals. This is done with `inputStates` and `layerStates` provided by `preparets` at an earlier stage.

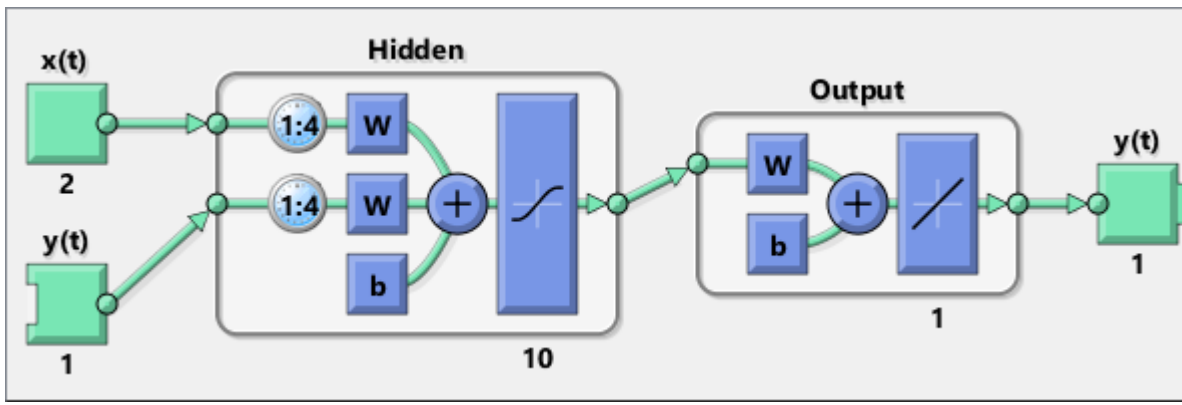
```
outputs = net(inputs,inputStates,layerStates);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)
```

```
performance =

    0.0042
```

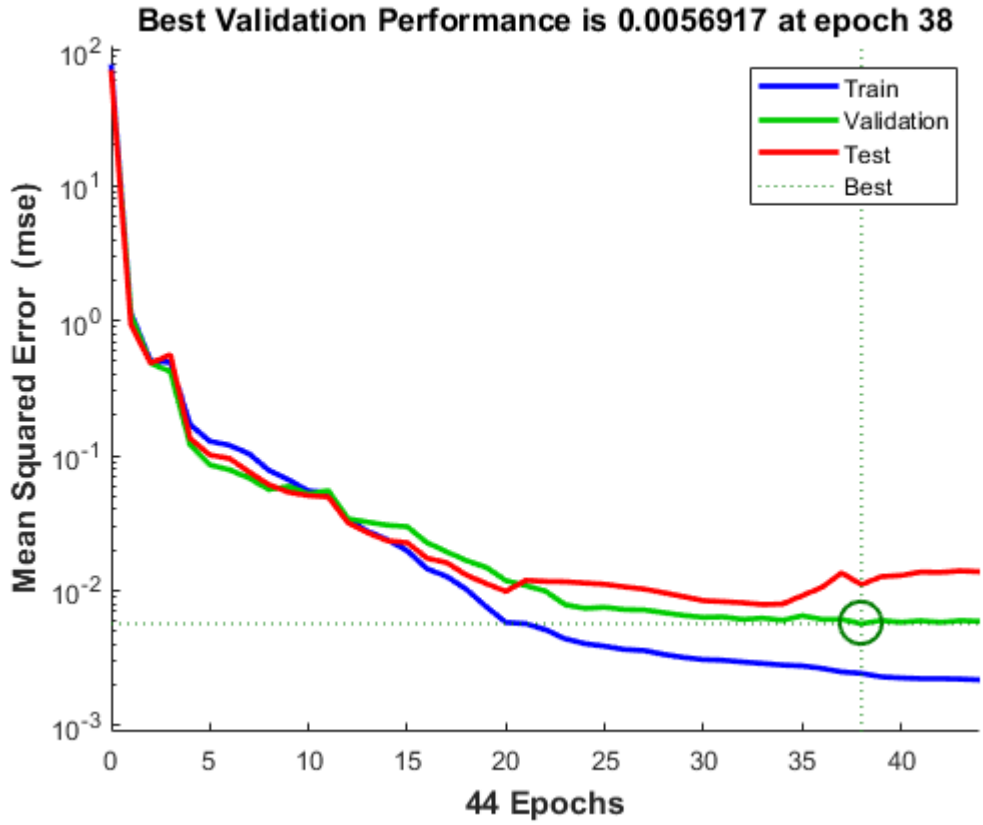
7 View the network diagram.

```
view(net)
```



8 Plot the performance training record to check for potential overfitting.

```
figure, plotperform(tr)
```



This figure shows that training and validation errors decrease until the highlighted epoch. It does not appear that any overfitting has occurred, because the validation error does not increase before this epoch.

All of the training is done in open loop (also called series-parallel architecture), including the validation and testing steps. The typical workflow is to fully create the network in open loop, and only when it has been trained (which includes validation and testing steps) is it transformed to closed loop for multistep-ahead prediction. Likewise, the R values in the GUI are computed based on the open-loop training results.

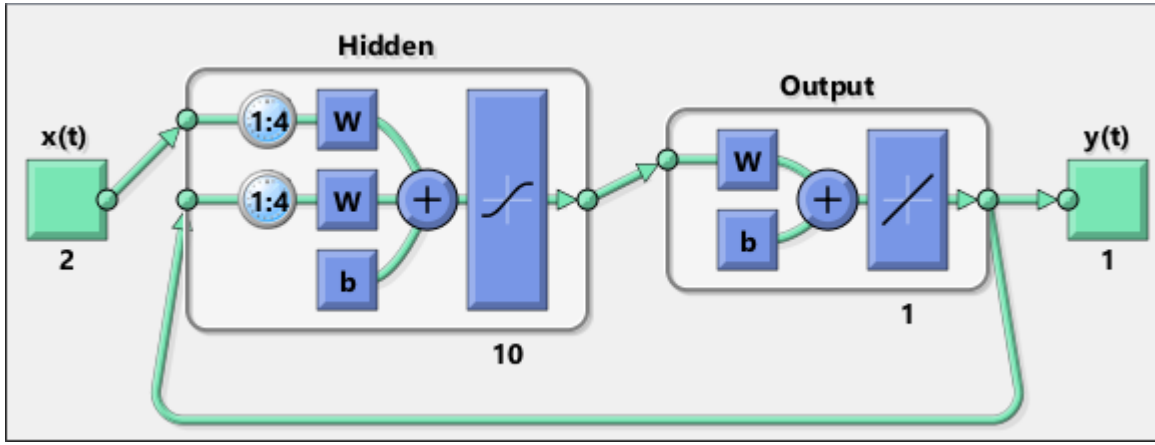
9 Close the loop on the NARX network. When the feedback loop is open on the NARX network, it is performing a one-step-ahead prediction. It is predicting the next value of $y(t)$ from previous values of $y(t)$ and $x(t)$. With the feedback loop closed, it can be

used to perform multi-step-ahead predictions. This is because predictions of $y(t)$ will be used in place of actual future values of $y(t)$. The following commands can be used to close the loop and calculate closed-loop performance

```
netc = closeloop(net);
netc.name = [net.name ' - Closed Loop'];
view(netc)
[xc,xic,aic,tc] = preparets(netc,inputSeries,{},targetSeries);
yc = netc(xc,xic,aic);
perfc = perform(netc,tc,yc)
```

```
perfc =

    2.8744
```

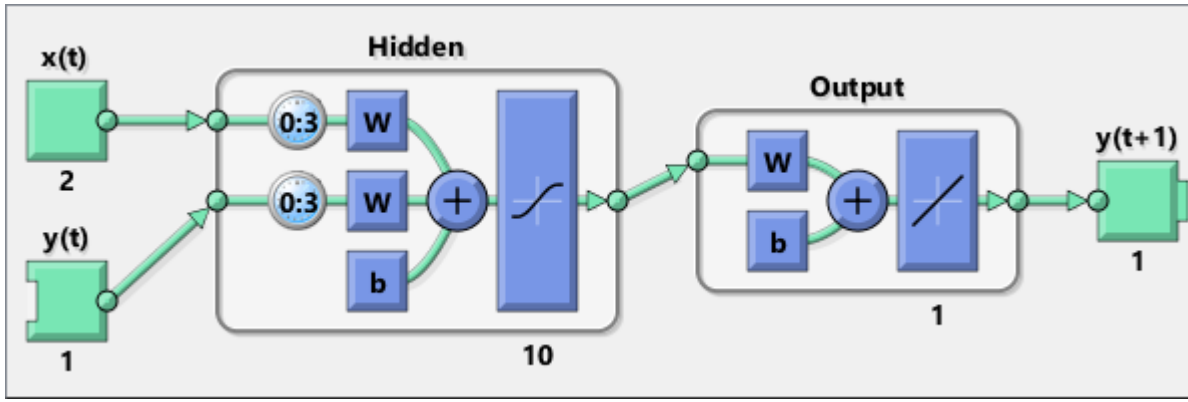


10 Remove a delay from the network, to get the prediction one time step early.

```
nets = removedelay(net);
nets.name = [net.name ' - Predict One Step Ahead'];
view(nets)
[xs,xis,ais,ts] = preparets(nets,inputSeries,{},targetSeries);
ys = nets(xs,xis,ais);
earlyPredictPerformance = perform(nets,ts,ys)
```

```
earlyPredictPerformance =
```

```
    0.0042
```



From this figure, you can see that the network is identical to the previous open-loop network, except that one delay has been removed from each of the tapped delay lines. The output of the network is then $y(t + 1)$ instead of $y(t)$. This may sometimes be helpful when a network is deployed for certain applications.

If the network performance is not satisfactory, you could try any of these approaches:

- Reset the initial network weights and biases to new values with `init` and train again (see “Initializing Weights” (init)).
- Increase the number of hidden neurons or the number of delays.
- Increase the number of training vectors.
- Increase the number of input values, if more relevant information is available.
- Try a different training algorithm (see “Training Algorithms”).

To get more experience in command-line operations, try some of these tasks:

- During training, open a plot window (such as the error correlation plot), and watch it animate.
- Plot from the command line with functions such as `plotresponse`, `ploterrcorr` and `plotperform`. (For more information on using these functions, see their reference pages.)

Also, see the advanced script for more options, when training from the command line.

Each time a neural network is trained, can result in a different solution due to different initial weight and bias values and different divisions of data into training, validation, and test sets. As a result, different neural networks trained on the same problem can give different outputs for the same input. To ensure that a neural network of good accuracy has been found, retrain several times.

There are several other techniques for improving upon initial solutions if higher accuracy is desired. For more information, see “Improve Shallow Neural Network Generalization and Avoid Overfitting”.

Train Shallow Networks on CPUs and GPUs

In this section...
“ Parallel Computing Toolbox ” on page 1-133
“Parallel CPU Workers” on page 1-134
“GPU Computing” on page 1-134
“Multiple GPU/CPU Computing” on page 1-135
“Cluster Computing with MATLAB Parallel Server” on page 1-135
“Load Balancing, Large Problems, and Beyond” on page 1-135

Parallel Computing Toolbox

Tip This topic describes shallow networks. For deep learning, see instead “Deep Learning with Big Data on GPUs and in Parallel”.

Neural network training and simulation involves many parallel calculations. Multicore CPUs, graphical processing units (GPUs), and clusters of computers with multiple CPUs and GPUs can all take advantage of parallel calculations.

Together, Deep Learning Toolbox and Parallel Computing Toolbox enable the multiple CPU cores and GPUs of a single computer to speed up training and simulation of large problems.

The following is a standard single-threaded training and simulation session. (While the benefits of parallelism are most visible for large problems, this example uses a small dataset that ships with Deep Learning Toolbox.)

```
[x, t] = bodyfat_dataset;
net1 = feedforwardnet(10);
net2 = train(net1, x, t);
y = net2(x);
```

Parallel CPU Workers

Intel® processors ship with as many as eight cores. Workstations with two processors can have as many as 16 cores, with even more possible in the future. Using multiple CPU cores in parallel can dramatically speed up calculations.

Start or get the current parallel pool and view the number of workers in the pool.

```
pool = gcp;
pool.NumWorkers
```

An error occurs if you do not have a license for Parallel Computing Toolbox.

When a parallel pool is open, set the `train` function's `'useParallel'` option to `'yes'` to specify that training and simulation be performed across the pool.

```
net2 = train(net1,x,t,'useParallel','yes');
y = net2(x,'useParallel','yes');
```

GPU Computing

GPUs can have thousands of cores on a single card and are highly efficient on parallel algorithms like neural networks.

Use `gpuDeviceCount` to check whether a supported GPU card is available in your system. Use the function `gpuDevice` to review the currently selected GPU information or to select a different GPU.

```
gpuDeviceCount
gpuDevice
gpuDevice(2) % Select device 2, if available
```

An "Undefined function or variable" error appears if you do not have a license for Parallel Computing Toolbox.

When you have selected the GPU device, set the `train` or `sim` function's `'useGPU'` option to `'yes'` to perform training and simulation on it.

```
net2 = train(net1,x,t,'useGPU','yes');
y = net2(x,'useGPU','yes');
```

Multiple GPU/CPU Computing

You can use multiple GPUs for higher levels of parallelism.

After opening a parallel pool, set both `'useParallel'` and `'useGPU'` to `'yes'` to harness all the GPUs and CPU cores on a single computer. Each worker associated with a unique GPU uses that GPU. The rest of the workers perform calculations on their CPU core.

```
net2 = train(net1,x,t,'useParallel','yes','useGPU','yes');
y = net2(x,'useParallel','yes','useGPU','yes');
```

For some problems, using GPUs and CPUs together can result in the highest computing speed. For other problems, the CPUs might not keep up with the GPUs, and so using only GPUs is faster. Set `'useGPU'` to `'only'`, to restrict the parallel computing to workers with unique GPUs.

```
net2 = train(net1,x,t,'useParallel','yes','useGPU','only');
y = net2(x,'useParallel','yes','useGPU','only');
```

Cluster Computing with MATLAB Parallel Server

MATLAB Parallel Server allows you to harness all the CPUs and GPUs on a network cluster of computers. To take advantage of a cluster, open a parallel pool with a cluster profile. Use the MATLAB **Home** tab **Environment** area **Parallel** menu to manage and select profiles.

After opening a parallel pool, train the network by calling `train` with the `'useParallel'` and `'useGPU'` options.

```
net2 = train(net1,x,t,'useParallel','yes');
y = net2(x,'useParallel','yes');

net2 = train(net1,x,t,'useParallel','yes','useGPU','only');
y = net2(x,'useParallel','yes','useGPU','only');
```

Load Balancing, Large Problems, and Beyond

For more information on parallel computing with Deep Learning Toolbox, see “Neural Networks with Parallel and GPU Computing”, which introduces other topics, such as how to manually distribute data sets across CPU and GPU workers to best take advantage of differences in machine speed and memory.

Distributing data manually also allows worker data to load sequentially, so that data sets are limited in size only by the total RAM of a cluster instead of the RAM of a single computer. This lets you apply neural networks to very large problems.

Sample Data Sets for Shallow Neural Networks

The Deep Learning Toolbox contains a number of sample data sets that you can use to experiment with shallow neural networks. To view the data sets that are available, use the following command:

```
help nndatasets
```

```
  Neural Network Datasets
  -----------------------

  Function Fitting, Function approximation and Curve fitting.

  Function fitting is the process of training a neural network on a
  set of inputs in order to produce an associated set of target outputs.
  Once the neural network has fit the data, it forms a generalization of
  the input-output relationship and can be used to generate outputs for
  inputs it was not trained on.

   simplefit_dataset     - Simple fitting dataset.
   abalone_dataset       - Abalone shell rings dataset.
   bodyfat_dataset       - Body fat percentage dataset.
   building_dataset      - Building energy dataset.
   chemical_dataset      - Chemical sensor dataset.
   cho_dataset           - Cholesterol dataset.
   engine_dataset        - Engine behavior dataset.
   vinyl_dataset         - Vinyl bromide dataset.

  ----------

  Pattern Recognition and Classification

  Pattern recognition is the process of training a neural network to assign
  the correct target classes to a set of input patterns.  Once trained the
  network can be used to classify patterns it has not seen before.

   simpleclass_dataset     - Simple pattern recognition dataset.
   cancer_dataset          - Breast cancer dataset.
   crab_dataset            - Crab gender dataset.
   glass_dataset           - Glass chemical dataset.
   iris_dataset            - Iris flower dataset.
   ovarian_dataset         - Ovarian cancer dataset.
   thyroid_dataset         - Thyroid function dataset.
   wine_dataset            - Italian wines dataset.
```

```
 digitTrain4DArrayData   - Synthetic handwritten digit dataset for
                           training in form of 4-D array.
 digitTrainCellArrayData - Synthetic handwritten digit dataset for
                           training in form of cell array.
 digitTest4DArrayData    - Synthetic handwritten digit dataset for
                           testing in form of 4-D array.
 digitTestCellArrayData  - Synthetic handwritten digit dataset for
                           testing in form of cell array.
 digitSmallCellArrayData - Subset of the synthetic handwritten digit
                           dataset for training in form of cell array.

----------

Clustering, Feature extraction and Data dimension reduction

Clustering is the process of training a neural network on patterns
so that the network comes up with its own classifications according
to pattern similarity and relative topology.  This is useful for gaining
insight into data, or simplifying it before further processing.

 simplecluster_dataset - Simple clustering dataset.

The inputs of fitting or pattern recognition datasets may also clustered.

----------

Input-Output Time-Series Prediction, Forecasting, Dynamic modeling
Nonlinear autoregression, System identification and Filtering

Input-output time series problems consist of predicting the next value
of one time series given another time series. Past values of both series
(for best accuracy), or only one of the series (for a simpler system)
may be used to predict the target series.

 simpleseries_dataset  - Simple time series prediction dataset.
 simplenarx_dataset    - Simple time series prediction dataset.
 exchanger_dataset     - Heat exchanger dataset.
 maglev_dataset        - Magnetic levitation dataset.
 ph_dataset            - Solution PH dataset.
 pollution_dataset     - Pollution mortality dataset.
 refmodel_dataset      - Reference model dataset
 robotarm_dataset      - Robot arm dataset
 valve_dataset         - Valve fluid flow dataset.
```

```
----------

Single Time-Series Prediction, Forecasting, Dynamic modeling,
Nonlinear autoregression, System identification, and Filtering

Single time series prediction involves predicting the next value of
a time series given its past values.

 simplenar_dataset     - Simple single series prediction dataset.
 chickenpox_dataset    - Monthly chickenpox instances dataset.
 ice_dataset           - Global ice volume dataset.
 laser_dataset         - Chaotic far-infrared laser dataset.
 oil_dataset           - Monthly oil price dataset.
 river_dataset         - River flow dataset.
 solar_dataset         - Sunspot activity dataset
```

Notice that all of the data sets have file names of the form `name_dataset`. Inside these files will be the arrays `nameInputs` and `nameTargets`. You can load a data set into the workspace with a command such as

```
load simplefit_dataset
```

This will load `simplefitInputs` and `simplefitTargets` into the workspace. If you want to load the input and target arrays into different names, you can use a command such as

```
[x,t] = simplefit_dataset;
```

This will load the inputs and targets into the arrays `x` and `t`. You can get a description of a data set with a command such as

```
help maglev_dataset
```

Shallow Neural Networks Glossary

ADALINE	Acronym for a linear neuron: ADAptive LINear Element.
adaption	Training method that proceeds through the specified sequence of inputs, calculating the output, error, and network adjustment for each input vector in the sequence as the inputs are presented.
adaptive filter	Network that contains delays and whose weights are adjusted after each new input vector is presented. The network adapts to changes in the input signal properties if such occur. This kind of filter is used in long distance telephone lines to cancel echoes.
adaptive learning rate	Learning rate that is adjusted according to an algorithm during training to minimize training time.
architecture	Description of the number of the layers in a neural network, each layer's transfer function, the number of neurons per layer, and the connections between layers.
backpropagation learning rule	Learning rule in which weights and biases are adjusted by error-derivative (delta) vectors backpropagated through the network. Backpropagation is commonly applied to feedforward multilayer networks. Sometimes this rule is called the *generalized delta rule*.
backtracking search	Linear search routine that begins with a step multiplier of 1 and then backtracks until an acceptable reduction in performance is obtained.
batch	Matrix of input (or target) vectors applied to the network simultaneously. Changes to the network weights and biases are made just once for the entire set of vectors in the input matrix. (The term *batch* is being replaced by the more descriptive expression “concurrent vectors.”)
batching	Process of presenting a set of input vectors for simultaneous calculation of a matrix of output vectors and/or new weights and biases.

Bayesian framework	Assumes that the weights and biases of the network are random variables with specified distributions.
BFGS quasi-Newton algorithm	Variation of Newton's optimization algorithm, in which an approximation of the Hessian matrix is obtained from gradients computed at each iteration of the algorithm.
bias	Neuron parameter that is summed with the neuron's weighted inputs and passed through the neuron's transfer function to generate the neuron's output.
bias vector	Column vector of bias values for a layer of neurons.
Brent's search	Linear search that is a hybrid of the golden section search and a quadratic interpolation.
cascade-forward network	Layered network in which each layer only receives inputs from previous layers.
Charalambous' search	Hybrid line search that uses a cubic interpolation together with a type of sectioning.
classification	Association of an input vector with a particular target vector.
competitive layer	Layer of neurons in which only the neuron with maximum net input has an output of 1 and all other neurons have an output of 0. Neurons compete with each other for the right to respond to a given input vector.
competitive learning	Unsupervised training of a competitive layer with the instar rule or Kohonen rule. Individual neurons learn to become feature detectors. After training, the layer categorizes input vectors among its neurons.
competitive transfer function	Accepts a net input vector for a layer and returns neuron outputs of 0 for all neurons except for the winner, the neuron associated with the most positive element of the net input **n**.
concurrent input vectors	Name given to a matrix of input vectors that are to be presented to a network simultaneously. All the vectors in

	the matrix are used in making just one set of changes in the weights and biases.
conjugate gradient algorithm	In the conjugate gradient algorithms, a search is performed along conjugate directions, which produces generally faster convergence than a search along the steepest descent directions.
connection	One-way link between neurons in a network.
connection strength	Strength of a link between two neurons in a network. The strength, often called weight, determines the effect that one neuron has on another.
cycle	Single presentation of an input vector, calculation of output, and new weights and biases.
dead neuron	Competitive layer neuron that never won any competition during training and so has not become a useful feature detector. Dead neurons do not respond to any of the training vectors.
decision boundary	Line, determined by the weight and bias vectors, for which the net input *n* is zero.
delta rule	See **Widrow-Hoff learning rule**.
delta vector	The delta vector for a layer is the derivative of a network's output error with respect to that layer's net input vector.
distance	Distance between neurons, calculated from their positions with a distance function.
distance function	Particular way of calculating distance, such as the Euclidean distance between two vectors.
early stopping	Technique based on dividing the data into three subsets. The first subset is the training set, used for computing the gradient and updating the network weights and biases. The second subset is the validation set. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and

	biases at the minimum of the validation error are returned. The third subset is the test set. It is used to verify the network design.
epoch	Presentation of the set of training (input and/or target) vectors to a network and the calculation of new weights and biases. Note that training vectors can be presented one at a time or all together in a batch.
error jumping	Sudden increase in a network's sum-squared error during training. This is often due to too large a learning rate.
error ratio	Training parameter used with adaptive learning rate and momentum training of backpropagation networks.
error vector	Difference between a network's output vector in response to an input vector and an associated target output vector.
feedback network	Network with connections from a layer's output to that layer's input. The feedback connection can be direct or pass through several layers.
feedforward network	Layered network in which each layer only receives inputs from previous layers.
Fletcher-Reeves update	Method for computing a set of conjugate directions. These directions are used as search directions as part of a conjugate gradient optimization procedure.
function approximation	Task performed by a network trained to respond to inputs with an approximation of a desired function.
generalization	Attribute of a network whose output for a new input vector tends to be close to outputs for similar input vectors in its training set.
generalized regression network	Approximates a continuous function to an arbitrary accuracy, given a sufficient number of hidden neurons.
global minimum	Lowest value of a function over the entire range of its input parameters. Gradient descent methods adjust weights and biases in order to find the global minimum of error for a network.

golden section search	Linear search that does not require the calculation of the slope. The interval containing the minimum of the performance is subdivided at each iteration of the search, and one subdivision is eliminated at each iteration.
gradient descent	Process of making changes to weights and biases, where the changes are proportional to the derivatives of network error with respect to those weights and biases. This is done to minimize network error.
hard-limit transfer function	Transfer function that maps inputs greater than or equal to 0 to 1, and all other values to 0.
Hebb learning rule	Historically the first proposed learning rule for neurons. Weights are adjusted proportional to the product of the outputs of pre- and postweight neurons.
hidden layer	Layer of a network that is not connected to the network output (for instance, the first layer of a two-layer feedforward network).
home neuron	Neuron at the center of a neighborhood.
hybrid bisection-cubic search	Line search that combines bisection and cubic interpolation.
initialization	Process of setting the network weights and biases to their original values.
input layer	Layer of neurons receiving inputs directly from outside the network.
input space	Range of all possible input vectors.
input vector	Vector presented to the network.
input weight vector	Row vector of weights going to a neuron.
input weights	Weights connecting network inputs to layers.
Jacobian matrix	Contains the first derivatives of the network errors with respect to the weights and biases.

Kohonen learning rule	Learning rule that trains a selected neuron's weight vectors to take on the values of the current input vector.
layer	Group of neurons having connections to the same inputs and sending outputs to the same destinations.
layer diagram	Network architecture figure showing the layers and the weight matrices connecting them. Each layer's transfer function is indicated with a symbol. Sizes of input, output, bias, and weight matrices are shown. Individual neurons and connections are not shown.
layer weights	Weights connecting layers to other layers. Such weights need to have nonzero delays if they form a recurrent connection (i.e., a loop).
learning	Process by which weights and biases are adjusted to achieve some desired network behavior.
learning rate	Training parameter that controls the size of weight and bias changes during learning.
learning rule	Method of deriving the next changes that might be made in a network *or* a procedure for modifying the weights and biases of a network.
Levenberg-Marquardt	Algorithm that trains a neural network 10 to 100 times faster than the usual gradient descent backpropagation method. It always computes the approximate Hessian matrix, which has dimensions *n*-by-*n*.
line search function	Procedure for searching along a given search direction (line) to locate the minimum of the network performance.
linear transfer function	Transfer function that produces its input as its output.
link distance	Number of links, or steps, that must be taken to get to the neuron under consideration.
local minimum	Minimum of a function over a limited range of input values. A local minimum might not be the global minimum.

log-sigmoid transfer function

Squashing function of the form shown below that maps the input to the interval (0,1). (The toolbox function is `logsig`.)

$$f(n) = \frac{1}{1 + e^{-n}}$$

Manhattan distance

The Manhattan distance between two vectors **x** and **y** is calculated as

```
D = sum(abs(x-y))
```

maximum performance increase

Maximum amount by which the performance is allowed to increase in one iteration of the variable learning rate training algorithm.

maximum step size

Maximum step size allowed during a linear search. The magnitude of the weight vector is not allowed to increase by more than this maximum step size in one iteration of a training algorithm.

mean square error function

Performance function that calculates the average squared error between the network outputs **a** and the target outputs **t**.

momentum

Technique often used to make it less likely for a backpropagation network to get caught in a shallow minimum.

momentum constant

Training parameter that controls how much momentum is used.

mu parameter

Initial value for the scalar μ.

neighborhood

Group of neurons within a specified distance of a particular neuron. The neighborhood is specified by the indices for all the neurons that lie within a radius d of the winning neuron i^*:

$$Ni(d) = \{j, d_{ij} \leq d\}$$

net input vector

Combination, in a layer, of all the layer's weighted input vectors with its bias.

neuron	Basic processing element of a neural network. Includes weights and bias, a summing junction, and an output transfer function. Artificial neurons, such as those simulated and trained with this toolbox, are abstractions of biological neurons.
neuron diagram	Network architecture figure showing the neurons and the weights connecting them. Each neuron's transfer function is indicated with a symbol.
ordering phase	Period of training during which neuron weights are expected to order themselves in the input space consistent with the associated neuron positions.
output layer	Layer whose output is passed to the world outside the network.
output vector	Output of a neural network. Each element of the output vector is the output of a neuron.
output weight vector	Column vector of weights coming from a neuron or input. (See also **outstar learning rule**.)
outstar learning rule	Learning rule that trains a neuron's (or input's) output weight vector to take on the values of the current output vector of the postweight layer. Changes in the weights are proportional to the neuron's output.
overfitting	Case in which the error on the training set is driven to a very small value, but when new data is presented to the network, the error is large.
pass	Each traverse through all the training input and target vectors.
pattern	A vector.
pattern association	Task performed by a network trained to respond with the correct output vector for each input vector presented.
pattern recognition	Task performed by a network trained to respond when an input vector close to a learned vector is presented. The

	network “recognizes” the input as one of the original target vectors.
perceptron	Single-layer network with a hard-limit transfer function. This network is often trained with the perceptron learning rule.
perceptron learning rule	Learning rule for training single-layer hard-limit networks. It is guaranteed to result in a perfectly functioning network in finite time, given that the network is capable of doing so.
performance	Behavior of a network.
performance function	Commonly the mean squared error of the network outputs. However, the toolbox also considers other performance functions. Type `nnets` and look under performance functions.
Polak-Ribiére update	Method for computing a set of conjugate directions. These directions are used as search directions as part of a conjugate gradient optimization procedure.
positive linear transfer function	Transfer function that produces an output of zero for negative inputs and an output equal to the input for positive inputs.
postprocessing	Converts normalized outputs back into the same units that were used for the original targets.
Powell-Beale restarts	Method for computing a set of conjugate directions. These directions are used as search directions as part of a conjugate gradient optimization procedure. This procedure also periodically resets the search direction to the negative of the gradient.
preprocessing	Transformation of the input or target data before it is presented to the neural network.
principal component analysis	Orthogonalize the components of network input vectors. This procedure can also reduce the dimension of the input vectors by eliminating redundant components.

quasi-Newton algorithm — Class of optimization algorithm based on Newton's method. An approximate Hessian matrix is computed at each iteration of the algorithm based on the gradients.

radial basis networks — Neural network that can be designed directly by fitting special response elements where they will do the most good.

radial basis transfer function — The transfer function for a radial basis neuron is

$$radbas(n) = e^{-n^2}$$

regularization — Modification of the performance function, which is normally chosen to be the sum of squares of the network errors on the training set, by adding some fraction of the squares of the network weights.

resilient backpropagation — Training algorithm that eliminates the harmful effect of having a small slope at the extreme ends of the sigmoid squashing transfer functions.

saturating linear transfer function — Function that is linear in the interval (-1,+1) and saturates outside this interval to -1 or +1. (The toolbox function is `satlin`.)

scaled conjugate gradient algorithm — Avoids the time-consuming line search of the standard conjugate gradient algorithm.

sequential input vectors — Set of vectors that are to be presented to a network one after the other. The network weights and biases are adjusted on the presentation of each input vector.

sigma parameter — Determines the change in weight for the calculation of the approximate Hessian matrix in the scaled conjugate gradient algorithm.

sigmoid — Monotonic S-shaped function that maps numbers in the interval (-∞,∞) to a finite interval such as (-1,+1) or (0,1).

simulation — Takes the network input **p**, and the network object `net`, and returns the network outputs **a**.

spread constant — Distance an input vector must be from a neuron's weight vector to produce an output of 0.5.

squashing function — Monotonically increasing function that takes input values between -∞ and +∞ and returns values in a finite interval.

star learning rule — Learning rule that trains a neuron's weight vector to take on the values of the current input vector. Changes in the weights are proportional to the neuron's output.

sum-squared error — Sum of squared differences between the network targets and actual outputs for a given input vector or set of vectors.

supervised learning — Learning process in which changes in a network's weights and biases are due to the intervention of any external teacher. The teacher typically provides output targets.

symmetric hard-limit transfer function — Transfer that maps inputs greater than or equal to 0 to +1, and all other values to -1.

symmetric saturating linear transfer function — Produces the input as its output as long as the input is in the range -1 to 1. Outside that range the output is -1 and +1, respectively.

tan-sigmoid transfer function — Squashing function of the form shown below that maps the input to the interval (-1,1). (The toolbox function is `tansig`.)

$$f(n) = \frac{1}{1 + e^{-n}}$$

tapped delay line — Sequential set of delays with outputs available at each delay output.

target vector — Desired output vector for a given input vector.

test vectors — Set of input vectors (not used directly in training) that is used to test the trained network.

topology functions — Ways to arrange the neurons in a grid, box, hexagonal, or random topology.

training	Procedure whereby a network is adjusted to do a particular job. Commonly viewed as an offline job, as opposed to an adjustment made during each time interval, as is done in adaptive training.
training vector	Input and/or target vector used to train a network.
transfer function	Function that maps a neuron's (or layer's) net output **n** to its actual output.
tuning phase	Period of SOFM training during which weights are expected to spread out relatively evenly over the input space while retaining their topological order found during the ordering phase.
underdetermined system	System that has more variables than constraints.
unsupervised learning	Learning process in which changes in a network's weights and biases are not due to the intervention of any external teacher. Commonly changes are a function of the current network input vectors, output vectors, and previous weights and biases.
update	Make a change in weights and biases. The update can occur after presentation of a single input vector or after accumulating changes over several input vectors.
validation vectors	Set of input vectors (not used directly in training) that is used to monitor training progress so as to keep the network from overfitting.
weight function	Weight functions apply weights to an input to get weighted inputs, as specified by a particular function.
weight matrix	Matrix containing connection strengths from a layer's inputs to its neurons. The element $w_{i,j}$ of a weight matrix `W` refers to the connection strength from input `j` to neuron `i`.
weighted input vector	Result of applying a weight to a layer's input, whether it is a network input or the output of another layer.

Widrow-Hoff learning rule	Learning rule used to train single-layer linear networks. This rule is the predecessor of the backpropagation rule and is sometimes referred to as the delta rule.