Robust learning and generalization with support vector machines

2001 J. Phys. A: Math. Gen. 34 4377

(http://iopscience.iop.org/0305-4470/34/21/301)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **34** (2001) 4377–4388 www.iop.org/Journals/ja PII: S0305-4470(01)19481-3

Robust learning and generalization with support vector machines

Arnaud Buhot[1] **and Mirta B Gordon**[2,3]

[1] Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK
[2] Département de Recherche Fondamentale sur la Matière Condensée, CEA/Grenoble, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France
[3] Centre National de Recherche Scientifique, France

Received 15 November 2000

Abstract

In this paper, we study the typical learning properties of the recently proposed support vector machines (SVMs). The generalization error on linearly separable tasks, the capacity, the typical number of support vectors, the margin and the robustness or noise tolerance of a class of SVMs are determined in the framework of statistical mechanics. The robustness is shown to be closely related to the generalization properties of these machines.

PACS numbers: 0520, 0510, 0705M, 8436, 8718S

1. Introduction

Support vector machines (SVMs), recently proposed to solve the problem of learning classification tasks from examples, have aroused a great deal of interest due to the simplicity of their implementation, and to their remarkable performances on difficult tasks [1–3]. Classification of data is a very general problem, as many real-life applications, like pattern recognition, medical diagnosis etc, may be cast as classification tasks. In the last few years, much work has been done to understand how high-performance learning may be achieved, mainly within the paradigm of neural networks. These are systems composed of interconnected neurons, which are two-state units like Ising spins. As in magnets, the neuron's state is determined by the sign of the weighted sum of its inputs, which acts as an external field, and of the states of its neighbours. Learning with neural networks means determining their connectivity and the weights of the connections. The aim is to classify correctly not only the examples, or training patterns, constituting our data or training set but also new data, as we expect that the learning system will be able to generalize.

A single neuron connected to its inputs, the *simple perceptron*, is the elementary neural network. It separates the input patterns into two classes by a hyperplane orthogonal to a vector whose components are the connection weights. Thus, the simple perceptron can learn without errors only linearly separable (LS) tasks. Most classification problems are not LS, requiring learning machines with more degrees of freedom. However, the relationship between the machine's complexity, its learning capacity and its generalization ability is still an open

0305-4470/01/214377+12$30.00

problem. Feedforward layered networks, called *multilayered perceptrons*, are the most popular learning machines. Their architecture is usually found through a trial and error procedure, in which the weights are determined with backpropagation [4], a learning algorithm that performs a gradient descent on a cost function. Its main drawback is that it usually gets trapped in metastable states [5, 6]. Growth heuristics that avoid using backpropagation have also been proposed [7, 8].

SVMs [1–3] are an alternative solution to the problem of learning from examples. The idea underlying SVM is to map the patterns from the input space to a new space, the *feature space*, through a nonlinear transformation chosen *a priori*. Provided that the dimension of the feature space is large enough, the image of the training patterns will be LS, i.e. learnable by a simple perceptron in the feature space. It is well known that, if the training set is LS, there is an infinite number of error-free separating hyperplanes. Among them, the *maximal stability perceptron* (MSP) has weights that maximize the distance of the patterns closest to the separating hyperplane. The SVM weight vector is that of the MSP in feature space. The patterns closest to the separating hyperplane are called *support vectors* (SVs); their distance to the hyperplane is the *maximal stability* or SV margin. The important point is that the SVs determine uniquely the MSP. Their number is proportional to the number of training patterns, and *not* to the dimension of the feature space (which may be huge). Thus, increasing the feature-space dimension does not necessarily increase the number of parameters to be learned, a fact that makes the SVM very attractive for applications. For example, in the problem of digit recognition [1], the input space of dimension 256 needs to be mapped onto a space of dimension $256^7 \sim 10^{16}$, but the number of SVs is as low as 422. However, in spite of the high performance reached by SVMs in realistic problems [1, 2], a clear theoretical understanding of their properties is still lacking even if a few attempts have been made [9–13].

In this paper, we determine theoretically some of the learning properties of a class of SVMs in order to get a better understanding of their relevant parameters. The paper is organized as follows: in section 2, we introduce the class of SVMs considered. They are defined by a particular family of mappings between the input space and the feature space. In section 3, we consider the statistical mechanics of these SVMs and we address several important questions about these machines. The generalization error in the particular case of learning an LS task is shown to decrease more slowly than that of a simple perceptron (in input space) as a function of the number of training patterns. The capacity increases proportionally to the dimension of the feature space. The number of SVs and the SV margin present interesting scaling with the number of features. In section 4, we introduce the probability of misclassification of training patterns corrupted after learning, which is shown to be a decreasing function of the SV margin. This property, that we call robustness or noise tolerance, may account for the good generalization performance of SVMs in applications. Finally, in section 5, we discuss our results and give some conclusions.

2. A family of support vector machines

2.1. Mapping to the feature space

We focus on SVMs defined by a nonlinear transformation Φ that maps the N-dimensional input space onto a $(k+1)N$-dimensional feature space through

$$\boldsymbol{\xi} \to \Phi(\boldsymbol{\xi}) = \{\boldsymbol{\xi}, \phi(\lambda_1)\,\boldsymbol{\xi}, \ldots, \phi(\lambda_k)\,\boldsymbol{\xi}\} \tag{1}$$

where the λ_i are functions of $\boldsymbol{\xi}$. The components $\phi(\lambda_i)\,\boldsymbol{\xi}$ $(i = 1, \ldots, k)$ are the *new features* that hopefully should make the task LS in feature space.

In the following we consider odd functions ϕ, and $\lambda_i = \boldsymbol{\xi} \cdot \boldsymbol{B}_i$ where $\{\boldsymbol{B}_i\}_{i=1,\ldots,k}$ is a set of k unitary orthogonal vectors ($\boldsymbol{B}_i \cdot \boldsymbol{B}_j = \delta_{ij}$). For example, the k first generators $\{e_1, e_2, \ldots, e_k\}$ of the input space ($e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$, ...) are one possible realization of the $\boldsymbol{B}_i$. In the thermodynamic limit considered below, any set of k randomly selected normalized vectors $\boldsymbol{B}_i$ satisfies the orthogonality constraint with probability one, and the slight correlation between the new features can be neglected.

The functions $\phi(\lambda) = \text{sign}(\lambda)$ and $\phi(\lambda) = \lambda$ are of particular interest. With the former, simple scaling laws for the SVM will be deduced exactly from the properties of the MSP in input space. If $k = N$, an SVM using the latter can implement all the possible discriminating surfaces of second order in input space, defined by the quadratic kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}(1 + \boldsymbol{x} \cdot \boldsymbol{y})$. More complicated transformations Φ, equivalent to higher-order surfaces, are defined by other kernels (for examples, see [1, 11]).

The weights of the connections of the SVM correspond to a $(k + 1)N$-dimensional vector $\boldsymbol{J} = \{\boldsymbol{J}_0, \boldsymbol{J}_1, \ldots, \boldsymbol{J}_k\}$. Hereafter we consider normalized weights, $\boldsymbol{J} \cdot \boldsymbol{J} = (k + 1)N$ without any lack of generality, but we *do not* impose any constraint on the normalization of each N-dimensional vector $\boldsymbol{J}_i$. The space of weights is similar to that of a multilayer perceptron with one hidden layer composed of $k + 1$ units. In the following, we will compare some of the learning properties of the SVM with k new features to those of monolayer perceptrons with $k + 1$ units in the hidden layer. This is why we restrict ourselves to mappings with $k \ll N$ new features.

2.2. Learning strategy

The output of the SVM with weights $\boldsymbol{J}$ to a pattern $\boldsymbol{\xi}$ is $\sigma = \text{sign}(\boldsymbol{J} \cdot \Phi(\boldsymbol{\xi}))$. The aim of learning is to determine a vector $\boldsymbol{J}$ such that the patterns of the training set are correctly classified (we restrict ourselves to problems where error-free learning is possible). Any vector $\boldsymbol{J}$ that meets these conditions separates linearly, in the feature space, the image by Φ of patterns with output +1 from those with output -1. Due to the nonlinearity of the mapping, this separation is not linear in input space. In the following, we restrict ourselves to quadratic kernels for simplicity.

We assume that we are given a training set $\mathcal{L}_\alpha$ of P independent N-dimensional vectors, the *training patterns* $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\ldots,P}$ and their corresponding classes $\tau^\mu = \pm 1$. The patterns are supposed to be drawn with a probability density

$$P(\boldsymbol{\xi}) = (2\pi)^{-N/2} \exp\left(-\frac{\boldsymbol{\xi}^2}{2}\right) \tag{2}$$

and the classes τ are given by an unknown function $\tau(\boldsymbol{\xi})$ called the supervisor or teacher.

The stability of a training pattern $\boldsymbol{\xi}^\mu$ of class τ^μ in feature space is defined as

$$\gamma^\mu = \frac{\tau^\mu \, \boldsymbol{J} \cdot \Phi(\boldsymbol{\xi}^\mu)}{\sqrt{(k + 1)N}}. \tag{3}$$

Geometrically, $|\gamma^\mu|$ is the distance of the image $\Phi(\boldsymbol{\xi}^\mu)$ of pattern $\boldsymbol{\xi}^\mu$ from the hyperplane orthogonal to $\boldsymbol{J}$ and passing through the origin. Then, the aim of the learning process is to determine a vector $\boldsymbol{J}$ such that $\sigma^\mu = \tau^\mu$, or equivalently $\gamma^\mu > 0$, for all μ. If these conditions are satisfied, we can define the margin as follows:

$$\kappa(\boldsymbol{J}) = \inf_\mu \gamma^\mu \tag{4}$$

which corresponds to the distance of the closest patterns from the separating hyperplane. The solution of the SVM consists in the MSP weight vector $\boldsymbol{J}^*$ with the largest margin called the

maximal stability or SV margin:

$$\kappa_{\max}(\boldsymbol{J}^*) = \max_{\boldsymbol{J}} \kappa(\boldsymbol{J}). \tag{5}$$

The training patterns at distance $\kappa_{\max}$ from the hyperplane are the SVs. It has been shown that the MSP weight vector is a linear combination of the SVs [1, 14],

$$\boldsymbol{J}^* = \sum_{\mu \in \mathrm{SV}} a^\mu \tau^\mu \Phi(\boldsymbol{\xi}^\mu). \tag{6}$$

The a^μ are positive parameters to be determined by the learning algorithm, which has to determine also the number of SVs. Generally, this number is small compared with the feature-space dimension, a fact that allows us to increase the latter considerably without increasing dramatically the number of parameters to be determined. The SVM in input space ($k = 0$) or *linear SVM* is the usual MSP, whose properties have been extensively studied (see [15] and references therein).

3. Statistical mechanics of support vector machines

We obtain the generic properties of the SVM through the by now standard replica approach [16]. Results are obtained in the thermodynamic limit, in which the input-space dimension and the number of training patterns go to infinity ($N \to +\infty$, $P \to +\infty$), keeping the reduced number of patterns (or training set size) $\alpha = P/N$ constant. In this limit, the SVM properties are independent of the training set. The appropriate cost function is

$$E(\boldsymbol{J}, \mathcal{L}_\alpha, \kappa) = \sum_{\mu=1}^{P} \Theta(\kappa - \gamma^\mu) \tag{7}$$

where Θ is the Heaviside function and $\mathcal{L}_\alpha$ represents the training set. This cost function counts the number of training patterns that have a stability smaller than κ in feature space. The largest value of κ that satisfies $E(\boldsymbol{J}^*, \mathcal{L}_\alpha, \kappa) = 0$ is the SV margin. The weight vector $\boldsymbol{J}^*$ defines the SVM. Its generic properties are determined by the free energy

$$f = \lim_{\beta \to +\infty} \lim_{N \to +\infty} -\frac{1}{\beta N} \langle \ln Z \rangle \tag{8}$$

where

$$Z = \int \mathrm{d}\boldsymbol{J}\, \delta\left((k+1)N - \boldsymbol{J}\cdot\boldsymbol{J}\right)\, \exp\left(-\beta E(\boldsymbol{J}, \mathcal{L}_\alpha, \kappa)\right) \tag{9}$$

is the partition function and β is an inverse temperature. In equation (8), the bracket represents the average over all the possible training sets $\mathcal{L}_\alpha$ at given α.

The learning problem consists in minimizing the cost function (or energy) (7). As a consequence, we are particularly interested in the zero-temperature limit (or $\beta \to +\infty$). If the problem is LS, then $f = 0$ for some $\kappa \geqslant 0$, meaning that error-free learning is possible. In general, the probability of error-free learning vanishes beyond some value of κ. The maximal value of κ for which $f = 0$ is the *typical* value of $\kappa_{\max}(k, \alpha)$.

The free energy is calculated using the replica trick [16–18]

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle Z^n \rangle. \tag{10}$$

3.1. Linearly separable task

We consider first the case of a teacher that is a simple perceptron in input space, of (unknown) N-dimensional normalized weights $\boldsymbol{K}$ ($\boldsymbol{K}\cdot\boldsymbol{K}=1$). Thus, the classes of the training patterns $\boldsymbol{\xi}^\mu$ are $\tau^\mu = \mathrm{sign}\,(\boldsymbol{K}\cdot\boldsymbol{\xi}^\mu)$. In this case, an error-free solution exists for all α, and we are interested in the generalization error $\epsilon_g(k,\alpha)$, which is the probability that the trained SVM misclassifies a new pattern $\boldsymbol{\xi}$. Clearly, we do not expect that an SVM with $k>0$ will perform well on this task, as it corresponds to a case where the *a priori* selected feature space is too complex. However, this may well be the case in real applications. We begin by considering this LS problem mainly because other properties considered below, like the capacity and robustness, can easily be deduced by disregarding, or setting to zero, some of the order parameters introduced here. These are

$$R^a = \frac{\boldsymbol{J}_0^a\cdot\boldsymbol{K}}{\sqrt{\boldsymbol{J}_0^a\cdot\boldsymbol{J}_0^a}} \tag{11a}$$

$$v_i^a = \frac{\boldsymbol{J}_i^a\cdot\boldsymbol{J}_i^a}{N} \tag{11b}$$

$$c_i^{ab} = \lim_{\beta\to+\infty}\beta\frac{(\boldsymbol{J}_i^a-\boldsymbol{J}_i^b)^2}{2N} \qquad (a\neq b) \tag{11c}$$

for $i=0,\ldots,k$. $\boldsymbol{J}^a$ and $\boldsymbol{J}^b$ are the weight vectors of replicas a and b. The cross-overlaps $\boldsymbol{J}_i^a\cdot\boldsymbol{J}_j^b$ ($i\neq j$) and $\boldsymbol{K}\cdot\boldsymbol{B}_i$ may be neglected for $k\ll N$, as they are of order $1/\sqrt{N}$. The parameters c_i^{ab} are a generalization of the parameter $x^{ab}=\lim_{\beta\to+\infty}\beta(1-q^{ab})$ in [17, 18]. In fact, Gardner and Derrida considered a simple perceptron ($k=0$) with normalized weights $\boldsymbol{J}_0$ ($\boldsymbol{J}_0\cdot\boldsymbol{J}_0=N$), so that $(\boldsymbol{J}_0^a-\boldsymbol{J}_0^b)^2/2N = 1-\boldsymbol{J}_0^a\cdot\boldsymbol{J}_0^b/N = 1-q^{ab}$ in their notations. We assume replica symmetry, i.e. $R^a=R$, $v_i^a=v_i$, $c_i^{ab}=c_i$ for all a,b. This assumption is valid whenever $\kappa_{\max}\geqslant 0$ (or $f=0$). The parameter R represents trivially the overlap between the first N components of vector $\boldsymbol{J}$ and the teacher $\boldsymbol{K}$. The overlap between $\boldsymbol{J}_i$ and $\boldsymbol{K}$ may be neglected for $i\geqslant 1$, since for odd functions ϕ and orthogonal vectors $\boldsymbol{B}_i$ the new features are uncorrelated. If ϕ were even, this would not be the case. The parameters v_i are proportional to the norm of the $\boldsymbol{J}_i$. The sense of the parameters c_i is more involved. They reflect how fast the fluctuations of $\boldsymbol{J}_i$ around the minimum of the cost function decrease as the temperature vanishes ($\beta\to+\infty$). In the case of a degenerate continuum of minima, these fluctuations decrease too slowly, and the c_i diverge. This is the case for $\kappa<\kappa_{\max}$.

A symmetry between the k vectors $\boldsymbol{J}_i$, $i\geqslant 1$, due to the invariance with respect to permutations of the $\boldsymbol{B}_i$, together with the fact that the $\boldsymbol{B}_i$ are uncorrelated with $\boldsymbol{K}$, allows us to take $v_i=v_1$ and $c_i=c_1$ for $i\geqslant 1$. Introducing $\tilde{v}_1=v_1/v_0$, where v_0 is determined by the normalization condition $\boldsymbol{J}\cdot\boldsymbol{J}/N=k+1=v_0+k\tilde{v}_1v_0$, $\tilde{c}_1=c_1/c_0$ and $\tilde{c}_0=c_0/(1+k)$, the free energy is $f(k,\alpha,\kappa)=\max_{\tilde{v}_1,\tilde{c}_1,\tilde{c}_0}\min_R g(k,\alpha,\kappa;\tilde{v}_1,\tilde{c}_1,\tilde{c}_0,R)$, with

$$\begin{aligned} g(k,\alpha,\kappa;\tilde{v}_1,\tilde{c}_1,\tilde{c}_0,R) = &-\frac{\tilde{c}_1(1-R^2)+k\tilde{v}_1}{2\tilde{c}_0\tilde{c}_1(1+k\tilde{v}_1)} \\ &+\frac{\alpha}{\tilde{c}_0}\int D\lambda_1\cdots\int D\lambda_k\int_{\kappa a-b}^{\kappa a}\mathrm{D}y\,\frac{(\kappa-y/a)^2}{e}H\left(-\frac{yR}{\sqrt{e}}\right) \\ &+2\alpha\int \mathrm{D}\lambda_1\cdots\int D\lambda_k\int_{-\infty}^{\kappa a-b}\mathrm{D}y\,H\left(-\frac{yR}{\sqrt{e}}\right). \end{aligned} \tag{12}$$

$\mathrm{D}y=\mathrm{d}y\,\exp(-y^2/2)/\sqrt{2\pi}$, $H(x)=\int_x^{+\infty}\mathrm{D}y$, and a, b, e represent

$$a=\sqrt{\frac{1+k\tilde{v}_1}{e+R^2}} \tag{13a}$$

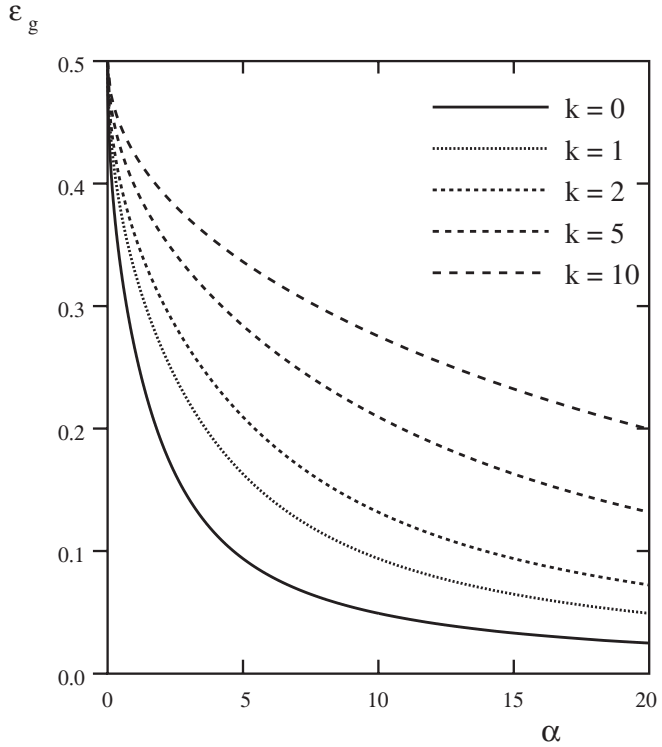

Figure 1. Generalization error ϵ_g in the case of learning an LS task by SVMs with respectively $k = 0, 1, 2, 5$ and 10 new features as a function of the training set size α. The SVMs correspond to the choice of the function $\phi(\lambda) = \text{sign}(\lambda)$ in the mapping to the feature space.

$$b = a\left[\tilde{c}_0\left(1 + \tilde{c}_1 \sum_{i=1}^{k} \phi^2(\lambda_i)\right)\right]^{1/2} \tag{13b}$$

$$e = 1 - R^2 + \tilde{v}_1 \sum_{i=1}^{k} \phi^2(\lambda_i). \tag{13c}$$

The generalization error $\epsilon_g(k, \alpha)$ is written

$$\epsilon_g(k, \alpha) = \frac{1}{\pi} \int \mathrm{D}\lambda_1 \cdots \int \mathrm{D}\lambda_k \ \arccos\left(\frac{R}{\sqrt{e + R^2}}\right) \tag{14}$$

where R and e extremize $g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0, R)$. The maximal stability $\kappa_{\max}(k, \alpha)$ is the largest value of κ that satisfies $\tilde{c}_0(\alpha, \kappa) = +\infty$ since f is non zero for finite values of $\tilde{c}_0$.

If $\phi(\lambda) = \text{sign}(\lambda)$, the extremization of (12) with respect to $\tilde{v}_1$ and $\tilde{c}_1$ gives $\tilde{v}_1 = 1 - R^2$ and $\tilde{c}_1 = 1$. Notice that for $R = 1$ (which corresponds to $\alpha = +\infty$), $\tilde{v}_1 = 0$ (thus, $v_1 = 0$) as expected: the new features are irrelevant because the task is LS. The fact that $\tilde{c}_1 = 1$ means that the fluctuations of $\boldsymbol{J}_0$ and $\boldsymbol{J}_i, i \geqslant 1$, have the same behaviour in the limit $\beta \to +\infty$ despite the fact that their norms are different ($\tilde{v}_1 \neq 1$). After introduction of these values for $\tilde{v}_1$ and $\tilde{c}_1$ in (12), we obtain

$$g(k, \alpha, \kappa; \tilde{c}_0, R) = g\left(0, \alpha/(k+1), \kappa; \tilde{c}_0, R\Big/\sqrt{1 + k(1 - R^2)}\right) \tag{15}$$

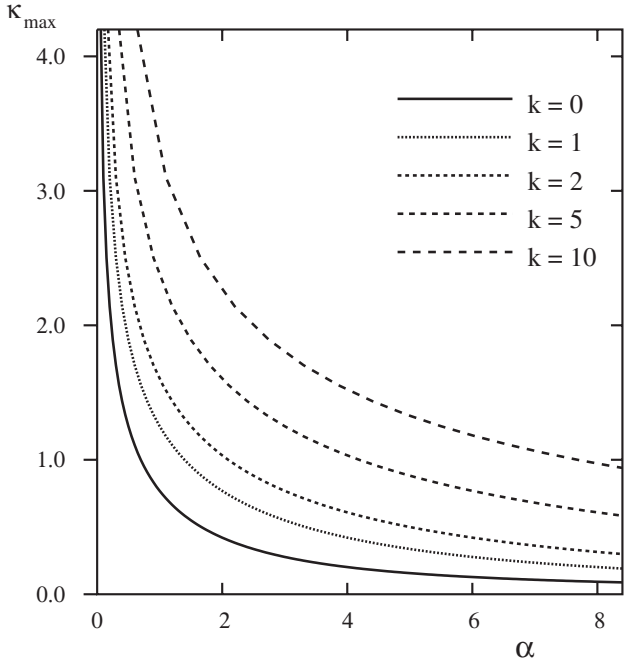

Figure 2. SV margin $\kappa_{\max}$ for the learning of an LS task by SVMs with respectively $k = 0, 1, 2, 5$ and 10 new features as a function of the training set size α. The SVMs correspond to the choice of the function $\phi(\lambda) = \text{sign}(\lambda)$ in the mapping to the feature space.

where the right-hand-side term corresponds to a simple perceptron trained with a training set of reduced size $\alpha/(k+1)$, having an overlap $R/\sqrt{1+k(1-R^2)}$ with the teacher. After introducing these values of the order parameters in (13c) and (14), we obtain $\epsilon_g(k, \alpha) = \epsilon_g(0, \alpha/(k+1))$. In figure 1 we represent the generalization error for the SVM with $k = 0, 1, 2, 5$ and 10 new features as a function of the training set size α. As expected, the generalization error of the SVM with $k > 0$ on an LS task is larger than that of the linear SVM. This is due to an entropic effect, as the SVMs phase space grows with k whereas the size of the space of functions considered, limited to the LS ones, remains the same. For large α, the generalization error vanishes as $0.5005\,(k+1)/\alpha$, to be compared to the linear SVM that has $\epsilon_g \sim 0.5005/\alpha$ [15].

From the above scaling (15), the SV margin is $\kappa_{\max}(k, \alpha) = \kappa_{\max}(0, \alpha/(k+1))$. We can see in figure 2 that the SV margin is an increasing function of the number k of introduced features, for all the training set sizes. This property will have an important effect on the robustness of the SVM against corruption of the patterns. For $\alpha \ll 1$, the SV margin is $\kappa_{\max}(k, \alpha) \sim \sqrt{(k+1)/\alpha}$ and for $\alpha \to +\infty$, $\kappa_{\max}(k, \alpha) \sim 0.226\sqrt{2\pi}(k+1)/\alpha$.

The number of SVs also follows from the distribution of stabilities $\rho(0, \alpha; \gamma)$ of the MSP in input space [15]. We obtain

$$\rho(k, \alpha; \gamma) = \rho_1(\gamma, k, \alpha)\,\Theta(\gamma - \kappa_{\max}) + \rho_0(k, \alpha)\,\delta(\gamma - \kappa_{\max}) \tag{16}$$

where

$$\rho_1(\gamma, k, \alpha) = \sqrt{\frac{2}{\pi}}\,H\left[-\frac{\gamma}{\tan\left(\pi\epsilon_g(k, \alpha)\right)}\right]\exp\left(-\frac{\gamma^2}{2}\right) \tag{17}$$

and $\rho_0(k, \alpha)$, the typical fraction of training patterns that belongs to the set of SVs, is such that

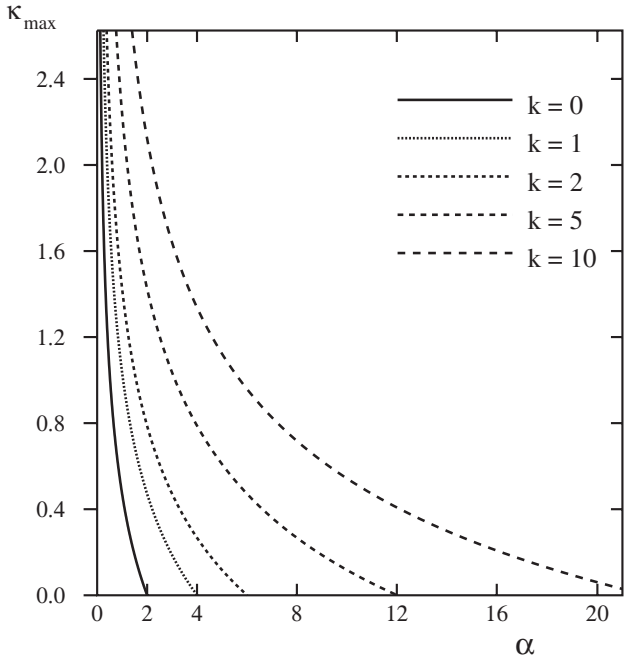

Figure 3. SV margin $\kappa_{\max}$ for the learning of a random task by SVMs with respectively $k = 0, 1, 2, 5$ and 10 new features as a function of the training set size α. The storage capacity $\alpha_c(k)$ corresponds to the training set size with a vanishing SV margin. The SVMs correspond to the choice of the function $\phi(\lambda) = \text{sign}(\lambda)$ in the mapping to the feature space.

$\rho(k, \alpha; \gamma)$ integrates to one. For $\alpha \ll 1$,

$$\rho_0(k, \alpha) \sim 1 - \sqrt{\frac{2\alpha}{\pi(k+1)}} \exp\left(-\frac{k+1}{2\alpha}\right) \tag{18}$$

meaning that in that limit almost all the training patterns are SVs. For $\alpha \to +\infty$,

$$\rho_0(k, \alpha) \sim 0.952\,\frac{k+1}{\alpha}. \tag{19}$$

The fraction of SVs vanishes when $\alpha \to +\infty$. However the typical number of SVs, $P_{\text{SV}} = \rho_0 P \sim 0.952(k+1)N$, is large and only slightly smaller than the feature-space dimension $(k+1)N$. This result is in contradiction with what is observed in applications, where the number of SVs is usually quite small compared to the feature-space dimension.

Vapnik showed that the fraction ρ_0 of SVs is an upper bound of the leave-one-out estimator of the generalization error [1, 2]. Our results show that the typical value of the generalization error as a function of the number of new features is also bounded by the fraction of training patterns that are SVs. In the large-training-set-size limit $\alpha \gg 1$ we find $\rho_0 \sim \epsilon_g \sim (k+1)/\alpha$. Only the prefactor differs (0.952 compared with 0.5005), showing that this bound is fair.

Solutions for other functions ϕ are more complicated, and we were not able to find a closed expression of $\epsilon_g(k, \alpha)$ for all α. It is however possible to show that the function $\phi(\lambda) = \text{sign}(\lambda)$ gives the smallest generalization error at a given k, at least for small α. Most of the properties obtained for this function ϕ remain valid for a general function ϕ. This is the case for the generalization error and the SV margin, which increase with k for a given α. The fact that for

large training set sizes ($\alpha \to \infty$) the number of SVs is close to the feature-space dimension, despite the fact that the fraction of SVs vanishes, is also independent of the function ϕ.

3.2. Random task

We turn now to the more interesting problem of the capacity, defined as the typical number of dichotomies that the SVM may implement, a quantity closely related to the VC dimension of the learning machine [19]. We consider training sets where the patterns' classes are given by a random teacher, that selects outputs +1 and −1 with the same probability 1/2. In this case, the order parameters are (11*b*) and (11*c*). The free energy is $f(k,\alpha,\kappa) = \max_{\tilde{v}_1,\tilde{c}_1,\tilde{c}_0} g(k,\alpha,\kappa;\tilde{v}_1,\tilde{c}_1,\tilde{c}_0)$ where $g(k,\alpha,\kappa;\tilde{v}_1,\tilde{c}_1,\tilde{c}_0)$ is obtained from (12) by setting $R = 0$:

$$g(k,\alpha,\kappa;\tilde{v}_1,\tilde{c}_1,\tilde{c}_0) = -\frac{\tilde{c}_1 + k\tilde{v}_1}{2\,\tilde{c}_0\,\tilde{c}_1\,(1+k\tilde{v}_1)} + \alpha \int \mathrm{D}\lambda_1 \cdots \int \mathrm{D}\lambda_k H(b-\kappa a)$$
$$+\frac{\alpha}{\tilde{c}_0} \int \mathrm{D}\lambda_1 \cdots \int \mathrm{D}\lambda_k \int_{\kappa a - b}^{\kappa a} \mathrm{D}y\ \frac{(\kappa - y/a)^2}{2\,e} \tag{20}$$

where a, b, e represent

$$a = \sqrt{\frac{1+k\tilde{v}_1}{e}} \tag{21a}$$

$$b = a\left[\tilde{c}_0\left(1+\tilde{c}_1\sum_{i=1}^{k}\phi^2(\lambda_i)\right)\right]^{1/2} \tag{21b}$$

$$e = 1+\tilde{v}_1\sum_{i=1}^{k}\phi^2(\lambda_i). \tag{21c}$$

The capacity $\alpha_c(k)$, the largest reduced number of patterns that the machine can learn without errors, corresponds to a vanishing SV margin, i.e. $\kappa_{\max}(k,\alpha_c(k)) = 0$. In this case, the extremae of $g(k,\alpha,0;\tilde{v}_1,\tilde{c}_1,\tilde{c}_0)$ correspond to $\tilde{c}_0(\alpha,\kappa) = +\infty$ and $\tilde{v}_1 = \tilde{c}_1$ for all the possible functions ϕ. This result means that the capacity is $\alpha_c = 2(k+1)$, independently of ϕ, provided that the new features are uncorrelated. This result generalizes to other feature spaces the value deduced by Cover [20] through a geometrical approach, and analysed thoroughly by Mitchison and Durbin [21] in the particular case of quadratic separating surfaces[4].

The phase space of the SVM with k new features has the same dimension as that of a monolayer perceptron with $k+1$ units in the hidden layer and with a fixed Boolean function between the hidden layer and the output. This is the case for the committee machine and the parity machine (where the Boolean functions are respectively the majority and the parity of the hidden units). Thus, it is interesting to compare the corresponding storage capacity. For example, the optimal capacity for the committee machine scales like $k\sqrt{\ln k}$, and that for the parity machine like $k \ln k$, for large k [22, 23]. The capacity of SVMs with k new features is smaller than that of multilayered perceptrons with one hidden layer of $k+1$ neurons. In practice it is not easy to reach the theoretical capacity of multilayered perceptrons, as the algorithm usually used to train them, called backpropagation, may get trapped in metastable states. (Notice, however, that it has recently been shown [8] that an incremental learning algorithm for the parity machine has a capacity close to the optimal, and avoids the main

[4] Notice that the quadratic separating surfaces correspond to an SVM with $\phi(\lambda) = \lambda$ and $k = N$. In this case, our calculation gives $\alpha_c = 2(N+1)$ instead of $N+1$. The difference comes from the correlations between the new features since the cross products $\xi_i\xi_j$ appear twice in the image of $\boldsymbol{\xi}$ by the mapping Φ. Our results are valid for $k \ll N$ when the correlations between new features may be neglected.

drawback of backpropagation.) The relative poor capacity of the SVMs is compensated by the fact that the corresponding learning algorithm is very efficient.

Let us now come back to the SV margin. It turns out that in the case $\phi(\lambda) = \mathrm{sign}(\lambda)$, the maximal stability $\kappa_{\max}(k, \alpha)$ scales trivially with k. The order parameters are $\tilde{v}_1 = \tilde{c}_1 = 1$ so that

$$g(k, \alpha, \kappa; \tilde{c}_0) = g(0, \alpha/(k+1), \kappa; \tilde{c}_0) \tag{22}$$

where the right-hand side corresponds to a simple perceptron of margin κ and reduced training set size $\alpha/(k+1)$ in input space. The maximal stability is thus $\kappa_{\max}(k, \alpha) = \kappa_{\max}(0, \alpha/(k+1))$. From [15] we deduce that for $\alpha \ll 1$,

$$\kappa_{\max}(k, \alpha) \sim \sqrt{\frac{k+1}{\alpha}} \tag{23}$$

and for $\alpha \to \alpha_c^-$,

$$\kappa_{\max}(k, \alpha) \sim \sqrt{\frac{\pi}{8}} \left(\frac{\alpha_c - \alpha}{\alpha_c} \right). \tag{24}$$

If $\phi(\lambda) = \lambda$, the property $\kappa_{\max}(k, \alpha) \sim \kappa_{\max}(0, \alpha/k)$ is valid for $\alpha \ll k$. As $\kappa_{\max}(0, \alpha)$ is a concave decreasing function of α [18], including new features may result in a large increase of the SV margin. We will see in the following section that this property is useful for the robustness of the SVM learning solution.

The number of SVs may also be determined through the distribution of stabilities (16) with

$$\rho_1(\gamma, k, \alpha) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\gamma^2}{2} \right). \tag{25}$$

The fraction of SVs is given by $\rho_0(k, \alpha) = H(-\kappa_{\max}(k, \alpha))$. For $\alpha \ll 1$, $\rho_0(k, \alpha) \sim 1$ and for $\alpha \simeq \alpha_c^-$, $\rho_0(k, \alpha) \sim 1/2$. For all the training set sizes for which a positive SV margin may be found, the number of SVs is a large fraction of the training patterns. This result seems in contradiction with the numerical applications. However, the fact that the fraction of SVs is larger than one-half is not surprising if we consider that this fraction is an upper bound of the generalization error, which in the case of learning a random task is obviously equal to one-half.

4. Robustness or noise-tolerance

In most classification problems we expect that similar patterns belong to the same class. In that case, having a large SV margin may be useful for the generalization performance. In particular, if slightly corrupted versions of the training patterns are presented to the trained SVM, its output should not change.

We consider an SVM that achieved error-free learning on a random task, with an SV margin $\kappa_{\max} > 0$. We assume that the training patterns are corrupted, after the learning process, through $\boldsymbol{\xi}^\mu \to \boldsymbol{\xi}^\mu + \boldsymbol{\eta}^\mu$, where $\boldsymbol{\eta}^\mu$ are randomly distributed vectors with probability distribution

$$P(\boldsymbol{\eta}) = (2\pi\Delta)^{-N/2} \exp\left(-\frac{\boldsymbol{\eta}^2}{2\Delta^2} \right). \tag{26}$$

The parameter Δ represents the amplitude of the perturbation. We will concentrate in the following on small perturbations ($\Delta \ll 1$).

We are interested in the classification error of the SVM on the corrupted patterns, defined as $\epsilon_r(k, \kappa, \Delta) = \sum_\mu (\sigma^\mu(\Delta) - \tau^\mu)^2/(4P)$ where τ^μ is the original pattern's class and $\sigma^\mu(\Delta)$

the SVM's output to the corrupted pattern. The dependence on α is implicitly included through $\kappa = \kappa_{\max}(k, \alpha)$. ϵ_r characterizes the robustness of the SVM with respect to a slight corruption of the training patterns ($\Delta \ll 1$). Input vectors close to a training pattern will be given the same class with probability $1 - \epsilon_r(\kappa, \Delta)$.

In the case of the linear SVM (the simple perceptron), a straightforward calculation gives

$$\epsilon_r(0, \kappa, \Delta) = H(-\kappa)H(\kappa/\Delta) + \int_{\kappa}^{+\infty} \mathrm{D}z\, H(z/\Delta). \tag{27}$$

If the margin is $\kappa = 0$, one-half of the training patterns have zero stability, and $\epsilon_r(0, 0, \Delta) > 1/4$. Thus, any small perturbation results in misclassifications. If $\kappa > 0$, then $\epsilon_r(0, \kappa, \Delta) \sim \exp(-\kappa^2/2\Delta^2)$ for small Δ, which means that if the margin is positive, we expect a small number of misclassifications. We can also notice that the error ϵ_r decreases with increasing stability κ, meaning that a better robustness and a smaller number of misclassifications are obtained by increasing the SV margin.

Consider next the general SVMs, in higher-dimensional feature spaces. If $\phi(\lambda) = \mathrm{sign}(\lambda)$ and $\kappa > 0$,

$$\epsilon_r(k, \kappa, \Delta) \sim \Delta \tag{28}$$

for small Δ. In comparison with the simple perceptron, the robustness of such an SVM is poor. This is due to the discontinuity of the function ϕ: a small perturbation of the input pattern may produce a strong perturbation on its stability. Notice also the absence of dependence upon κ, which means that in this case increasing the SV margin does not help to decrease the error.

In contrast, for continuous functions ϕ, like $\phi(\lambda) = \lambda$, and small Δ,

$$\epsilon_r(k, \kappa, \Delta) \sim \exp\left(-h(k)\frac{\kappa}{\Delta}\right) \tag{29}$$

where $h(k)$ is an increasing function of k. In this case, the error ϵ_r is small and this corresponds to a good robustness. We can notice that, like for the simple perceptron, the robustness increases with the SV margin (since ϵ_r decreases). Thus, continuous functions ϕ are preferable for improving the SVMs robustness or noise tolerance, and a large SV margin also improves the robustness.

5. Discussion and conclusion

We have presented a study of the typical properties of a class of SVMs. We determined, as a function of the number of new features and the number of training patterns, different characteristics of the SVMs like the fraction of SVs and the SV margin, and the robustness against pattern corruption, that give highlights on the behaviour of such SVMs.

In comparison with our mapping, that considered by Dietrich *et al* [11] introduced a normalizing factor for the quadratic features to minimize the importance of the new features compared to the input features. In the case of a teacher corresponding to a quadratic separating surface, the generalization error was shown to present an interesting crossover between learning of the linear and the quadratic features. The effect of this normalizing factor has been studied in great detail by Risau-Gusman and Gordon [12, 13]. One of its effects is to reduce the entropical contribution that leads, in the case of an LS task, to a generalization error in feature space much larger than that of a simple perceptron learning in input space.

We determined the fraction ρ_0 of SVs for the class of SVMs considered; this fraction seems to be not only an upper bound but also a fair estimation for the generalization error. In the case of an LS task, the asymptotic behaviour in the limit of large training set sizes is correctly predicted.

We showed that the SV margin increases with the number of new features. This may explain the good results in real applications. Assuming some continuity in the classification of the patterns, in other words, that two patterns close together belong to the same class with a large probability, the probability of misclassification of training patterns slightly corrupted (or noise tolerance) is then of great importance and directly related to the generalization error. In fact, in actual applications we expect that patterns belonging to different classes be set in clusters well disconnected one from the other. In that case, the separating surface should pass through regions with small density (or probability) of patterns, so that the number of SVs, which are patterns close to the separating surface, is small. This situation is very different from that of the LS task analysed in section 3.1, where the separating hyperplane between the two classes lies in a region where the probability of training patterns is large. In the case of the random task considered in section 3.2 the training patterns belonging to different classes were not well separated in clusters, but randomly mixed. We showed that the noise tolerance or robustness is enhanced by large margins in the case of SVMs with continuous mappings from the input space to the feature space. This explains why maximizing the margin is so important: the probability that the trained SVM will assign the same class to the corrupted and to the original training patterns is enhanced by large margins.

Acknowledgments

We would like to thank S Risau-Gusman for useful discussions. AB is supported by a Marie Curie Fellowship (HPMF-CT-1999-00328).

References

[1] Vapnik V N 1995 *The Nature of Statistical Learning Theory* (New York: Springer)
[2] Vapnik V N 1998 *Statistical Learning Theory* (New York: Wiley)
[3] Cortes C and Vapnik V N 1995 *Mach. Learn.* **20** 273–97
[4] Rumelhart D E, Hinton G E and Williams R J 1986 *Nature* **323** 533–6
[5] Finnoff W 1994 *Neural Comput.* **6** 285–95
[6] Magoulas G D, Vrahatis M N and Androulakis G S 1997 *Neural Networks* **10** 69–82
[7] Torres Moreno J M and Gordon M B 1998 *Neural Comput.* **10** 1017–40
[8] Buhot A and Gordon M B 2000 *J. Phys. A: Math. Gen.* **33** 1713–20
[9] Yoon H and Oh J-H 1998 *J. Phys. A: Math. Gen.* **31** 7771–84
[10] Buhot A and Gordon M B 1999 *Proc. ESANN'99* ed M Verleysen pp 201–6
[11] Dietrich R, Opper M and Sompolinsky H 1999 *Phys. Rev. Lett.* **82** 2975–8
[12] Risau-Gusman S and Gordon M B 2000 Generalization properties of finite size polynomial support vector machines *Phys. Rev.* E **62** 7092–9
[13] Risau-Gusman S and Gordon M B 2000 Hierarchical learning in polynomial support vector machines *Preprint* cond-mat/0010423
[14] Gerl U and Krey F 1997 *J. Physique (France)* **7** 303–27
[15] Gordon M B and Grempel D R 1995 *Europhys. Lett.* **29** 257–62
[16] Opper M and Kinzel W 1996 *Models of Neural Networks* vol 3, ed E Domany, J L van Hemmen and K Shulten (New York: Springer) pp 151–209
[17] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257–70
[18] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271–84
[19] Opper M 1994 *Phys. Rev. Lett.* **72** 2113–6
[20] Cover T M 1965 *IEEE Trans. Electromagn. Compat.* **14** 326–34
[21] Mitchison G J and Durbin R M 1989 *Biol. Cybern.* **60** 345
[22] Monasson R and Zecchina R 1995 *Phys. Rev. Lett.* **75** 2432–5
[23] Xiong Y, Oh J-H and Kwon C 1997 *Phys. Rev.* E **56** 4540–4