Margin Distribution Bounds on Generalization

John Shawe-Taylor
Royal Holloway, University of London
j.shawe-taylor@dcs.rhbnc.ac.uk

Nello Cristianini
University of Bristol
nello.cristianini@bristol.ac.uk

NeuroCOLT2 Technical Report Series

NC2-TR-1998-020

July, 1998[1]

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

For more information see the NeuroCOLT website
http://www.neurocolt.com
or email neurocolt@neurocolt.com

[1] Received 17-JUL-1998; revised 18-SEP-1998

Abstract

A number of results have bounded generalization of a classifier in terms of its margin on the training points. There has been some debate about whether the minimum margin is the best measure of the distribution of training set margin values with which to estimate the generalization. Freund and Schapire [6] have shown how a different function of the margin distribution can be used to bound the number of mistakes of an on-line learning algorithm for a perceptron, as well as an expected error bound. We show that a slight generalization of their construction can be used to give a pac style bound on the tail of the distribution of the generalization errors that arise from a given sample size.

1 Introduction

The idea that a large margin classifier might be expected to give good generalization is certainly not new [5, 12]. Despite this insight it was not until comparatively recently [10] that such a conjecture has been placed on a firm footing in the probably approximately correct (pac) model of learning. Learning in this model entails giving a bound on the generalization error which will hold with high confidence over randomly drawn training sets. In this sense it can be said to ensure robust learning, something that cannot be guaranteed by bounds on the expected error of a classifier.

Despite successes in extending this style of analysis to the agnostic case [1] and applying it to neural networks [1], boosting algorithms [9] and Bayesian algorithms [4], there has been concern that the measure of the distribution of margin values attained by the training set is largely ignored in a bound that depends only on its minimal value. Intuitively, there appeared to be something lost with a bound that depended so critically on the positions of possibly a small proportion of the training set, ignoring the margin attained by the majority of the points.

Freund and Schapire [6] (a similar technique was employed by Klasner and Simon [8] for rendering a real valued function learning algorithm noise tolerant) developed a measure of the margin distribution which they showed could be used to bound the expected generalization error more tightly than the minimal margin. The aim of this paper is to show that the same measure can also be used to provide a pac style bound on the generalization error. We will also develop an algorithm for a modified Kernel based linear machine which directly optimises the new measure.

2 Background Results

We first give some necessary definitions.

Definition 2.1 *Let H be a set of binary valued functions. We say that a set of points X is* shattered by *H if for all binary vectors b indexed by X, there is a function $f_b \in H$ realising b on X. The* Vapnik-Chervonenkis (VC) dimension,

$\mathrm{VCdim}(H)$, *of the set H is the size of the largest shattered set, if this is finite or infinity otherwise.*

Definition 2.2 *Let H be a set of real valued functions. We say that a set of points X is* γ-shattered by H *if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b indexed by X, there is a function $f_b \in H$ satisfying*

$$f_b(x) \begin{cases} \geq r_x + \gamma & \textit{if } b_x = 1 \\ \leq r_x - \gamma & \textit{otherwise.} \end{cases}$$

The fat shattering dimension fat_H *of the set H is a function from the positive real numbers to the integers which maps a value γ to the size of the largest γ-shattered set, if this is finite or infinity otherwise.*

We will make critical use of the following result contained in Shawe-Taylor et al [10] which involves the fat shattering dimension of the space of functions.

Theorem 2.3 *Consider a real valued function class $\mathcal{H}$ having fat shattering function bounded above by the function* $\mathrm{afat} : \Re \rightarrow \mathcal{N}$ *which is continuous from the right. Fix $\theta \in \Re$. Then with probability at least $1 - \delta$ a learner who correctly classifies m independently generated examples $\mathbf{z}$ with $h = T_\theta(f) \in T_\theta(\mathcal{H})$ such that $\mathrm{er}_{\mathbf{z}}(h) = 0$ and $\gamma = \min |f(\mathbf{x}_i) - \theta|$ will have error of h bounded from above by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log_2\left(\frac{8em}{k}\right) \log_2(32m) + \log_2\left(\frac{8m}{\delta}\right)\right),$$

where $k = \mathrm{afat}(\gamma/8) \leq em$.

Note how the fat shattering dimension at scale $\gamma/8$ plays the role of the VC dimension in this bound. This result motivates the use of the term effective VC dimension for this value. In order to make use of this theorem, we must have a bound on the fat shattering dimension and then calculate the margin of the classifier. We begin by considering bounds on the fat shattering dimension. The first bound on the fat shattering dimension of bounded linear functions in a finite dimensional space was obtained by Shawe-Taylor *et al.* [10]. Gurvits [7] generalised this to infinite dimensional Banach spaces. We will quote an improved version of this bound for Hilbert spaces which is contained in [2] (slightly adapted here for an arbitrary bound on the linear operators).

Theorem 2.4 *[2] Consider a Hilbert space and the class of linear functions L of norm less than or equal to B restricted to the sphere of radius R about the origin. Then the fat shattering dimension of L can be bounded by*

$$\mathrm{fat}_L(\gamma) \leq \left(\frac{BR}{\gamma}\right)^2.$$

In order to apply Theorems 2.3 and 2.4 we need to bound the radius of the sphere containing the points and the norm of the linear functionals involved. Clearly, scaling by these quantities will give the margin appropriate for application of the theorem.

3 Main Result

Let X be a Hilbert space. We define the following Hilbert space derived from X.

Definition 3.1 *Let $L_f(X)$ be the set of real valued functions f on X with support $\operatorname{supp}(f)$ finite, that is functions in $L_f(X)$ are non-zero only for finitely many points. We define the inner product of two functions $f, g \in L_f(X)$, by*

$$\langle f \cdot g \rangle = \sum_{x \in \operatorname{supp}(f)} f(x)g(x).$$

Note that the sum which defines the inner product is well-defined since the functions have finite support. Clearly the space is closed under addition and multiplication by scalars.

Now for any fixed $\Delta > 0$ we define an embedding of X into the Hilbert space $X \times L_f(X)$ as follows.

$$\tau_\Delta : x \mapsto X_\Delta = (x, \Delta\delta_x),$$

where $\delta_x \in L_f(X)$ is defined by

$$\delta_x(y) = \begin{cases} 1; & \text{if } y = x; \\ 0; & \text{otherwise.} \end{cases}$$

We begin by considering the case where Δ is fixed. In practice we wish to choose this parameter in response to the data. In order to obtain a bound over different values of Δ it will be necessary to apply the following theorem several times.

For a linear classifier $\mathbf{u}$ on X and threshold $b \in \Re$ we define

$$d((\mathbf{x}, y), (\mathbf{u}, b), \gamma) = \max\{0, \gamma - y(\langle \mathbf{u} \cdot \mathbf{x} \rangle - b)\}.$$

This quantity is the amount by which $\mathbf{u}$ fails to reach the margin γ on the point $(\mathbf{x}, y)$ or 0 if its margin is larger than γ. Similarly for a training set S, we define

$$D(S, (\mathbf{u}, b), \gamma) = \sqrt{\sum_{(\mathbf{x},y) \in S} d((\mathbf{x}, y), (\mathbf{u}, b), \gamma)^2}.$$

Theorem 3.2 *Fix $\Delta > 0$, $b \in \Re$. Consider a fixed but unknown probability distribution on the input space X with support in the ball of radius R about the origin. Then with probability $1 - \delta$ over randomly drawn training sets S of size m for all $\gamma > 0$ the generalization of a linear classifier $\mathbf{u}$ on X thresholded at b is bounded by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log_2\left(\frac{8em}{k}\right) \log_2(32m) + \log_2\left(\frac{8m}{\delta}\right)\right),$$

where

$$k = \left\lfloor \frac{64.5(R^2 + \Delta^2)(\|\mathbf{u}\|^2 + D(S, (\mathbf{u}, b), \gamma)^2/\Delta^2)}{\gamma^2} \right\rfloor,$$

provided $m \geq 2/\epsilon$ and $k \leq em$.

Proof: Consider the fixed mapping τ_Δ and the augmented linear functional over the space $X \times L_f(X)$,

$$\hat{\mathbf{u}} = \left(\mathbf{u}, \frac{1}{\Delta}\sum_{(\mathbf{x},y)\in S} d((\mathbf{x},y),(\mathbf{u},b),\gamma)y\delta_{\mathbf{x}}\right).$$

We claim that

1. for $\mathbf{x} \notin S$, $\langle \mathbf{u}\cdot\mathbf{x}\rangle = \langle \hat{\mathbf{u}}\cdot\tau_\Delta(\mathbf{x})\rangle$, and
2. the margin of $\hat{\mathbf{u}}$ with threshold b on the training set $\tau_\Delta(S)$ is γ.

Hence, the behaviour of the linear classifier $(\mathbf{u},b)$ can be characterised by the behaviour of $(\hat{\mathbf{u}},b)$, while $(\hat{\mathbf{u}},b)$ is a large margin classifier in the space $X\times L_f(X)$. Since for $x \in S$, $\|\tau(\mathbf{x})\|^2 \leq R^2+\Delta^2$ and $\|\hat{\mathbf{u}}\|^2 = \|\mathbf{u}\|^2 + D(S,(\mathbf{u},b),\gamma)^2/\Delta^2$, the result will then follow from an application of Theorems 2.3 and 2.4. Note that we have replaced the constant 64 by 64.5 to ensure the continuity from the right required by Theorem 2.3.

1. The first claim follows immediately from the observation that for $\mathbf{z}\notin S$,

$$\left\langle \sum_{(\mathbf{x},y)\in S} d((\mathbf{x},y),(\mathbf{u},b),\gamma)y\delta_{\mathbf{x}}\cdot\delta_{\mathbf{z}}\right\rangle = 0.$$

2. For $(\mathbf{x}',y')\in S$, we have

$$\begin{aligned} y'(\langle\hat{\mathbf{u}},\tau_\Delta(\mathbf{x}')\rangle - b) &= y'(\langle\mathbf{u},\mathbf{x}'\rangle - b) + y'\left\langle \sum_{(\mathbf{x},y)\in S} d((\mathbf{x},y),\mathbf{u},\gamma)y\delta_{\mathbf{x}}\cdot\delta_{\mathbf{x}'}\right\rangle \\ &\geq \gamma - d((\mathbf{x}',y'),\mathbf{u},\gamma) + d((\mathbf{x}',y'),\mathbf{u},\gamma) = \gamma. \end{aligned}$$

The theorem follows. ∎

We now apply this theorem several times to allow a choice of Δ which approximately minimises the expression for k. Note that the minimum of the expression (ignoring the constant and suppressing the denominator γ^2) is $(R+D)^2$ attained when $\Delta = \sqrt{RD}$.

Theorem 3.3 *Fix $b\in\Re$. Consider a fixed but unknown probability distribution on the input space X with support in the ball of radius R about the origin. Then with probability $1-\delta$ over randomly drawn training sets S of size m for all $\gamma>0$ such that $d((\mathbf{x},y),(\mathbf{u},b),\gamma)=0$, for some $(\mathbf{x},y)\in S$, the generalization of a linear classifier $\mathbf{u}$ on X satisfying $\|\mathbf{u}\|\leq 1$ is bounded by*

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k\log_2\left(\frac{8em}{k}\right)\log_2(32m) + \log_2\left(\frac{2m(28+\log_2(m))}{\delta}\right)\right),$$

where

$$k = \left\lfloor \frac{65[(R+D)^2 + 2.25RD]}{\gamma^2}\right\rfloor,$$

for $D = D(S,(\mathbf{u},b),\gamma)$, and provided $m\geq \max\{2/\epsilon, 6\}$ and $k\leq em$.

Proof: Consider a fixed set of values for Δ, $\Delta_1 = R\lfloor 2m^{0.25} - 1 \rfloor$, $\Delta_{i+1} = \Delta_i/2$, for $i = 2, \dots, t$, where t satisfies, $R/32 \geq \Delta_t > R/64$. Hence, $t \leq \log_2(128m^{0.25}) = 7 + 0.25\log_2(m)$. We apply Theorem 3.2 for each of these values of Δ, using $\delta' = \delta/t$ in each application. For a given value of γ and $D = D(S, \mathbf{u}, \gamma)$, it is easy to check that the value of k is minimal for $\Delta = \sqrt{RD}$ and is monotonically decreasing for smaller values of Δ and monotonically increasing for larger values. Note that $\sqrt{RD} \leq R\sqrt{2\sqrt{m-1}}$, as the largest absolute difference in the values of the linear function on two training points is $2R$ and since $d((\mathbf{x}, y), (\mathbf{u}, b), \gamma) = 0$, for some $(\mathbf{x}, y) \in S$, we must have $d((\mathbf{x}', y'), (\mathbf{u}, b), \gamma) \leq 2R$, for all $(\mathbf{x}', y') \in S$. Hence, as $2m^{0.25} - 1 > \sqrt{2}(m-1)^{0.25}$ for $m \geq 6$, we can find a value of Δ_i satisfying

$$\sqrt{RD}/2 \leq \Delta_i \leq \sqrt{RD},$$

provided $\sqrt{RD} \geq R/32$. The value of the expression

$$(R^2 + \Delta^2)(1 + D(S, \mathbf{u}, \gamma)^2/\Delta^2)$$

at the value Δ_i will be upper bounded by its value at $\Delta = \sqrt{RD}/2$. A routine calculation confirms that for this value of Δ, the expression is equal to $(R+D)^2 + 2.25RD$. Now suppose $\sqrt{RD} < R/32$. In this case we will show that

$$(R^2 + \Delta_t^2)(1 + D^2/\Delta_t^2) \leq \frac{130}{129}\left\{(R+D)^2 + 2.25RD\right\},$$

so that the application of Theorem 3.2 with $\Delta = \Delta_t$ covers this case once the constant 64.5 is replaced by 65. Recall that $R/32 \geq \Delta_t > R/64$ and note that $\sqrt{D/R} < 1/32$. We therefore have

$$\begin{aligned}
(R^2 + \Delta_t^2)(1 + D^2/\Delta_t^2) &\leq R^2(1 + 1/32^2)(1 + 64^2 D^2/R^2) \\
&\leq R^2\left(1 + \frac{1}{1024}\right)\left(1 + \frac{64^2}{32^4}\right) \\
&\leq R^2\left(1 + \frac{1}{1024}\right)\left(1 + \frac{1}{256}\right) \\
&< \frac{130}{129}R^2 \\
&\leq \frac{130}{129}\left\{(R+D)^2 + 2.25RD\right\}
\end{aligned}$$

as required. The result follows. ∎

4 Algorithmics

Theorem 3.3 suggests a different learning goal from the maximal margin hyperplane sought by the Support Vector Machine [3]. We should instead seek to minimise $D(S, (\mathbf{u}, b), \gamma)$ for a given fixed value of γ and subsequently minimise over different choices of γ. Vapnik has posed this problem in a slightly more general form [11, Section 5.5.1] as follows.

For non-negative variables $\xi_i \geq 0$, we minimise the function

$$F_\sigma(\xi) = \sum_{j=1}^{m} \xi_j^\sigma,$$

subject to the constraints:

$$\begin{aligned} y_j[\langle \mathbf{u} \cdot \mathbf{x}_j \rangle - b] &\geq 1 - \xi_j, \quad j = 1, \ldots, m \qquad (1) \\ \langle \mathbf{u} \cdot \mathbf{u} \rangle &\leq C. \qquad (2) \end{aligned}$$

He is most interested in values of σ close to 0 when F approximates the number of training set errors. If, however, we take $\sigma = 2$ and make the constraint (2) an equality constraint, the problem corresponds exactly to minimising $D(S, (\mathbf{u}, b), \gamma)$, where $\gamma = 1/\sqrt{C}$. This follows from considering the hyperplane $(\mathbf{u}', b') = (\mathbf{u}/\sqrt{C}, b/\sqrt{C})$ which has norm 1 and classifies the point $(\mathbf{x}_j, y_j)$ such that $d((\mathbf{x}_j, y_j), (\mathbf{u}', b'), \gamma) = \xi_j/\sqrt{C}$, so that

$$D(S, (\mathbf{u}', b'), \gamma) = \sqrt{F_2(\xi)/C}.$$

We now consider converting to the dual problem by introducing Lagrange multipliers α_0 for constraint (2) and $\alpha_j \geq 0$, $j = 1, \ldots, m$, for constraints (1). Setting the derivatives to zero and solving for $\mathbf{u}$ gives

$$\mathbf{u} = \frac{1}{2\alpha_0} \sum_{j=1}^{m} \alpha_j y_j \mathbf{x}_j.$$

Substituting into the other expressions and simplifying results in the following Lagrangian,

$$F(\alpha_0, \alpha) = -\frac{1}{4} \sum_{j=1}^{m} \alpha_j^2 + \sum_{j=1}^{m} \alpha_j - \frac{1}{4\alpha_0} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle - \alpha_0 C,$$

which must be maximised subject to the constraints, $\alpha_j \geq 0$, $j = 0, \ldots, m$, and

$$\sum_{j=1}^{m} \alpha_j y_j = 0.$$

It is convenient to use vector notation, with α denoting the vector of α_j, $j = 1, \ldots, m$, G the matrix with entries, $G_{ij} = y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$, and $\mathbf{1}$ the m vector with entries equal to 1. Using this notation we can write

$$F(\alpha_0, \alpha) = -\frac{1}{4} \alpha^T \alpha + \mathbf{1}^T \alpha - \frac{1}{4\alpha_0} \alpha^T G \alpha - \alpha_0 C.$$

We can optimise with respect to α_0 by computing $\frac{\partial F}{\partial \alpha_0}$ and setting it equal to zero.

$$\frac{\partial F(\alpha_0, \alpha)}{\partial \alpha_0} = \frac{1}{4\alpha_0^2} \alpha^T G \alpha - C = 0.$$

Hence, $\alpha_0 = \sqrt{\frac{1}{4C}\alpha^T G\alpha}$ and resubstituting

$$F(\alpha) = F(\alpha_0, \alpha) = -\frac{1}{4}\alpha^T\alpha + \mathbf{1}^T\alpha - \sqrt{C\alpha^T G\alpha} \quad (3)$$

$$\mathbf{u} = \sqrt{\frac{C}{\alpha^T G\alpha}}\sum_{j=1}^{m}\alpha_j y_j \mathbf{x}_j \quad (4)$$

Note that we can ignore the constant factor in the formula for $\mathbf{u}$ as this will not affect the classification, and in fact $\alpha^T G\alpha = ||\mathbf{u}||^2 = C$ once the optimal value has been found. The value of b can also be determined from the values of α. We wish to confirm that this optimisation problem is concave. We can evaluate the Hessian $H(F)$ of the function F as follows:

$$\begin{aligned} \mathrm{grad}(F) &= -\frac{1}{2}\alpha + \mathbf{1} - \frac{\sqrt{C}G\alpha}{\sqrt{\alpha^T G\alpha}}. \\ \text{Hence} \quad H(F) &= -\frac{1}{2}I - \frac{\sqrt{C}[(\alpha^T G\alpha)G - G\alpha\alpha^T G]}{(\alpha^T G\alpha)^{1.5}}. \end{aligned}$$

We wish to verify that $H(F)$ is concave, that is $\mathbf{x}^T H(F)\mathbf{x} \leq 0$ for all $\mathbf{x}$.

$$\mathbf{x}^T H(F)\mathbf{x} = -0.5||\mathbf{x}||^2 - C'[||\alpha||_G^2||\mathbf{x}||_G^2 - \langle\mathbf{x}\cdot\alpha\rangle_G^2]$$

where C' is a positive constant and $\langle .\cdot.\rangle_G$ and $||.||_G$ are the inner product and norm defined by the semi-definite matrix G. By the Cauchy-Schwartz inequality the expression in square brackets is non-negative, making the overall expression negative as required. Hence, the optimal solution can be found in polynomial time by applying a gradient based central path algorithm following grad(F) with an appropriate learning rate η.

Note further that a small change in $\gamma > 0$ only changes the value of $D(S, (\mathbf{u}, b), \gamma)$ by a small amount for a fixed $(\mathbf{u}, b)$. Hence, the optimal value of k can also only change by a small amount. Hence, solving the problem for a fine enough grid of values of γ and choosing the value which minimises k will give a value which will be within an arbitrarily small margin of the overall optimum.

Finally, note that the computation described in equation (3) can be performed using a Kernel inner product in place of the input space inner product, the technique that is used in the Support Vector Machine.

5 Conclusion

We have shown how an approach developed by Freund and Schapire [6] for mistake bounded learning can be adapted to give pac style bounds which depend on the margin distribution rather than the margin of the closest point to the hyperplane. The bounds obtained can be significantly better than previously obtained bounds, particularly when some of the points are misclassified and agnostic bounds would need to be applied were a classical analysis to be adopted.

We have gone on to show how the measure of the margin distribution that appears in the bound can be optimised by expressing the optimisation problem as a concave dual problem. This formulation also allows the problem to be solved in Kernel spaces such as those used with the Support Vector Machine.

We believe that this paper presents the first pac style bound for a margin distribution measure that is neither critically dependent on the nearest points to the hyperplane nor is an agnostic version of that approach. In addition, we believe it is the first paper to give a provably optimal algorithm for agnostic learning with hyperplanes, by showing that the criterion to optimized should not be the number of errors, but rather a more flexible criterion which could be termed a soft margin.

References

[1] Peter Bartlett, Pattern Classification in Neural Networks, IEEE Transactions on Information Theory, to appear.

[2] Peter Bartlett and John Shawe-Taylor, Generalization Performance of Support Vector Machines and Other Pattern Classifiers, *In* 'Advances in Kernel Methods - Support Vector Learning', Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.), MIT Press, Cambridge, USA, 1998.

[3] C. Cortes and V. Vapnik, Support-Vector Networks, *Machine Learning*, 20(3):273-297, September 1995

[4] Nello Cristianini, John Shawe-Taylor, and Peter Sykacek, Bayesian Classifiers are Large Margin Hyperplanes in a Hilbert Space, in Shavlik, J., ed., *Machine Learning: Proceedings of the Fifteenth International Conference*, Morgan Kaufmann Publishers, San Francisco, CA.

[5] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, New York: Wiley, 1973.

[6] Yoav Freund and Robert E. Schapire, Large Margin Classification Using the Perceptron Algorithm, Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 1998.

[7] Leonid Gurvits, A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In *Proceedings of Algorithm Learning Theory, ALT-97*, and as NECI Technical Report, 1997.

[8] Norbert Klasner and Hans Ulrich Simon, From Noise-Free to Noise-Tolerant and from On-line to Batch Learning, *Proceedings of the Eighth Annual Conference on Computational Learning Theory, COLT'95*, 1995, pp. 250–257.

[9] R. Schapire, Y. Freund, P. Bartlett, W. Sun Lee, Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. In D.H. Fisher,

Jr., editor, *Proceedings of International Conference on Machine Learning, ICML'97*, pages 322–330, Nashville, Tennessee, July 1997. Morgan Kaufmann Publishers.

[10] John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, Martin Anthony, Structural Risk Minimization over Data-Dependent Hierarchies, to appear in *IEEE Trans. on Inf. Theory*, and NeuroCOLT Technical Report NC-TR-96-053, 1996.
(`ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports`).

[11] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[12] Vladimir N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, 1982.

[13] Vladimir N. Vapnik, Esther Levin and Yann Le Cunn, Measuring the VC-dimension of a learning machine, *Neural Computation,* 6:851–876, 1994.