ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 201-206

Statistical mechanics of support vector machines

Arnaud Buhot and Mirta B. Gordon *
Départment de Recherche Fondamentale sur la Matière Condensée
CEA-Grenoble, 17 rue des Martyrs,
38054 Grenoble Cedex 9, France

Abstract. We present a theoretical study of the properties of a class of Support Vector Machines within the framework of Statistical Mechanics. We determine their capacity, the margin, the number of support vectors and the distribution of distances of the patterns to the separating hyperplane in feature-space.

1. Introduction

In this paper we investigate the learning properties of Support Vector Machines (SVMs) with the tools of Statistical Mechanics. We restrict to a class of non-linear mappings between the input and the feature-space that include, as a particular case, the quadratic SVM. We consider learning of random input-output relations, to determine the typical capacity (the maximum number of learnable patterns), and learning of tasks that are linearly separable (LS) in input space, to analyse the generalization performance. We find that the capacity is proportional to the feature-space dimension. As long as the training set size remains below the machine's capacity, the margin and the number of SVs increase with the feature-space dimension. The generalization error on LS tasks learned using non-linear feature spaces increases with the complexity of the feature space due to entropic effects. The paper is organized as follows: in section 2, we describe the SVMs and the particular feature-space considered. We present the replica calculation with our main results in section 3 and our conclusions in section 4.

2. The feature-space

We assume that we are given a set of P *training patterns* in a N-dimensional space. The input vectors $\mathbf{x}^\mu$ $(\mu = 1, \cdots, P)$ are supposed to be drawn with a probability density $P(\mathbf{x}) = (2\pi)^{-N/2} \exp\left(-\mathbf{x}^2/2\right)$. Their corresponding classes are $y^\mu = \pm 1$. The aim is to map the input space to a feature space where the training set is LS. We consider the following non-linear mapping:

* also member of CNRS

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 201-206

$$\mathbf{x} \longrightarrow \Phi(\mathbf{x}) \equiv \{\phi_0 \mathbf{x}, \phi_1 \mathbf{x}, \cdots, \phi_k \mathbf{x}\}, \quad (1)$$

where $\phi_0 \equiv 1$, and the ϕ_i $(1 \leq i \leq k)$ are *odd* functions $\phi_i = \phi(\lambda_i)$ of $\lambda_i \equiv \mathbf{x} \cdot \mathbf{B}_i$, with the $\mathbf{B}_i$ $(i = 1, \cdots, k \leq N)$ being a set of k orthonormal vectors $(\mathbf{B}_i \cdot \mathbf{B}_j = \delta_{ij})$. With this choice the *features* are weakly correlated. In the thermodynamic limit considered below, any set of $k \leq N$ randomly selected normalized vectors $\mathbf{B}_i$ satisfies the orthogonality constraint with probability one. Depending on the function ϕ and the vectors $\mathbf{B}_i$, the mapping (1) generates different families of SVMs. If $k = 0$, we have the Maximal Stability Perceptron (MSP), or *linear SVM*, whose properties have been extensively studied (see [1] and references therein). If $k = N$, choosing $\phi(\lambda) = \lambda$ and the input space generators for the $\mathbf{B}_i$ ($\mathbf{B}_i = \mathbf{e}_i$ with $\mathbf{e}_1 = (1, 0, \cdots, 0)$, $\mathbf{e}_2 = (0, 1, \cdots, 0)$, etc.), corresponds to the quadratic SVM. Another choice of theoretical interest is $\phi(\lambda) = \text{sign}(\lambda)$.

The output of the SVM to a pattern $\mathbf{x}$ is $\sigma = \text{sign}\,[\mathbf{w} \cdot \Phi(\mathbf{x})]$, where $\mathbf{w} = \{\mathbf{w}_0, \mathbf{w}_1, \cdots, \mathbf{w}_k\}$ is a $(1+k)N$-dimensional vector. Hereafter we consider normalized weights, $\mathbf{w} \cdot \mathbf{w} = (1+k)N$ without any lack of generality, but we *do not* impose any constraint to the normalization of each N-dimensional vector $\mathbf{w}_i$. We restrict to solutions without threshold.

The aim of learning is to determine a vector $\mathbf{w}$ such that $\sigma^\mu = y^\mu$ or, equivalently, such that

$$\gamma^\mu = \frac{y^\mu \; \mathbf{w} \cdot \Phi(\mathbf{x}^\mu)}{\sqrt{(1+k)N}} \geq 0 \quad \forall \mu. \quad (2)$$

Notice that $|\gamma^\mu|$ is the distance of pattern μ to the hyperplane normal to $\mathbf{w}$. Any vector $\mathbf{w}$ that meets conditions (2) separates linearly in feature-space the images of training patterns with output $+1$ from those with output -1. As we consider solutions without thresholds, the separating hyperplane passes through the origin. Due to the non-linearity of Φ, the separation is not linear in input space. The distance of the training patterns closest to the hyperplane defines the hyperplane's *stability* or *margin* κ. The *Optimal Hyperplane* [2] $\mathbf{w}^*$ has maximal stability, $\kappa_{\max}$:

$$\kappa_{\max}(\mathbf{w}^*) = \max_{\mathbf{w}} \inf_{\mu} \gamma^\mu = \max_{\mathbf{w}} \kappa. \quad (3)$$

i.e. it is the MSP in feature-space. $\mathbf{w}^*$ is a linear combination of the patterns at distance $\kappa_{\max}$ [2, 3], which are the *Support Vectors* (SV): $\mathbf{w}^* = \sum_{\mu \in SV} a^\mu y^\mu \Phi(\mathbf{x}^\mu)$. The a^μ are positive parameters to be determined by the learning algorithm, which has also to find out which patterns are the SVs, whose number P_{sv} is unknown. In [2], the optimal hyperplane is defined as the vector $\widetilde{\mathbf{w}}$ that minimizes $L(\widetilde{\mathbf{w}}) = \widetilde{\mathbf{w}} \cdot \widetilde{\mathbf{w}}$ under the conditions:

$$y^\mu(\widetilde{\mathbf{w}} \cdot \Phi(\mathbf{x}^\mu) + b) \geq 1 \qquad \forall \mu = 1, \cdots, P \quad (4)$$

It is easy to show that $\mathbf{w}^* = \widetilde{\mathbf{w}} \sqrt{(1+k)N / L(\widetilde{\mathbf{w}})}$.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 201-206

3. Replica calculation

We studied the generic properties of the SVMs defined by the mapping (1), through the by now standard replica approach [4]. The 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* $\alpha \equiv P/N$ constant. In this limit, the SVM properties become independent of the particular training set realization, a fact known as self-averaging. The appropriate cost function, whose minimum is the solution to the learning problem, is

$$E(\mathbf{w}, \mathcal{L}_\alpha, \kappa) = \sum_{\mu=1}^{P} \Theta(\kappa - \gamma^\mu) \tag{5}$$

where Θ is the Heaviside step function and $\mathcal{L}_\alpha$ represents the training set. (5) counts the number of training patterns μ that have $\gamma^\mu < \kappa$ in feature-space. The largest value of κ that satisfies $E(\mathbf{w}^*, \mathcal{L}_\alpha, \kappa) = 0$ is the SVM's maximal margin. The weight vector $\mathbf{w}^*$ defines the SVM. Its generic properties are determined by the zero temperature free energy

$$f(k, \alpha, \kappa) = \lim_{N \to +\infty} \lim_{\beta \to +\infty} -\frac{1}{\beta N} \langle \ln Z \rangle, \tag{6}$$

where $Z = \int dP(\mathbf{w}) \exp\left[-\beta E(\mathbf{w}, \mathcal{L}_\alpha, \kappa)\right]$ is the partition function, $dP(\mathbf{w}) = d\mathbf{w}\ \delta\left[(1+k)N - \mathbf{w} \cdot \mathbf{w}\right]$ and β is the inverse temperature. In Eq.(6), the bracket stands for the average over all the possible training sets $\mathcal{L}_\alpha$ for a given α. The free energy (6) is calculated using the replica trick $\langle \ln Z \rangle = \lim_{n \to 0} \ln \langle Z^n \rangle / n$.

We first consider the case of learning a random input-output relation, in which the classes y of the training patterns are randomly selected to be $+1$ or -1 with the same probability $1/2$. In the following, we describe the main steps of the calculation. The reader not interested in these details may jump to the next paragraph, where the main results are presented and discussed. If $f = 0$ for $\kappa \geq 0$, the training set is LS with probability one. The largest value of κ for which $f = 0$ is the *typical* value of $\kappa_{\max}(k, \alpha)$. In this problem, the pertinent order parameters are

$$v_i^a = \frac{\mathbf{w}_i^a \cdot \mathbf{w}_i^a}{N}, \tag{7}$$

$$c_i^{ab} = \lim_{\beta \to +\infty} \beta \frac{(\mathbf{w}_i^a - \mathbf{w}_i^b)^2}{2N} \qquad (a \neq b), \tag{8}$$

where $\mathbf{w}^a$ and $\mathbf{w}^b$ are the weight vectors of replicas a and b respectively, and $i = 0, \cdots, k$. The cross-overlaps $\mathbf{w}_i^a \cdot \mathbf{w}_j^b / N$ $(i \neq j)$ may be neglected, as they are of order $1/\sqrt{N}$ due to the fact that features i and j are uncorrelated [5]. These order parameters generalize to $k \neq 0$ the ones introduced by Gardner

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 201-206

and Derrida [6, 7] in their seminal papers on the single perceptron ($k = 0$) with normalized weights $\mathbf{w}_0^a$ ($\mathbf{w}_0^a \cdot \mathbf{w}_0^a = N$). The parameters c_i^{ab} are a generalization of Gardner-Derrida's parameter $x^{ab} = \lim_{\beta \to +\infty} \beta(1 - q^{ab})$, as $(\mathbf{w}_0^a - \mathbf{w}_0^b)^2/2N = 1 - \mathbf{w}_0^a \cdot \mathbf{w}_0^b/N = 1 - q^{ab}$ in their notations. As we do not impose the norms of the $\mathbf{w}_i^a$ but only the global norm $\mathbf{w} \cdot \mathbf{w} = (1+k)N$, the parameters v_i^a, absent in their formulation, appear naturally here. We assume replica symmetry, *i.e.* $v_i^a = v_i$ and $c_i^{ab} = c_i$ for all a, b. Then, the order parameters have a quite intuitive meaning: the norm of the $\mathbf{w}_i$ do not depend on the replica index, and the c_i reflect how fast the fluctuations of $\mathbf{w}_i$ 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 very slowly, and the c_i diverge. This is the case for $0 \leq \kappa \leq \kappa_{\max}$. The general properties of the SVMs are invariant under permutations of the $\mathbf{B}_i$. This symmetry allows us to take $v_i = v_1$ and $c_i = c_1$ for $i \geq 1$. Introducing $\tilde{v}_1 = v_1/v_0$, where v_0 is determined by the normalization condition $\mathbf{w} \cdot \mathbf{w}/N = 1 + k = v_0 + k\tilde{v}_1 v_0$, and $\tilde{c}_1 = c_1/c_0$, the free energy writes $f(k, \alpha, \kappa) = \max_{\tilde{v}_1, \tilde{c}_1, c_0} g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, c_0)$. The maximization of g determine the values of the order parameters (7), which in turn give the properties of the optimal hyperplane.

The capacity $\alpha_c(k)$ is the largest reduced number of patterns that the machine with k features can learn without errors. At $\alpha = \alpha_c(k)$, the maximal margin vanishes, *i.e.* $\kappa_{\max}(k, \alpha_c(k)) = 0$. In this case, the extrema of $g(k, \alpha, 0; \tilde{v}_1, \tilde{c}_1, c_0)$ correspond to $c_0(\alpha, \kappa) = +\infty$ and $\tilde{v}_1 = \tilde{c}_1$ for all the possible odd functions ϕ. Notice that our assumption of replica symmetry is consistent, as the replica symmetric solution is stable until $c_0(\alpha, \kappa) = +\infty$, or equivalently for $0 \leq \kappa \leq \kappa_{\max}(k, \alpha)$, which is the region where error-free learning is possible. Our result means that the capacity is $\alpha_c = 2(1+k)$, *independently* of the particular choice of ϕ, provided that the new features are uncorrelated. This generalizes to more general feature-spaces the result obtained for quadratic separating surfaces by Cover [8], who found through a geometrical approach that $\alpha_c = 2N$. Quadratic classifiers correspond to SVMs with $\phi(\lambda) = \lambda$ and $k = N$.

Contrary to the capacity, the typical maximal margin depends on the particular mapping ϕ implemented by the SVM. 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 $g(k, \alpha, \kappa; c_0) = (1+k) g(0, \alpha/(1+k), \kappa; c_0)$, where the RHS corresponds to a single perceptron of stability κ in input space. The maximal margin for these mappings is thus given by $\kappa_{\max}(k, \alpha) = \kappa_{\max}(0, \alpha/(1+k))$. From [1] we deduce that for $\alpha \ll 1+k$, $\kappa_{\max}(k, \alpha) \sim \sqrt{(1+k)/\alpha}$, and for $\alpha \to \alpha_c^-$, $\kappa_{\max}(k, \alpha) \sim \sqrt{\pi/8}\ (\alpha_c - \alpha)/\alpha_c$. Although we were unable to find a closed form of the maximal margin for the mapping $\phi(\lambda) = \lambda$, the property that $\kappa_{\max}(k, \alpha) \sim \kappa_{\max}(0, \alpha/k)$ is verified for $k \gg \max(\alpha, 1)$. More generally, as $\kappa_{\max}(0, \alpha)$ is a concave decreasing function of α [7], it is possible to increase the margin by including new features, *i.e.*, by increasing k.

The typical fraction of training patterns that are SVs, $\rho_{sv}(k, \alpha) = P_{sv}/P$, is a quantity of great importance, since Vapnik [2] showed that it is an upper bound to the generalization error ϵ_g, the probability of making a mistake in the classification of a new pattern (see Theorem 5.2 p.135 in [2]). We determine the distribution of distances (2) of the patterns to the optimal hyperplane, $\rho(k,\alpha;\gamma)$, which has a delta peak at the position of the SVs, whose weight is exactly ρ_{sv}. In fact, $\rho(k,\alpha;\gamma)$ follows from the MSP's distribution [1], $\rho(0,\alpha;\gamma)$. The dependence on the mapping $\Phi(\lambda)$ is implicit in $\kappa_{\max}(k,\alpha)$. We obtain

$$\rho(k,\alpha;\gamma) = \frac{\exp\left(-\gamma^2/2\right)}{\sqrt{2\pi}}\,\Theta\left[\gamma-\kappa_{\max}\right] + \rho_{sv}(k,\alpha)\,\delta\left[\gamma-\kappa_{\max}\right] \tag{9}$$

where $\rho_{sv}(k,\alpha)$ is such that $\rho(k,\alpha;\gamma)$ integrates to one. For $\alpha \ll 1+k$, $\rho_{sv}(k,\alpha) \sim 1-\sqrt{\alpha/2\pi(1+k)}\exp(-(1+k)/2\alpha)$, meaning that in the limit of very small α almost all the training patterns are SVs. $\rho_{sv}(k,\alpha)$ decreases with increasing α. For $\alpha \to \alpha_c^-$, $\rho_{sv}(k,\alpha) \to 1/2$, *i.e.* when the reduced training set size gets close to the capacity, one half of the training patterns are SVs. Since $\alpha_c = 2(1+k)$, in this limit the number of SVs is equal to the feature-space dimension. In the case of learning a random task, the fact that $\rho_{sv}(k,\alpha) > 1/2$ is consistent with Vapnik's bound, since the generalization is impossible and $\epsilon_g = 1/2$.

In the following, we consider that the learned task is LS in input space. Despite the fact that this task is too simple to be representative of realistic applications, its study provides insight on the properties of SVMs in a case where the number of SV should be meaningful for predicting the generalization error. The calculation of the free energy (6) in this case is somewhat more involved than in the random task, as it includes a new order parameter besides those in equations (7) and (8). We do not detail here the calculations, but present the main results. For $\alpha \ll 1+k$, the behavior of $\kappa_{\max}(k,\alpha)$ and $\rho_{sv}(k,\alpha)$ are the same as for random outputs. This is not surprising, since, at small α, the SVM does not have enough information to realize that the task is LS. More interesting is the behavior for $\alpha \gg 1+k$. In this case, $\kappa_{\max}(k,\alpha) \sim 0.226\sqrt{2\pi}(1+k)/\alpha$, $\rho_{sv}(k,\alpha) \sim 0.952(1+k)/\alpha$, and $\epsilon_g(k,\alpha)$ vanishes as $0.5005\,(1+k)/\alpha$. Thus, the typical number of SVs is only slightly smaller than the feature-space dimension. It is interesting to notice that the bound given by Vapnik for the generalization error is in good agreement with our results. As a linear SVM learning a LS task has $\epsilon_g(0,\alpha) \sim 0.5005/\alpha$ [1], we see that the overfitting arising when the task is learnt by too complex machines ($k \neq 0$) produces an increase in the generalization error and in the number of SVs which is proportional to the number of superfluous parameters. In other words, the SVM is unable to find the solution without quadratic components, *i.e.* with $\mathbf{w}_i = 0$ for $i \neq 0$. This is an entropic effect and is expected to arise whenever the mapping defining the feature space is more complex than the task to be learned.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 201-206

4. Conclusions

We have presented the typical properties of a general class of Support Vector Machines. The first result obtained for this type of SVMs is the capacity. This capacity, strongly related to the VC dimension, is shown to be proportional to the feature-space dimension, generalizing Cover's well known result for quadratic feature-spaces. Our second result shows that the SV-margin and the number of Support Vectors both increase with the feature space dimension. This behaviour is valid in both cases considered: learning a random task and a task that is LS in the input space. The fact that the SV-margin increases with the feature space dimension is not surprising. Moreover, it can be shown that it increases the robustness of the solution with respect to input noise.

In real applications, it is commonly observed that the number of SVs saturates when the size of the feature space increases. This is different from what we find for a random task and a LS task. One reason for this desagreement may come from the unrealistic distribution of training patterns considered. In the particular case of the Linearly Separable task, the gaussian distribution considered in this calculation has a large probability that points lie on the separating surface. In realistic applications, we expect the points to be distributed around prototypes, each prototype corresponding to a given class, and with small overlaps between the distributions around different prototypes.

References

[1] M. B. Gordon and D. R. Grempel, *Europhys. Lett.*, **29**, 257-262 (1995).

[2] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York (1995).

[3] U. Gerl and F. Krey, *J. Phys. I France*, **7**, 303-327 (1997).

[4] M. Opper and W. Kinzel, in *Models of Neural networks III*, E. Domany, J. L. van Hemmen and K. Shulten (Eds.), pp. 151-209 (Springer Verlag, New York, 1996).

[5] As $(1+k)N \equiv \mathbf{w} \cdot \mathbf{w}$ and $\langle \mathbf{w} \cdot \mathbf{w} \rangle = (1+k)N \sum_{\mu \in SV} (a^\mu)^2$, then typically $(a^\mu)^2 \sim P_{sv}^{-1}$. It follows that $\langle (\mathbf{w}_i \cdot \mathbf{w}_j)^2 \rangle \sim N(1 + N/P_{sv})$, for all $k < N$. Thus, $\langle \mathbf{w}_i \cdot \mathbf{w}_j \rangle / N \sim 1/\sqrt{N}$ is negligible.

[6] E. Gardner, *J. Phys. A: Math. Gen.*, **21**, 257-270 (1988).

[7] E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.*, **21**, 271-284 (1988).

[8] T. M. Cover, *IEEE Trans. Elect. Comp.*, **14**, 326-334 (1965).