Robust Bounds on Generalization from the Margin Distribution

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-1998-029

October, 1998[1]

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

[1]Received 29-OCT-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 [8] 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. Algorithms arising from the approach are related to those of Cortes and Vapnik [5]. We generalise the basic result to function classes with bounded fat-shattering dimension and the 1-norm of the slack variables which gives rise to Vapnik's box constraint algorithm. We also extend the results to the regression case and obtain bounds on the probability that a randomly chosen test point will have error greater than a given value. The bounds apply to the ϵ-insensitive loss function proposed by Vapnik for Support Vector Machine regression. A special case of this bound gives a bound on the probabilities in terms of the least squares error on the training set showing a quadratic decline in probability with margin.

1 Introduction

For classification by thresholding a real valued function the margin of a training point is the amount by which its output is on the right side of the threshold or, if misclassified, minus the amount by which it fails to be correctly classified. In the case of linear hyperplanes with unit weight vectors, this value can also be seen as the distance of the input point from the hyperplane. The margin of a classifier is the minimum margin over the training set.

The idea that a large margin classifier might be expected to give good generalization is certainly not new [7, 19]. Despite this insight it was not until comparatively recently [12] 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 reliable 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 [2] and applying it to neural networks [2], boosting algorithms [11], perceptron decision trees [13] and Bayesian algorithms [6], 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. Attempts to address this problem have been made in for example [11], but they treat points that fail to meet the larger margin as errors and fall back on agnostic bounds for the generalization error. In contrast our results apply to the case where there are training set errors, but have the form of bounds with no training set errors.

The question of how to handle the situation of non linearly separable data has received a lot of attention (see [4] for a review of some of the methods suggested). The problem is that minimising the number of training errors is NP-complete and so the various methods adopted are inherently heuristic relative to the best bounds previously available for bounding the generalization error. By showing that the generalization error can be bounded in terms of a quantity that can be optimized by a polynomial time algorithm, we provide a solution to a long-standing conundrum of perceptron learning.

The analysis is based on work of Freund and Schapire [8] (a similar technique was employed by Klasner and Simon [10] for rendering a real valued function learning algorithm noise tolerant), who 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. In this paper we show that the same measure can be used to obtain a pac style bound for linear functions. This result provides a formal justification for the soft margin heuristic introduced by Vapnik to render Support Vector machines noise-tolerant [18]. The same theoretical approach is then applied to more general non-linear classes of functions with bounded fat-shattering dimension.

Algorithms arising from the approach are related to those of Cortes and Vapnik [5] and directly justify the original proposal made to minimise the 2-norm of the slack variables. We generalise the basic result to function classes with bounded fat-shattering dimension and the 1-norm of the slack variables which gives rise to Vapnik's box constraint algorithm. Finally, application to regression is considered, which includes results for standard least squares as a special case.

The paper is structured as follows. In the next section we will summarise the results in O notation to give a flavour of what the paper aims to achieve. In Section 3 relevant background material and definitions are introduced. This is followed by a section describing the results for classification using linear functions. This is the simplest case considered and provides insight into the basic techniques employed. Section 5 describes the algorithm consequences of these results for Support Vector Machine classification algorithms. We then proceed to generalize the results to non-linear function spaces in Section 6. The penultimate section considers the problem of regression and shows how the results obtained for classification readily generalize to this case.

2 Summary of Results

The results in this section will be given in the $\tilde{O}$ notation indicating asymptotics ignoring log factors. The aim is to give the flavour of the results obtained which might otherwise be obscured by the detailed technicalities of the proofs and precise bounds obtained.

The first case considered is that of classification using linear function classes that include the use of kernel functions such as those used in the Support Vector Machine. For this case consider a margin γ about the separating hyperplane and set $(d(\gamma)_{(x,y)})_{(x,y)\in S})$ to be the vector for training set S to be the vector of the amounts by which the training points fail to achieve the margin γ. We bound the probability ϵ of misclassification of a randomly chosen test point by (see Theorem 4.3)

$$\epsilon \leq \tilde{O}\left(\frac{(R+\|d\|_2)^2}{|S|\gamma^2}\right),$$

where R is the radius of a ball about the origin which contains the support of the input probability distribution.

The results are generalized to non-linear function classes using a characterisation of their capacity at scale γ known as the fat shattering dimension $\mathrm{fat}(\gamma)$. In this case the bound obtained has the form (see Theorem 6.13)

$$\epsilon \leq \tilde{O}\left(\frac{\mathrm{fat}(\gamma/16)+\|d\|_2^2/\gamma^2}{|S|}\right),$$

This result is generalized to obtain a bound in terms of the 1-norm of the vector d (see Corollary 6.14)

$$\epsilon \leq \tilde{O}\left(\frac{\mathrm{fat}(\gamma/16)+\|d\|_1/\gamma^2}{|S|}\right),$$

which could of course also be applied to the linear case using a bound on the fat shattering dimension for this case.

Finally, the problem of estimating errors of regressors is addressed with the techniques developed. We bound the probability ϵ that for a randomly chosen test point the absolute error is greater than a given value θ. In this case we define a vector $(\partial_{(x,y)})_{(x,y)\in S}$ of amounts by which the error on the training examples exceeds $\theta - \gamma$. Note that $\|\partial(\theta)\|_2^2$ is simply the least squares error on the training set. We then bound the probability ϵ by (see Theorem 7.2)

$$\epsilon \leq \tilde{O}\left(\frac{\mathrm{fat}(\gamma/16) + \|\partial(\gamma)\|_2^2/\gamma^2}{|S|}\right).$$

These results can be used for Support Vector Regression and give a way of choosing the optimal size $\theta - \gamma$ of the tube for the insensitive loss function. In addition they can be applied to standard least square regression by setting $\gamma = \theta$ to obtain the bound (see Corollary 7.4)

$$\epsilon \leq \tilde{O}\left(\frac{\mathrm{fat}(\theta/16) + \|\partial(\theta)\|_2^2/\theta^2}{|S|}\right).$$

3 Background Results

We consider learning from examples, initially of a binary classification. We denote the domain of the problem by X and a sequence of inputs by $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$. A training sequence is typically denoted by $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times \{-1, 1\})^m$ and the set of training examples by S. By $\mathrm{Er}_{\mathbf{z}}(h)$ we denote the number of classification errors of the function h on the sequence $\mathbf{z}$. As we will typically be classifying by thresholding real valued functions we introduce the notation $T_\theta(f)$ to denote the function giving output 1 if f has output greater than or equal to θ and -1 otherwise. For a class $\mathcal{H}$ the class $T_\theta(\mathcal{H})$ is the set of derived classification functions.

We first give some necessary definitions.

Definition 3.1 *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.*

The following theorem is well known in a number of different forms. We quote the result here as a bound on the generalization error rather than as a required sample size for given generalization.

Theorem 3.2 *[12] Let H_i, $i = 1, 2, \ldots$ be a sequence of hypothesis classes mapping X to $\{0, 1\}$ such that $\mathrm{VCdim}(H_i) = i$, and let P be a probability distribution on X. Let p_d be any set of positive numbers satisfying $\sum_{d=1}^{\infty} p_d = 1$. With probability $1 - \delta$ over m independent examples drawn according to P, for any d for*

which a learner finds a consistent hypothesis h in H_d, the generalization error of h is bounded from above by

$$\epsilon(m,d,\delta) = \frac{4}{m}\left(d\ln\left(\frac{2em}{d}\right) + \ln\left(\frac{1}{p_d}\right) + \ln\left(\frac{4}{\delta}\right)\right),$$

provided $d \leq m$.

We now introduce the generalization of the VC dimension which makes it possible to generalize Theorem 3.2 to large margin classification.

Definition 3.3 *Let $\mathcal{H}$ be a set of real valued functions. We say that a set of points X is* γ-shattered by $\mathcal{H}$ *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 $f_b \in \mathcal{H}$ satisfying*

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

The fat shattering dimension $\mathrm{fat}_{\mathcal{H}}$ *of the set $\mathcal{H}$ 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 [12] which involves the fat shattering dimension of the space of functions.

Theorem 3.4 *Consider a real valued function class $\mathcal{H}$ having fat shattering function bounded above by the function* $\mathrm{afat} : \Re \to \mathcal{N}$ *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 $\gamma = \min|f(x_i) - \theta|$ 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.* [12]. Gurvits [9] generalised this to infinite dimensional Banach spaces. We will quote an improved version of this bound for Hilbert spaces which is contained in [3] (slightly adapted here for an arbitrary bound on the linear operators).

Theorem 3.5 *[3] 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 3.4 and 3.5 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.

4 Linear Function Classes

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

Definition 4.1 *Let $L_f(X)$ be the set of real valued functions f on X with support* $\mathrm{supp}(f)$ *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*

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

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

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

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

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

Theorem 4.2 *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 with $\|\mathbf{u}\| = 1$, thresholded at b is bounded by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log_2\left(\frac{8em}{k}\right)\log_2(32m) + \log_2\left(\frac{720m\log_2(1 + mR^2/\Delta^2)}{\delta}\right)\right),$$

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 $m \geq 2/\epsilon$, $k \leq em$ and there is no discrete probability on misclassified training points.

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

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

We claim that

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

Hence, the off training set 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 $x \in S$, $\|\tau(x)\|^2 \leq R^2 + \Delta^2$ 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 3.4 and 3.5 provided that there are no misclassified training points with discrete probability. Since Theorem 3.4 can only be applied for a fixed space of functions we must apply the two theorems for a discrete set of values for the bound B on the norm of the linear functions. This corresponds to choosing a discrete set of values for the product $(BR)^2$ of Theorem 3.5. We will choose the arithmetic series $\alpha^i(R^2+\Delta^2)$, for $i = 1, \ldots, \ell = 90\log_2(1+mR^2/\Delta^2)$, where α is chosen so that $\alpha^\ell(R^2 + \Delta^2) = (R^2 + \Delta^2)(1 + mR^2/\Delta^2)$ which is an upper bound on the product $\|\tau(x)\|^2\|\hat{\mathbf{u}}\|^2$, since $D^2 \leq mR^2$. Hence, $\alpha = 2^{1/90}$ and it is readily verified that $64.005 \times \alpha < 64.5$. This implies that if we replace the constant 64 of Theorem 3.4 by 64.005 to ensure the continuity from the right, then for the observed value of $\|\tau(x)\|^2\|\hat{\mathbf{u}}\|^2$ there is an application of the theorem for a value of $(BR)^2$ within a factor α of this value and the required bound holds. Note that for each application of the theorem we must replace the δ of Theorem 3.4 by $\delta' = \delta/\ell$ in order to ensure that all applications hold uniformly with probability $1 - \delta$.

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

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

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

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

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 4.3 *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 $d((x,y),(\mathbf{u},b),\gamma) = 0$, for some $(x,y) \in S$, the generalization of a linear classifier $\mathbf{u}$ on X satisfying $\|\mathbf{u}\| \leq 1$ is bounded by*

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k \log_2\left(\frac{8em}{k}\right)\log_2(32m) + \log_2\left(\frac{180m(21+\log_2(m))^2}{\delta}\right)\right),$$

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 $m \geq \max\{2/\epsilon, 6\}$, $k \leq em$ and there is no discrete probability on misclassified training points.

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 $i = 2,\ldots,t$, where t satisfies, $R/32 \geq \Delta_t > R/64$. Hence, $t \leq \log_2(128m^{0.25}) = 0.25(28 + \log_2(m))$. We apply Theorem 4.2 for each of these values of Δ, using $\delta' = \delta/t$ in each application. Note that we have also loosely upper bounded the expression $(28 + \log_2(m))\log_2(1 + mR^2/\Delta^2)$ by $(21+\log_2(m))^2$ 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 $d((x,y),(\mathbf{u},b),\gamma) = 0$, for some $(x,y) \in S$, we must have $d((x',y'),(\mathbf{u},b),\gamma) \leq 2R$, for all $(x',y') \in S$. 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 4.2 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. ∎

5 Algorithmics

The theory developed in the previous section provides a principled answer to a long standing question: what is the "best" linear separation of a set of points that are not linearly separable? Many heuristics have been proposed (see [4] for a review), mainly aimed at reducing the empirical risk, but most of them suffer from computational problems. The question is a practically interesting one, expecially after the revival of perceptron-like systems due to the success of Support Vector Machines [5, 18]. The inability of the original Support Vector Machines to deal with noise (and tolerate outliers) is a serious practical limitation, expecially because - when combined with the use of kernels - it can easily lead to overfitting. The solution developed for Support Vector machines is a heuristic known as the "Soft Margin", which will be described below. The bound proven in the previous section implies the following algorithm: minimize $D(S, (\mathbf{u}, b), \gamma)$ for a given fixed value of γ, and subsequently minimize the bound over different choices of γ. This would ensure that the hyperplane coincides with the minimum of the upper bound on the generalization error. Moreover, as we will see, it can be found in polynomial time. The approach taken by Vapnik [18, Section 5.5.1] for his Soft Margin Classifer is similar, albeit with totally different motivations: in order to minimize the training error of the output hypothesis (an NP-complete task) he approximates it with the quantity $\sum_{j=1}^{m} d_j^{\sigma}$,

which tends to the training error for $\sigma \to 0$. This gives rise to the following algorithm: for non-negative variables $d_i \geq 0$, minimize the function

$$F_\sigma(\mathbf{d}) = \sum_{j=1}^{m} d_j^\sigma,$$

subject to the constraints:

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

which can be solved in polynomial time for $\sigma = 1$ or $\sigma = 2$. This constrained optimization problem is then solved by minimizing the following quantity (problem (1)):

$$\alpha \sum_{j=1}^{m} d_j^\sigma + \frac{1}{2}\langle \mathbf{u} \cdot \mathbf{u} \rangle$$

for different, fixed, values of α. A suitable value of α is then usually chosen by means of a validation set. Once translated into dual variables, this problem turns out to be a quadratic programming problem for each fixed value of α, and can be solved efficiently using standard methods.

The algorithm which follows from the theory presented in the previous section can, in contrast, be described as follows (problem (2)): minimize

$$\|\mathbf{d}\|^2 = \sum_{j=1}^{m} d_j^2$$

subject to constraints

$$\begin{aligned} y_j[\langle \mathbf{u} \cdot \mathbf{x}_j \rangle - b] &\geq 1 - d_j, \quad j = 1, \ldots, m & (3) \\ \langle \mathbf{u} \cdot \mathbf{u} \rangle &= C & (4) \end{aligned}$$

which corresponds exactly to minimizing $D(S, (\mathbf{u}, b), \gamma)$, where $\gamma = \frac{1}{\sqrt{C}}$. This follows from considering the hyperplane $(\mathbf{u}', b') = (\mathbf{u}/\sqrt{C}, b/\sqrt{C})$ which has norm 1 and classifies the point (x_j, y_j) such that

$$d((x_j, y_j), (\mathbf{u}', b'), \gamma) = d_j/\sqrt{C},$$

so that $D(S, (\mathbf{u}', b'), \gamma) = \sqrt{F_2(\mathbf{d})/C}$. Once translated into dual variables, problem (2) gives rise to a convex maximization problem [14]

$$F(\lambda_0, \overline{\lambda}) = -\frac{1}{4}\sum_{j=1}^{m} \lambda_j^2 + \sum_{j=1}^{m} \lambda_j - \frac{1}{4\lambda_C} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle - \lambda_0 C,$$

which must be solved subject to the constraints, $\lambda_j \geq 0$, $j = 0, \ldots, m$, and $\sum_{j=1}^{m} \lambda_j y_j = 0$ for each value of $\gamma = \frac{1}{\sqrt{C}}$, giving the optimal (according to our bound) hyperplane of fixed norm $||w|| = \frac{1}{\gamma}$. Its solution can be found in

polynomial time by applying a gradient based path algorithm following grad(F) with an appropriate learning rate η, but this convex optimization problem is more difficult than a standard quadratic programming one. The best γ is then chosen again using the bound derived in the previous section, namely:

$$w^* = \text{argmin}_\gamma \min_{||w||=1/\gamma} \left(\frac{R+D}{\gamma}\right)^2$$

We will now show that the same result can be obtained by solving the (simpler) quadratic problem used by Vapnik, with $\sigma = 2$ and α is optimised with respect to our bound. The idea is that the class of functions defined by problem (1) for $\alpha \in \Re^+$ is identical to the class of functions defined by problem (2) for $\gamma \in \Re^+$. The optimal function according to our bound is hence the same in both classes. First we need to prove a technical lemma, and state some definitions.

Lemma 5.1 *The hyperplane implicitly defined by the optimization problem (1) depends continuously on the parameter α.*

Proof: This follows from the fact that the dual function equivalent to problem (1) (once maximized in the positive quadrant for each value of α) [5]

$$\mathcal{W}(\overline{\lambda}) = \sum \lambda_i - \frac{1}{2}\sum_{i,j} y_i y_j \lambda_i \lambda_j K(x_i, x_j) - \frac{1}{\alpha}\sum \lambda_i^2 + \lambda_C \sum \lambda_i y_i$$

is continuous both in $\overline{\lambda}$ and in α, and is strictly convex in $\overline{\lambda}$ for any fixed value of α. Strict convexity follows from the fact that its Hessian

$$H_{ij} = \frac{\partial F_i}{\partial \lambda_{j_j}} = y_{j_i} y_{j_j} K(x_{j_i}, x_{j_j}) - \frac{1}{\alpha} = H_{ji}, i, j \geq 1,$$

is negative definite. ∎

Definition 5.2 *We define $\mathcal{W}_\alpha$ to be the set of the solutions of problem (1) for all values of α, and $\mathcal{W}_\gamma$ to be the set of the solutions of problem (2) for all values of γ. Formally:*

$$\mathcal{W}_\alpha = \{\mathbf{u} \in \Re^k | \exists \alpha \in \Re^+, \mathbf{u} = argmin_\alpha \alpha \sum_{j=1}^{m} d_j^2 + \frac{1}{2}\langle \mathbf{u}, \mathbf{u} \rangle\}$$

$$\mathcal{W}_\gamma = \{\mathbf{u} \in \Re^k | \exists \gamma \in \Re^+, \mathbf{u} = argmin_{||\mathbf{u}||=1/\gamma} \left(\frac{R+D}{\gamma}\right)^2\}$$

Theorem 5.3 *The sets of functions $\mathcal{W}_\alpha$ and $\mathcal{W}_\gamma$ defined above are equivalent.*

Proof: let w_α be the solution of the problem (1) for a fixed value of α. Then, $||w_\alpha|| \to 0$ if $\alpha \to 0$, and $||w_\alpha|| \to \infty$ if $\alpha \to \infty$. By lemma 5.1 we know that the function $||w(\alpha)||$ is continuous in α, and hence the norm of the solutions

of problem (1) ranges through all possible positive values for suitably chosen values of α. Since $||w|| = \frac{1}{\sqrt{C}}$, considering the solution for value of α in problem (1) is equivalent solving problem (2) with $C = \|w_\alpha\|$. This implies that for each function $w_\alpha \in \mathcal{W}_\alpha$ there exists a value of γ such that the corresponding $w_\gamma \in \mathcal{W}_\gamma$, and $w_\alpha = w_\gamma$. ∎

An obvious, but important, consequence of this theorem is the following corollary:

Corollary 5.4 *The minima of the bound on $\mathcal{W}_\alpha$ and $\mathcal{W}_\gamma$ coincide:*

$$min_{w_\alpha \in \mathcal{W}_\alpha} \left(\frac{R+D}{\gamma} \right)^2 = min_{w_\gamma \in \mathcal{W}_\gamma} \left(\frac{R+D}{\gamma} \right)^2$$

This means that the optimal separating hyperplane (according to our bound) can be found by solving the quadratic optimization problem (2) with $\sigma = 2$ for different values of α, and choosing the value of α which minimizes the bound itself. This analysis provides a theoretical justification to the Soft Margin heuristics (using the 2-norm of the vector $\overline{d}$ described in the appendix of [5]), as well as a theoretically sound way to choose the optimal value of α in that case. In the next sections we will generalize the theoretical results given so far, and this will lead to a further description of soft-margin heuristics for Support Vector Machines as well as non-linear function classes.

6 Non-linear Function Spaces

6.1 Further Background Results

Before we can quote the next lemma, we need another definition.

Definition 6.1 *Let (X, d) be a (pseudo-) metric space, let A be a subset of X and $\epsilon > 0$. A set $B \subseteq X$ is an* ϵ-cover *for A if, for every $a \in A$, there exists $b \in B$ such that $d(a, b) \leq \epsilon$. The* ϵ-covering number *of A, $\mathcal{N}_d(\epsilon, A)$, is the minimal cardinality of an ϵ-cover for A (if there is no such finite cover then it is defined to be ∞). We will say the cover is proper if $B \subseteq A$.*

Note that we have used less than or equal to in the definition of a cover. This is somewhat unconventional, but will not change the bounds we use. It does, however, prove technically useful in the proofs. The idea is that B should be finite but approximate all of A with respect to the pseudometric d. we will use the l^∞ distance over a finite sample $\mathbf{x} = (x_1, \ldots, x_m)$ for the pseudo-metric in the space of functions,

$$d_{\mathbf{x}}(f, g) = \max_i |f(x_i) - g(x_i)|.$$

We write $\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) = \mathcal{N}_{d_{\mathbf{x}}}(\epsilon, \mathcal{F})$ We will consider the covers to be chosen from the set of all functions with the same domain as $\mathcal{F}$ and range the reals.

We now quote a lemma from [12] which follows immediately from a result of Alon *et al.* [1].

Corollary 6.2 *[12] Let $\mathcal{F}$ be a class of functions $X \to [a,b]$ and P a distribution over X. Choose $0 < \epsilon < 1$ and let $d = \mathrm{fat}_{\mathcal{F}}(\epsilon/4)$. Then*

$$\sup_{\mathbf{x} \in X^m} \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) \leq 2 \left(\frac{4m(b-a)^2}{\epsilon^2} \right)^{d \log_2(2em(b-a)/(d\epsilon))}.$$

We define a clipping function π_γ.

$$\pi_\gamma(\alpha) := \begin{cases} \theta & \text{if } \alpha > \theta \\ \theta - 2.01\gamma & \text{if } \alpha < \theta - 2.01\gamma \\ \alpha & \text{otherwise,} \end{cases}$$

and let $\pi_\gamma(\mathcal{F}) = \{\pi_\gamma(f) : f \in \mathcal{F}\}$. The choice of the threshold θ is arbitrary but will be fixed before any analysis is made. If the value of θ needs to be included explicitly we will denote the clipping function by π_γ^θ.

For a monotonic function $f(\gamma)$ we define

$$f(\gamma^-) = \lim_{\alpha \to 0^+} f(\gamma - \alpha),$$

that is the left limit of f at γ. Note that the minimal cardinality of an ϵ-cover is a monotonically decreasing function of ϵ, as is the fat shattering dimension as a function of γ.

Definition 6.3 *We say that a class of functions $\mathcal{F}$ is* sturdy *its images under the evaluation maps*

$$\tilde{x}_{\mathcal{F}} : \mathcal{F} \longrightarrow \Re, \quad \tilde{x}_{\mathcal{F}} : f \mapsto f(x)$$

are compact subsets of $\Re$ for all $x \in X$.

Note that this definition differs slightly from that introduced in [15]. The current definition is more general, but at the same time simplifies the proof of the required properties.

Lemma 6.4 *Let $\mathcal{F}$ be a sturdy class of functions. Then for each $N \in \mathbb{N}$ and any fixed sequence $\mathbf{x} \in X^m$, the infimum $\gamma_N = \inf\{\gamma | \mathcal{N}(\gamma, \mathcal{F}, \mathbf{x}) = N\}$, is attained.*

Proof: The straightforward proof follows exactly the proof of Lemma 2.6 of [15]. ∎

Corollary 6.5 *Let $\mathcal{F}$ be a sturdy class of functions. Then for each $N \in \mathbb{N}$ and any fixed sequence $\mathbf{x} \in X^m$, the infimum $\gamma_N = \inf\{\gamma | \mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}), \mathbf{x}) = N\}$, is attained.*

Proof: Suppose that the assertion does not hold for some $\mathbf{x} \in X^m$ and $N \in \mathbb{N}$. Let $N' = \mathcal{N}(\gamma_N, \pi_{\gamma_N}(\mathcal{F}), \mathbf{x}) > N$. Consider the following supremum

$$\gamma^{N'} = \sup\{\gamma | \mathcal{N}(\gamma, \pi_{\gamma_N}(\mathcal{F}), \mathbf{x}) = N'\}.$$

Since the assertion does not hold we have $\gamma^{N'} \geq \gamma_N$. By the lemma we must have $\gamma^{N'} > \gamma_N$, since otherwise the infimum of the γ required for the next size of cover will not be attained. Hence, there exists $\gamma' > \gamma_N$ with $\mathcal{N}(\gamma', \pi_{\gamma_N}(\mathcal{F}), \mathbf{x}) = N'$. Let $\gamma = (\gamma' + \gamma_N)/2$. Note that $\mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}), \mathbf{x}) \leq N$. Let B be a minimal cover in this case. Claim that B is also a γ' cover for $\pi_{\gamma_N}(\mathcal{F})$ in the $d_{\mathbf{x}}$ metric. To show this consider $f \in \mathcal{F}$ and let $f_i \in B$ be within γ of $\pi_\gamma(f)$ in the $d_{\mathbf{x}}$ metric. Hence, for all $x \in \mathbf{x}$, $|f_i(x) - \pi_\gamma(f)(x)| \leq \gamma$. But this implies that

$$\begin{aligned} |f_i(x) - \pi_{\gamma_N}(f)(x)| &\leq \gamma + (\gamma - \gamma_N) \\ &= \gamma'. \end{aligned}$$

Hence, we have $\mathcal{N}(\gamma', \pi_{\gamma_N}(\mathcal{F}), \mathbf{x}) \leq N$, a contradiction. ∎

We will make use of the following lemma, which in the form below is due to Vapnik [17, page 168].

Lemma 6.6 *Let X be a set and S a system of sets on X, and P a probability measure on X. For $\mathbf{x} \in X^m$ and $A \in S$, define $\nu_{\mathbf{x}}(A) := |\mathbf{x} \cap A|/m$. If $m > 2/\epsilon$, then*

$$P^m \left\{ \mathbf{x} : \sup_{A \in S} |\nu_{\mathbf{x}}(A) - P(A)| > \epsilon \right\} \leq 2P^{2m} \left\{ \mathbf{xy} : \sup_{A \in S} |\nu_{\mathbf{x}}(A) - \nu_{\mathbf{y}}(A)| > \epsilon/2 \right\}.$$

The following two theorems are essentially quoted from [12] but they have been reformulated here in terms of the covering numbers involved. The difference will be apparent if Theorem 6.8 is compared with Theorem 3.4 quoted in Section 3.

Lemma 6.7 *Suppose $\mathcal{F}$ is a sturdy set of functions that map from X to $\Re$. Then for any distribution P on X, and any $k \in \mathbb{N}$ and any $\theta \in \Re$*

$$P^{2m} \left\{ \mathbf{xy} : \exists f \in \mathcal{F}, r = \max_j \{f(x_j)\}, 2\gamma < \theta - r, \lceil \log_2(\mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}), \mathbf{xy})) \rceil = k, \right.$$
$$\left. \frac{1}{m} |\{i | f(y_i) \geq \theta\}| > \epsilon(m, k, \delta) \right\} < \delta,$$

where $\epsilon(m, k, \delta) = \frac{1}{m}(k + \log_2 \frac{2}{\delta})$.

Proof: We have omitted the detailed proof since it is essentially the same as the corresponding proof in [12] with the simplification that Corollary 6.2 is not required and the property of sturdiness ensures by Corollary 6.5 that we can find a γ_k cover where

$$\gamma_k = \inf\{\gamma | \mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}), \mathbf{xy}) = 2^k\}$$

which can be used for all γ satisfying $\lceil \log_2(\mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}), \mathbf{xy})) \rceil = k$. Note also that an inequality is required $2\gamma < \theta - r$, as we have coverings using closed rather than open balls. ∎

Theorem 6.8 *Consider a sturdy real valued function class $\mathcal{F}$ having a uniform bound on the covering numbers*

$$\mathcal{N}(\gamma^-, \pi_{\gamma^-}(\mathcal{F}), \mathbf{x}) \leq \mathcal{B}(\ell, \gamma),$$

for all $\mathbf{x} \in X^\ell$, for all ℓ. Fix $\theta \in \Re$. If a learner correctly classifies m independently generated examples $\mathbf{z}$ with $h = T_\theta(f) \in T_\theta(\mathcal{F})$ such that $\mathrm{er}_{\mathbf{z}}(h) = 0$ and $\gamma = \min |f(x_i) - \theta|$, then with confidence $1 - \delta$ the expected error of h is bounded from above by

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

where $k = \lceil \log_2 \mathcal{B}(2m, \gamma/2) \rceil$.

Proof: Making use of lemma 6.6 we will move to the double sample and stratify by k. By the union bound, it thus suffices to show that $\sum_{k=1}^{2m} P^{2m}(J_k) < \delta/2$, where

$$\begin{aligned} J_k = \{\mathbf{xy} \quad : \quad & \exists h = T_\theta(f) \in T_\theta(\mathcal{F}), \mathrm{Er}_{\mathbf{x}}(h) = 0, k = \lceil \log_2 \mathcal{B}(2m, \gamma/2) \rceil, \\ & \gamma = \min |f(x_i) - \theta|, \mathrm{Er}_{\mathbf{y}}(h) \geq m\epsilon(m, k, \delta)/2\}. \end{aligned}$$

(The largest value of k we need consider is $2m$, since for larger values the bound will in any case be trivial). It is sufficient if $P^{2m}(J_k) \leq \frac{\delta}{4m} = \delta'$. We will in fact work with the set

$$\begin{aligned} J_k(\gamma') = \{\mathbf{xy} \quad : \quad & \exists h = T_\theta(f) \in T_\theta(\mathcal{F}), \mathrm{Er}_{\mathbf{x}}(h) = 0, k = \lceil \log_2 \mathcal{N}(\gamma'/2, \pi_{\gamma'/2}(\mathcal{F}), \mathbf{xy}) \rceil, \\ & \gamma' < \min |f(x_i) - \theta|, \mathrm{Er}_{\mathbf{y}}(h) \geq m\epsilon(m, k, \delta)/2\}. \end{aligned}$$

We will show that for any $\gamma' < \gamma$, we have $P^{2m}(J_k(\gamma')) \leq \delta'$. Hence, considering the limit $\gamma' \to \gamma$ from below, the result will follow.
Consider $\hat{\mathcal{F}} = \hat{\mathcal{F}}_\theta$. The probability distribution on $\hat{X} = X \times \{0, 1\}$ is given by P on X with the second component determined by the target value of the first component. Note that for a point $y \in \mathbf{y}$ to be misclassified, it must have $\hat{f}(\hat{y}) \geq \theta > \max\{\hat{f}(\hat{x}) : \hat{x} \in \hat{\mathbf{x}}\} + \gamma$, so that

$$\begin{aligned} J_k(\gamma') \subseteq \Big\{\hat{\mathbf{x}}\hat{\mathbf{y}} \in (X \times \{0,1\})^{2m} \quad : \quad & \exists \hat{f} \in \hat{\mathcal{F}}, r = \max\{\hat{f}(\hat{x}) : \hat{x} \in \hat{\mathbf{x}}\}, \gamma' < \theta - r, \\ & k = \lceil \log \mathcal{N}(\gamma'/2, \pi_{\gamma'/2}(\mathcal{F}), \mathbf{xy}) \rceil, \left|\{\hat{y} \in \hat{\mathbf{y}} : \hat{f}(\hat{y}) \geq \theta\}\right| \geq m\epsilon(m, k, \delta)/2 \Big\} \end{aligned}$$

Replacing γ by $\gamma'/2$ in Lemma 6.7 and appealing to Lemma 6.6 we obtain $P^{2m}(J_k(\gamma')) \leq \delta'$, for

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k + \log(2/\delta')\right),$$

as required. Note that the condition of Lemma 6.6 are satisfied by ϵ and m. ∎

6.2 Margin distribution and fat shattering

In this section we will generalise the results of Section 4 to function classes for which a bound on their fat-shattering dimension is known. The basic trick is to bound the covering numbers of the sum of two function classes in terms of the covering numbers of the individual classes. If $\mathcal{F}$ and $\mathcal{G}$ a real valued function classes defined on a domain X we denote by $\mathcal{F}+\mathcal{G}$ the function class

$$\mathcal{F}+\mathcal{G} = \{f+g | f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Lemma 6.9 *Let $\mathcal{F}$ and $\mathcal{G}$ be two real valued function classes both defined on a domain X. Suppose $\mathcal{G}$ has range $[a,b]$. Then we can bound the cardinality of a minimal γ cover of $\mathcal{F}+\mathcal{G}$ by*

$$\mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}+\mathcal{G}), \mathbf{x}) \leq \mathcal{N}(\gamma/2, \pi^{\theta-a}_{\gamma+(b-a)/2}(\mathcal{F}), \mathbf{x})\mathcal{N}(\gamma/2, \mathcal{G}, \mathbf{x}).$$

Proof: Fix $\eta \in (0,\gamma)$ and let B (respectively C) be a minimal η (respectively $\gamma-\eta$) cover of $\pi^{\theta-a}_{\gamma+(b-a)/2}(\mathcal{F})$ (respectively $\mathcal{G}$) in the $d_{\mathbf{x}}$ metric. Consider the set of functions $B+C$. For any $f+g \in \mathcal{F}+\mathcal{G}$, there is an $f_i \in B$ within η of $\pi^{\theta-a}_{\gamma+(b-a)/2}(f)$ in the $d_{\mathbf{x}}$ metric and a $g_j \in C$ within $\gamma-\eta$ of g in the same metric. For $x \in \mathbf{x}$ we claim

$$|\pi_\gamma(f+g)(x) - \pi_\gamma(f_i+g_j)(x)| \leq \gamma. \tag{5}$$

Hence, $\pi_\gamma(B+C)$ forms a γ cover of $\pi_\gamma(\mathcal{F}+\mathcal{G})$. Since

$$|B+C| \leq \mathcal{N}(\eta, \pi^{\theta-a}_{\gamma+(b-a)/2}(\mathcal{F}), \mathbf{x})\mathcal{N}(\gamma-\eta, \mathcal{G}, \mathbf{x}),$$

the result follows by setting $\eta = \gamma/2$. To justify the claim, assume first that $\theta - 2\gamma \leq (f+g)(x) \leq \theta$. This implies that

$$\theta - 2\gamma - b \leq \theta - 2\gamma - g(x) \leq f(x) \leq \theta - g(x) \leq \theta - a.$$

Hence, in this case using the fact that π_γ only reduces distances,

$$\begin{aligned}
|\pi_\gamma(f+g)(x) - \pi_\gamma(f_i+g_j)(x)| &\leq |(f+g)(x) - (f_i+g_j)(x)| \\
&= |(\pi^{\theta-a}_{\gamma+(b-a)/2}(f)+g)(x) - (f_i+g_j)(x)| \\
&\leq |\pi^{\theta-a}_{\gamma+(b-a)/2}(f)(x) - f_i(x)| + |g(x) - g_j(x)| \\
&\leq \eta + \gamma - \eta = \gamma.
\end{aligned}$$

If on the other hand $(f+g)(x) \geq \theta$, we need only show that $(f_i+g_j)(x) \geq \theta - \gamma$ in order for (5) to be satisfied. But we have $f_i(x) \geq \min\{f(x), \theta-a\} - \eta$, while $g_j(x) \geq g(x) - (\gamma-\eta)$. Hence,

$$\begin{aligned}
(f_i+g_j)(x) &\geq \min\{(f+g)(x), g(x)+\theta-a\} - \gamma \\
&\geq \theta - \gamma.
\end{aligned}$$

Finally, if $(f+g)(x) \leq \theta - 2\gamma$, we must show that $(f_i+g_j)(x) \leq \theta - \gamma$ to satisfy equation (5). In this case $f_i(x) \leq \max\{f(x), \theta - 2\gamma - b\} + \eta$, while $g_j(x) \leq g(x) + (\gamma-\eta)$. Hence,

$$\begin{aligned}
(f_i+g_j)(x) &\leq \max\{(f+g)(x), g(x)+\theta-2\gamma-b\} + \gamma \\
&\leq \theta - \gamma.
\end{aligned}$$

as required. ∎

Before proceeding we need a further technical lemma to show that the property of sturdiness is preserved under the addition operator.

Lemma 6.10 *Let $\mathcal{F}$ and $\mathcal{G}$ be sturdy real valued function classes. Then $\mathcal{F}+\mathcal{G}$ is also sturdy.*

Proof: Consider $x \in X$. $\tilde{x}_{\mathcal{F}}(\mathcal{F})$ is a compact subset of $\Re$ as is $\tilde{x}_{\mathcal{G}}(\mathcal{G})$. Note that

$$\tilde{x}_{\mathcal{F}+\mathcal{G}}(\mathcal{F}+\mathcal{G}) = \tilde{x}_{\mathcal{F}}(\mathcal{F}) + \tilde{x}_{\mathcal{G}}(\mathcal{G}),$$

where the addition of two sets A and B of real numbers is defined

$$A + B = \{a + b | a \in A, b \in B\}.$$

Since, $\tilde{x}_{\mathcal{F}}(\mathcal{F}) \times \tilde{x}_{\mathcal{G}}(\mathcal{G})$ is a compact set of $\Re^2$ and $+$ is a continuous function from $\Re^2$ to $\Re$, we have that $\tilde{x}_{\mathcal{F}}(\mathcal{F}) + \tilde{x}_{\mathcal{G}}(\mathcal{G})$ being the image of a compact set under $+$ is also compact. ∎

Definition 6.11 *Fix a threshold $\theta \in \Re$. For a function f on X we define*

$$d((x,y), f, \gamma) = \max\{0, \gamma - y(f(x) - \theta)\}.$$

This quantity is the amount by which f fails to reach the margin γ on the point (x,y) or 0 if its margin is larger than γ. Let $g_f \in L_f(X)$ be the function

$$g_f = \sum_{(x,y)\in S} d((x,y), f, \gamma) y \delta_x.$$

Proposition 6.12 *Fix $\theta \in \Re$. Let $\mathcal{F}$ be a sturdy class of real-valued functions with range $[a,b] \subset \Re$ having a uniform bound on the covering numbers*

$$\mathcal{N}(\gamma^-, \pi^{\theta+A}_{2\gamma^-+A}(\mathcal{F}), \mathbf{x}) \leq \mathcal{B}(\ell, \gamma, A),$$

for all $\mathbf{x} \in X^\ell$, for all ℓ. Let $\mathcal{G}$ be a sturdy subset of $L_f(X)$ with the uniform bound on the covering numbers,

$$\mathcal{N}(\gamma^-, \mathcal{G}, \mathbf{x}) \leq \mathcal{A}(\ell, \gamma),$$

for $\mathbf{x} \in \Delta^\ell$, where $\Delta = \{\delta_x | x \in X\}$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1-\delta$ over randomly drawn training sets S of size m for all $\gamma > 0$ the generalization of a function $f \in \mathcal{F}$ thresholded at θ satisfying $g_f \in \mathcal{G}$ is bounded by

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

where

$$k = \lceil \log_2 \mathcal{B}(2m, \gamma/4, A) + \log_2 \mathcal{A}(2m, \gamma/4) \rceil,$$

where $A \geq \sup\{\langle g, \delta_x \rangle | g \in \mathcal{G}, x \in X\}$, provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: Consider the fixed mapping τ_1. We extend the function class $\mathcal{F}$ to act on the space $X \times L_f(X)$ by its action on X. We similarly extend the function class $\mathcal{G}$ by composing with a projection. We claim that

1. for $x \notin S$, $f(x) = (f + g_f)(x)$, and
2. the margin of $f + g_f$ with threshold θ on the training set $\tau_1(S)$ is γ.

Hence, the off training set behaviour of the classifier f can be characterised by the behaviour of $f+g_f$, while $f+g_f$ is a large margin classifier in the space $X \times L_f(X)$. In order to bound the generalization error we will apply Theorem 6.8 for $\mathcal{F}+\mathcal{G}$ which gives a bound in terms of the covering numbers. These we will bound using Lemma 6.9. The space $\mathcal{F}+\mathcal{G}$ is sturdy by Lemma 6.10, since both $\mathcal{F}$ and $\mathcal{G}$ are. Note that the range of $\mathcal{G}$ is contained in $[-A, A]$ on the input domain. In this case we obtain the following bound on the covering numbers,

$$\begin{aligned}\lim_{\alpha\to 0^+} \log_2\left(\mathcal{N}((\gamma-\alpha)/2, \pi_{(\gamma-\alpha)/2}(\mathcal{F}+\mathcal{G}), \mathbf{x})\right) &\leq \lim_{\alpha\to 0^+} \log_2\left(\mathcal{N}((\gamma-\alpha)/4, \pi^{\theta+A}_{(\gamma-\alpha)/2+A}(\mathcal{F}), \mathbf{x})\right) \\ &\quad + \lim_{\alpha\to 0^+} \log_2\left(\mathcal{N}((\gamma-\alpha)/4, \mathcal{G}, \mathbf{x})\right) \\ &\leq \log_2(\mathcal{B}(2m, \gamma/4, A)) + \log_2(\mathcal{A}(2m, \gamma/4)),\end{aligned}$$

as required.

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

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

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

$$\begin{aligned} y'((f+g_f)(x') - \theta) &= y'(f(x') - \theta) + y'\left\langle \sum_{(x,y)\in S} d((x,y), f, \gamma) y' \delta_x \cdot \delta_{x'} \right\rangle \\ &\geq \gamma - d((x',y'), f, \gamma) + d((x',y'), f, \gamma) = \gamma. \end{aligned}$$

The theorem follows. ∎

For a training set S, we define

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

Theorem 6.13 *Let $\mathcal{F}$ be a sturdy class of real-valued functions with range $[a, b]$ and fat shattering dimension bounded by $\mathrm{fat}_{\mathcal{F}}(\gamma)$. Fix $\theta \in \Re$ and a scaling of the output range $\eta \in \Re$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1-\delta$ over randomly drawn training*

sets S of size m for all $b-a>\gamma>0$ the generalization of a function $f\in\mathcal{F}$ thresholded at θ is bounded by

$$\epsilon(m,k,\delta)=\frac{2}{m}\left(k\log_2\left(65m\left(1+\tilde{D}\right)^2\right)\log_2\left(9em\left(1+\tilde{D}\right)\right)+\log_2\left(\frac{64m^{1.5}(b-a)}{\delta\eta}\right)\right),$$

where

$$k=\left[\mathrm{fat}_{\mathcal{F}}(\gamma^-/16)+64\tilde{D}^2\right]\quad\textit{and}\quad\tilde{D}=2(D(S,f,\gamma)+\eta)/\gamma,$$

provided $m\geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: We define a sequence of function classes $\mathcal{G}_j\subset L_f(X)$ to be the linear functionals with norm at most B_j on the space $L_f(X)$. We will apply Proposition 6.12 for each class G_j. Note that the range of $\mathcal{G}_j$ is $[-B_j,B_j]$ on the input domain. Note also that the image of $\mathcal{G}_j$ under the evaluation map is a closed bounded subset of the reals and hence is compact. It follows that G_j is sturdy. We choose $B_j=j\eta$, for $j=1,\ldots,\ell=\sqrt{m}(b-a)/\eta$. Hence, $B_\ell=\sqrt{m}(b-a)\geq D(S,f,\gamma)$, for all $f\in\mathcal{F}$ and all $\gamma<b-a$. Hence, for any value of $D=D(S,f,\gamma)$ obtained there is a value of B_j satisfying $D\leq B_j<D+\eta$. Substituting the upper bound $D+\eta$ for this B_j will give the result, when we use $\delta'=\delta/\ell$ and bound the covering numbers of the component function classes using Corollary 6.2 and Theorem 3.5. In this case we obtain the following bounds on the covering numbers,

$$\begin{aligned}\lim_{\alpha\to 0^+}\log_2\left(\mathcal{N}((\gamma-\alpha)/4,\pi^{\theta+B_j}_{\gamma-\alpha+B_j}(\mathcal{F}),\mathbf{x})\right)&\leq 1+d_1\log_2\left(\frac{260m(\gamma/2+B_j)^2}{\gamma^2}\right)\\&\qquad\log_2\left(\frac{18em(\gamma/2+B_j)}{d_1\gamma}\right)\\&=:\log_2(\mathcal{B}(2m,\gamma/4,B_j))\end{aligned}$$

where $d_1=\mathrm{fat}_{\mathcal{F}}(\gamma^-/16)$, and

$$\begin{aligned}\lim_{\alpha\to 0^+}\log_2\left(\mathcal{N}((\gamma-\alpha)/4,\mathcal{G}_j,\mathbf{x})\right)&\leq 1+d_2\log_2\left(\frac{260mB_j^2}{\gamma^2}\right)\log_2\left(\frac{18emB_j}{d_2\gamma}\right)\\&=:\log_2(\mathcal{A}(2m,\gamma/4))\end{aligned}$$

where $d_2=(16B_j/\gamma)^2$. Hence, in this case we can bound $\lceil\log_2\mathcal{B}(2m,\gamma/4,B_j)+\log_2\mathcal{A}(2m,\gamma/4)\rceil$ by

$$\begin{aligned}\lceil\log_2\mathcal{B}(2m,\gamma/4,B_j)+\log_2\mathcal{A}(2m,\gamma/4)\rceil&\leq 3+\left[\mathrm{fat}_{\mathcal{F}}(\gamma^-/16)+\left(\frac{16B_j}{\gamma}\right)^2\right]\\&\qquad\log_2 65m(1+2B_j/\gamma)^2\log_2 9em(1+2B_j/\gamma)\end{aligned}$$

giving the result where the 3 contributes a factor of 8 into the argument of the final log term. ∎

The theorem can of course be applied for linear function classes, using the bound on the fat shattering dimension given in Theorem 3.5. The bound obtained is very comparable, though a lot less clean than Theorem 4.3.

For a training set S, we define

$$D'(S,f,\gamma) = \sum_{(x,y)\in S} d((x,y),f,\gamma).$$

This is the l_1 sum of the slack variables which is optimised in Vapnik's box constraint maximal margin hyperplane algorithm. The following Corollary shows that optimising this quantity does indeed lead to good generalization.

Corollary 6.14 *Let $\mathcal{F}$ be a sturdy class of real-valued functions with range $[a,b]$ and fat shattering dimension bounded by $\mathrm{fat}_{\mathcal{F}}(\gamma)$. Fix $\theta \in \Re$ and a scaling of the output range $\eta \in \Re$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1-\delta$ over randomly drawn training sets S of size m for all $b-a > \gamma > 0$ the generalization of a function $f \in \mathcal{F}$ thresholded at θ is bounded by*

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k\log_2\left(65m\left(1+\tilde{D}\right)^2\right)\log_2\left(9em\left(1+\tilde{D}\right)\right) + \log_2\left(\frac{64m^{1.5}(b-a)}{\delta\eta}\right)\right),$$

where

$$k = \left\lfloor \mathrm{fat}_{\mathcal{F}}(\gamma^-/16) + 64\tilde{D}^2 \right\rfloor \quad \textit{and} \quad \tilde{D} = 2(\sqrt{D'(S,f,\gamma)(b-a)} + \eta)/\gamma,$$

provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: The corollary follows by observing that

$$\begin{aligned} D(S,f,\gamma) &= \sqrt{\sum_{(x,y)\in S} d((x,y),f,\gamma)^2} \\ &\leq \sqrt{(b-a)\sum_{(x,y)\in S} d((x,y),f,\gamma)} \\ &= \sqrt{D'(S,f,\gamma)(b-a)} \end{aligned}$$

and applying the theorem. ∎

If we choose the hyperplane to minimise $D'(S,f,\gamma)$ and apply the Corollary, we will necessarily obtain a weaker bound than we would if we minimised $D(S,f,\gamma)$ and then applied the Theorem. In the case of linear function classes, better bounds for the generalization in terms of D and D' should be obtained using recent results which bound the covering numbers for different norms directly [21].

It is worth noting that we can apply Corollary 6.14 to the case of linear functions with norm 1 and recover a result similar to Theorem 4.3. The bound would involve an expression R^2+D^2 rather than $(R+D)^2$, which appears preferable. The constants, however, are significantly worse so that overall the bound will not be as tight.

7 Regression

In order to apply the results of the last section to the regression case we formulate the error estimation as a classification problem. Consider a real-valued function class $\mathcal{F}$ and a target real-valued function $t(x)$. For $f \in \mathcal{F}$ we define the function $e(f)$ and the class $e(\mathcal{F})$,

$$\begin{aligned} e(f)(x) &= |f(x) - t(x)|, \\ e(\mathcal{F}) &= \{e(f) | f \in \mathcal{F}\}. \end{aligned}$$

For a training point $(x, y) \in X \times \Re$ we define

$$\partial((x, y), f, \gamma) = \max\{0, |f(x) - y| - (\theta - \gamma)\}.$$

This quantity is the amount by which f exceeds the error margin $\theta - \gamma$ on the point (x, y) or 0 if f is within $\theta - \gamma$ of the target value. Hence, this is the ϵ insensitive loss measure considered by Vapnik with $\epsilon = \theta - \gamma$. Let $g_f \in L_f(X)$ be the function

$$g_f = -\sum_{(x,y)\in S} \partial((x, y), f, \gamma)\delta_x.$$

Proposition 7.1 *Fix $\theta \in \Re$, $\theta > 0$. Let $\mathcal{F}$ be a sturdy class of real-valued functions with range $[a, b] \subset \Re$ having a uniform bound on the covering numbers*

$$\mathcal{N}(\gamma^-, \mathcal{F}, \mathbf{x}) \leq \mathcal{B}(m, \gamma),$$

for all $\mathbf{x} \in X^m$. Let $\mathcal{G}$ be a sturdy subset of $L_f(X)$ with the uniform bound on the covering numbers,

$$\mathcal{N}(\gamma^-, \mathcal{G}, \mathbf{x}) \leq \mathcal{A}(m, \gamma),$$

for $\mathbf{x} \in \Delta^m$, where $\Delta = \{\delta_x | x \in X\}$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1 - \delta$ over randomly drawn training sets S of size m for all $\gamma > 0$ the probability that a function $f \in \mathcal{F}$ has error greater than θ with respect to target function t on a randomly chosen input is bounded by

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

where

$$k = \left\lceil \log_2 \mathcal{B}(2m, \gamma/4) + \log_2 \mathcal{A}(2m, \gamma/4) \right\rceil,$$

where $A \geq \sup\{\langle g, \delta_x\rangle | g \in \mathcal{G}, x \in X\}$, provided $m \geq 2/\epsilon$, there is no discrete probability on training points with error greater than θ and $g_{e(f)} \in \mathcal{G}$

Proof: The result follows from an application of Proposition 6.12 to the function class $e(\mathcal{F})$, noting that we treat all training examples as negative, and hence correct classification corresponds to having error less than θ. Finally, we can bound the covering numbers

$$\mathcal{N}(\gamma, \pi^{\theta+A}_{2\gamma+A}(e(\mathcal{F})), \mathbf{x}) \leq \mathcal{N}(\gamma, \mathcal{F}, \mathbf{x}) \leq \mathcal{B}(m, \gamma).$$

The result follows. ∎

For a training set S, we define

$$\mathcal{D}(S, f, \gamma) = \sqrt{\sum_{(x,y)\in S} \partial((x, y), f, \gamma)^2}.$$

The above result can be used to obtain a bound in terms of the observed value of $D(S, f, \gamma)$ and the fat shattering dimension of the function class.

Theorem 7.2 *Let $\mathcal{F}$ be a sturdy class of real-valued functions with range $[a, b]$ and fat shattering dimension bounded by $\mathrm{fat}_{\mathcal{F}}(\gamma)$. Fix $\theta \in \Re$, $\theta > 0$ and a scaling of the output range $\eta \in \Re$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1 - \delta$ over randomly drawn training sets S of size m for all γ with $\theta \geq \gamma > 0$ the probability that a function $f \in \mathcal{F}$ has error larger than θ on a randomly chosen input is bounded by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log_2\left(65m\left(\frac{b-a}{\gamma}\right)^2\right)\log_2\left(9em\left(\frac{b-a}{\gamma}\right)\right) + \log_2\left(\frac{64m^{1.5}(b-a)}{\delta\eta}\right)\right),$$

where

$$k = \left[\mathrm{fat}_{\mathcal{F}}(\gamma^-/16) + 64\tilde{\mathcal{D}}^2\right] \quad \textit{and} \quad \tilde{\mathcal{D}} = 2(\mathcal{D}(S, f, \gamma) + \eta)/\gamma,$$

provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: The proof follows the same pattern as that of Theorem 6.13, with the exception that the bounds on the covering numbers must use the full range of the function class $\mathcal{F}$ in the log factors. ∎

Corollary 7.3 *Let $\mathcal{F}$ be a the set of linear functions with norm 1 restricted to inputs in a ball of radius R about the origin. Fix $\theta \in \Re$, $\theta > 0$ and a scaling of the output range $\eta \in \Re$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1-\delta$ over randomly drawn training sets S of size m for all γ, with $\theta \geq \gamma > 0$ the probability that a function $f \in \mathcal{F}$ has error larger than θ on a randomly chosen input is bounded by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log_2\left(260m\left(\frac{R}{\gamma}\right)^2\right)\log_2\left(18em\frac{R}{\gamma}\right) + \log_2\left(\frac{128m^{1.5}R}{\delta\eta}\right)\right),$$

where

$$k = \left[256R^2/\gamma^2 + 64\tilde{\mathcal{D}}^2\right] \quad \textit{and} \quad \tilde{\mathcal{D}} = 2(\mathcal{D}(S, f, \gamma) + \eta)/\gamma,$$

provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: The range of linear functions with unit weight vectors when restricted to the unit ball is $[-R, R]$. Their fat shattering dimension is bounded by Theorem 3.5. The result follows. ∎

Note that we obtain a generalization bound for standard least squares regression by taking $\gamma = \theta$ in Theorem 7.2. In this case $\mathcal{D}(S, f, \theta)$ is the least squares error on the training set, while the bound gives the probability of a randomly chosen input having error greater than θ. This is summarised in the following corollary.

Corollary 7.4 *Let $\mathcal{F}$ be a sturdy class of real-valued functions with range $[a, b]$ and fat shattering dimension bounded by $\mathrm{fat}_{\mathcal{F}}(\gamma)$. Fix $\theta \in \Re$, $\theta > 0$ and a scaling of the output range $\eta \in \Re$. Consider a fixed but unknown probability distribution on the input space X. Then with probability $1 - \delta$ over randomly drawn training sets S of size m the probability that a function $f \in \mathcal{F}$ has error larger than θ on a randomly chosen input is bounded by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log_2\left(65m\left(\frac{b-a}{\theta}\right)^2\right)\log_2\left(9em\left(\frac{b-a}{\theta}\right)\right) + \log_2\left(\frac{64m^{1.5}(b-a)}{\delta\eta}\right)\right),$$

where

$$k = \left[\mathrm{fat}_{\mathcal{F}}(\theta^-/16) + 64\tilde{\mathcal{D}}^2\right] \quad \textit{and} \quad \tilde{\mathcal{D}} = 2\frac{\sqrt{\sum_{(x,y)\in S}(f(x)-y)^2} + \eta}{\theta},$$

provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

As mentioned in the section dealing with classification we could bound the generalization in terms of other norms of the vector of slack variables

$$(\partial((x, y), f, \gamma))_{(x,y)\in S}\,.$$

The aim of this paper, however, is not to list all possible results, it is rather to illustrate how such results can be obtained.

Another application of these results is to choose the best ϵ for the ϵ insensitive loss function for Support Vector Regression. This problem has usually been solved by using a validation set, but Corollary 7.3 could be used by choose the value of ϵ which gives the best bound on the generalization. We assume here that a target accuracy θ has been set and we wish to minimise the probability that the error exceeds this value. The optimum will be the ϵ which minimises

$$\frac{R^2 + \mathcal{D}(S, f_\epsilon, \theta - \epsilon)^2}{(\theta - \epsilon)^2},$$

where f_ϵ is the solution obtained when using the ϵ-insensitive loss function.

8 Conclusions

We have shown how an approach developed by Freund and Schapire [8] 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 in which the square root of the sample size replaces the sample size in the denominator. The bound is also more robust that that derived for the maximal margin hyperplane where a single point can have a dramatic effect on the hyperplane produced.

We have gone on to show how optimizing the measure of the margin distribution that appears in the bound corresponds to an algorithm proposed by Cortes and Vapnik [5]. 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 optimizing the generalization performance of agnostic learning with hyperplanes, by showing that the criterion to be minimised should not be the number of training errors, but rather a more flexible criterion which could be termed a 'soft margin'. The problem of finding a more informative and theoretically well-founded measure of the margin distribution has been an open problem for some time. This paper suggests one candidate for such a measure which has the advantage of being robust in the sense that it is not critically sensitive to the behaviour of individual training points.

The results have been further generalized to non-linear function classes with bounded fat-shattering dimensions, other norms on the vector of shortfalls of individual training points and to the regression case. For regression one byproduct is a bound in terms of the least square error on the training set of the probability that a randomly drawn test point will have error greater than a given value.

References

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi and David Haussler, "Scale-sensitive Dimensions, Uniform Convergence, and Learnability," *Journal of the ACM* **44**(4), 615–631, (1997)

[2] Peter L. Bartlett, "The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network," *IEEE Trans. Inf. Theory*, **44**(2), 525–536, (1998).

[3] 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.

[4] C. Campbell, Constructive Learning Techniques for Designing Neural Network Systems, *in* (ed CT Leondes) Neural Network Systems Technologies and Applications. San Diego: Academic Press. 1997.

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

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, Martin Anthony, Structural Risk Minimization over Data-Dependent Hierarchies, *IEEE Trans. on Inf. Theory*, **44**(5), 1926–1940, (1998), and NeuroCOLT Technical Report NC-TR-96-053, 1996.
(`ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports`).

[13] John Shawe-Taylor and Nello Cristianini, Data Dependent Structural Risk Minimization for Perceptron Decision Trees, Proceedings of the Eleventh Conference on Neural Information Processing Systems, NIPS'97. Advances in Neural Information Processing Systems 10 Michael I. Jordan, Michael J. Kearns, and Sara A. Solla (eds.) Cambridge, MA: MIT Press (1998), pp. 336–342.

[14] John Shawe-Taylor and Nello Cristianini, Margin Distribution Bounds on Generalization, Submitted to EuroCOLT'99, 1998.

[15] John Shawe-Taylor and Robert C. Williamson, Generalization Performance of Classifiers in Terms of Observed Covering Numbers, Submitted to EuroCOLT'99, 1998.

[16] University of California, Irvine - Machine Learning Repository, `http://www.ics.uci.edu/ mlearn/MLRepository.html`

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

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

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

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

[21] Robert C. Williamson, Alex J. Smola and Bernhard Schölkopf, "Entropy Numbers, Operators and Support Vector Kernels," submitted to EuroCOLT'99. See also "Generalization Performance of Regularization Networks
and Support Vector Machines *via* Entropy Numbers of Compact Operators," `http://spigot.anu.edu.au/people/williams/papers/P100.ps` submitted to *IEEE Transactions on Information Theory*, July 1998.