Further Results on the Margin Distribution

John Shawe-Taylor
Department of Computer Science,
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
j.shawe-taylor@dcs.rhbnc.ac.uk

Nello Cristianini
Dept of Engineering Mathematics
University of Bristol
Bristol BS8 1TR, UK
nello.cristianini@bristol.ac.uk

February 10, 1999

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 [7] 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. Shawe-Taylor and Cristianini [13] showed 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. We show that in the linear case the approach can be viewed as a change of kernel and that the algorithms arising from the approach are exactly those originally proposed by Cortes and Vapnik [4]. We generalise the basic result to function classes with bounded fat-shattering dimension and the l_1 measure for slack variables which gives rise to Vapnik's box constraint algorithm. Finally, application to regression is considered, which includes standard least squares as a special case.

1 Introduction

The idea that a large margin classifier might be expected to give good generalization is certainly not new [6]. 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 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 [2] and applying it to neural networks [2], boosting algorithms [11] and Bayesian algorithms [5], 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 in a bound that depended so critically on the positions of possibly a small proportion of the training set.

Shawe-Taylor and Cristianini [13] following an approach used by Freund and Schapire [7] for on-line learning showed that a measure of the margin distribution can be used to provide pac style bounds on the generalization error.

In this paper we show that in the linear case we can view the technique as a change of kernel and that algorithms arising from the approach correspond exactly to those originally proposed by Cortes and Vapnik [4] as heuristics for agnostic learning. We further generalise the basic result to function classes with bounded fat-shattering dimension and the l_1 measure for slack variables which gives rise to Vapnik's box constraint algorithm. Finally, application to regression is considered. Special applications of our results include a justification for using the square loss in training back-propagation networks, as well as bounds for the probability of exceeding a certain error margin for standard least squares regressors.

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 of real-valued functions $\mathcal{H}$ the class $T_\theta(\mathcal{H})$ is the set of derived classification functions.

Definition 1.1 *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 $f_b(x) \geq r_x + \gamma$, if $b_x = 1$ and $f_b(x) \leq r_x - \gamma$, otherwise. 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.*

2 Linear Function Classes

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 [8] 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 2.1 *[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 $\mathrm{fat}_L(\gamma) \leq \left(\frac{BR}{\gamma}\right)^2$.*

Definition 2.2 *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: $\tau_\Delta : x \mapsto (x, \Delta\delta_x)$, where $\delta_x \in L_f(X)$ is defined by $\delta_x(y) = 1$, if $y = x$ and 0, otherwise.

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 2.3 *[13] 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.

This theorem is applied 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 2.4 *[13] 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.

3 Algorithmics

The theory developed in the previous section provides a way to transform a non linearly separable problem into a separable one by mapping the data to a higher dimensional space, a technique that can be viewed as using a kernel in a similar way to Support Vector Machines.

Is it possible to give an effective algorithm for learning a large margin hyperplane in this augmented space? This would automatically give an algorithm for optimizing the margin distribution in the original space. It turns out that not only is the answer yes, but also that such an algorithm already exists.

The mapping τ defined in the previous section implicitly defines a kernel as follows:

$$k(x,x') = \langle \tau_\Delta(x), \tau_\Delta(x') \rangle = \langle (x, \Delta\delta_x), (x', \Delta\delta_{x'}) \rangle = \langle x, x' \rangle + \Delta^2 \langle \delta_x, \delta_{x'} \rangle = \langle x, x' \rangle + \Delta^2 \delta_x(x')$$

By using these kernels, the decision function of a SV machine would be:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b = \sum_{i=1}^{m} \alpha_i y_i \left[\langle x, x_i \rangle + \Delta^2 \delta_x(x') \right] + b$$

and the lagrange multipliers α would be obtained by solving the following QP problem: minimize in the positive quadrant the lagrangian

$$L \;=\; \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j [\langle x_i, x_j \rangle + \Delta^2 \delta_i(j)] \\
&= \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \Delta^2 \sum_{i=1}^{m} \alpha_i^2
\end{aligned}
$$

This is exacly the dual QP problem that one obtains by solving the soft margin problem for the case $\sigma = 2$, as stated by Cortes and Vapnik [4], minimise $\frac{1}{2}\langle u, u \rangle + C \sum_{i=1}^{m} \xi_i^2$ subject to $y_j[\langle u, x_j \rangle - b] \geq 1 - \xi_j$ and $\xi_j \geq 0$. The solution obtained is

$$
L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - C \sum_{i=1}^{m} \alpha_i^2
$$

which makes clear how the trade off parameter C in their formulation is related to the kernel parameter Δ. Another way to look at this technique is the following: doing soft margin, or enlarging the margin distribution, is equivalent to replacing the covariance matrix K with the covariance, $K' \leftarrow K + \lambda I$, which has a heavier diagonal. Again, the trade off parameter λ is simply related to Δ and C in the previous formulations. So rather than using a soft margin algorithm, one can use a (simpler) hard margin algorithm after adding λI to the covariance matrix.

This technique is well known in classical statistics, where it is sometimes called the "shrinkage method" (see Ripley [10]). In the context of regression it is better known as Ridge Regression, and leads to a form of weight decay. It is a regularization technique in the sense of Tychonov. Another way to describe it, is that it reduces the number of effective free parameters, as measured by the trace of K. Note finally that from an algorithmical point of view these kernels still give a positive definite matrix, and a better conditioned problem.

4 Non-linear Function Spaces

Definition 4.1 *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 4.2 *[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*

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

Let $\pi_\gamma(\alpha)$ be the identity function in the range $[\theta - 2.01\gamma, \theta]$, with output θ for larger values and $\theta - 2.01\gamma$ for smaller ones, and let $\pi_\gamma(\mathcal{F}) = \{\pi_\gamma(f)\colon f \in \mathcal{F}\}$. 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 4.3 *Let*

$$\tilde{\mathbf{x}}\colon \mathcal{F} \longrightarrow \Re^m, \quad \tilde{\mathbf{x}}\colon f \mapsto (f(x_1), f(x_2), \ldots, f(x_m))$$

denote the multiple evaluation map induced by $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$. We say that a class of functions $\mathcal{F}$ is sturdy *if for all $m \in \mathbb{N}$ and all $\mathbf{x} \in X^m$ the image $\tilde{\mathbf{x}}(\mathcal{F})$ of $\mathcal{F}$ under $\tilde{\mathbf{x}}$ is a compact subset of $\Re^m$.*

Lemma 4.4 *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.*

Corollary 4.5 *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 $|f_i(x) - \pi_{\gamma_N}(f)(x)| \leq \gamma + (\gamma - \gamma_N) = \gamma'$. Hence, we have $\mathcal{N}(\gamma', \pi_{\gamma_N}(\mathcal{F}), \mathbf{x}) \leq N$, a contradiction. ∎

The following two theorems are essentially quoted from [12] but they have been reformulated here in terms of the covering numbers involved.

Lemma 4.6 *Suppose $\mathcal{F}$ is a sturdy set of functions that map from X to $\Re$ with a uniform bound on the covering numbers $\mathcal{N}(\gamma, \pi_\gamma(\mathcal{F}), \mathbf{x}) \leq \mathcal{B}(m, \gamma)$, for all $\mathbf{x} \in X^m$. Then for any distribution P on X, and any $k \in \mathbb{N}$ and any $\theta \in \Re$*

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

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 4.2 is not required and the property of sturdiness ensures by Corollary 4.5 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 $\lceil \log_2(\mathcal{B}(2m,\gamma))\rceil = k$. ∎

Theorem 4.7 *Consider a sturdy real valued function class $\mathcal{F}$ having a uniform bound on the covering numbers $\mathcal{N}(\gamma^-, \pi_{\gamma^-}(\mathcal{F}), \mathbf{x}) \leq \mathcal{B}(m,\gamma)$, for all $\mathbf{x} \in X^m$. 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*

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k + \log_2\left(\frac{8m}{\delta}\right)\right), \quad \textit{where } k = \lceil \log_2 \mathcal{B}(2m, \gamma/2)\rceil.$$

Proof: The proof is again identical to the proof of Theorem 3.12 in [12] except that Lemma 4.6 is used in place of the corresponding result of [12]. ∎

4.1 Margin distribution and fat shattering

In this section we will generalise the results of Section 2 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 4.8 *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: The relatively straightforward proof is given in the appendix. ∎

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

Lemma 4.9 *Let $\mathcal{F}$ and $\mathcal{G}$ be sturdy real valued function classes. Then $\mathcal{F}+\mathcal{G}$ is also sturdy.*

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

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

where the addition of two sets A and B of real vectors is defined $A+B = \{a+b | a \in A, b \in B\}$. Since, $\tilde{\mathbf{x}}_{\mathcal{F}}(\mathcal{F}) \times \tilde{\mathbf{x}}_{\mathcal{G}}(\mathcal{G})$ is a compact set of $\Re^{2m}$ and $+$ is a continuous function from $\Re^{2m}$ to $\Re^m$, 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 4.10 *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 4.11 *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 $\mathcal{N}(\gamma^-, \pi^{\theta+A}_{2\gamma^-+A}(\mathcal{F}), \mathbf{x}) \leq \mathcal{B}(m, \gamma, A)$, 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 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 $k = \lceil \log_2 \mathcal{B}(2m, \gamma/4, A) + \log_2 \mathcal{A}(2m, \gamma/4) \rceil$, and $A \geq \sup\{\langle g, \delta_x \rangle | g \in \mathcal{G}, x \in X\}$, 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 4.7 for $\mathcal{F}+\mathcal{G}$ which gives a bound in terms of the covering numbers. These we will bound using Lemma 4.8. The space $\mathcal{F}+\mathcal{G}$ is sturdy by Lemma 4.9, 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,

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

as required. The proof of the first and second claims is as in Theorem 2.3. ∎

For a training set S, we define $D(S,f,\gamma) = \sqrt{\sum_{(x,y)\in S} d((x,y),f,\gamma)^2}$.

Theorem 4.12 *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 $k = \left\lceil \mathrm{fat}_{\mathcal{F}}(\gamma/16) + 64\tilde{D}^2\right\rceil$ and $\tilde{D} = 2(D(S,f,\gamma)+\eta)/\gamma$, 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 4.11 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 4.2 and Theorem 2.1. In this case we obtain the following bounds on the covering numbers,

$$\begin{aligned}\log_2\left(\mathcal{N}(\gamma/4, \pi^{\theta+B_j}_{\gamma+B_j}(\mathcal{F}), \mathbf{x})\right) &\leq 1 + d_1\log_2\left(\frac{256m(\gamma/2+B_j)^2}{\gamma^2}\right)\log_2\left(\frac{16em(\gamma/2+B_j)}{d_1\gamma}\right)\\ &=: \log_2(\mathcal{B}(2m,\gamma/4,B_j))\end{aligned}$$

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

$$\begin{aligned}\log_2\left(\mathcal{N}(\gamma/4, \mathcal{G}_j, \mathbf{x})\right) &\leq 1 + d_2\log_2\left(\frac{256mB_j^2}{\gamma^2}\right)\log_2\left(\frac{16emB_j}{d_2\gamma}\right)\\ &=: \log_2(\mathcal{A}(2m,\gamma/4))\end{aligned}$$

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

$$\begin{aligned}\lceil\log_2\mathcal{B}(2m,\gamma/4,B_j)+\log_2\mathcal{A}(2m,\gamma/4)\rceil &\leq 3 + \left\lceil \mathrm{fat}_{\mathcal{F}}(\gamma/16) + \left(\frac{16B_j}{\gamma}\right)^2\right\rceil\\ &\quad \log_2 64m(1+2B_j/\gamma)^2\log_2 8em(1+2B_j/\gamma)\end{aligned}$$

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

The obvious choice of non-linear function class would be neural networks. Bartlett [2] shows that by placing a bound on the weights we guarantee a bound on the fat-shattering dimension. Hence, to optimize the generalization performance we should minimise the quantity $D(S,f,\gamma)$. The backpropagation algorithm with weight decay, optimizes a trade-off between the fat-shattering dimension and the value $D(S,f,1)$, assuming an output value in the range $[-1,1]$ and a least squares training error. Hence, the theorem gives a more direct justification for the backpropagation algorithm than [2], while at the same time suggesting that one could try optimising the value of $D(S,f,\gamma)$, for $\gamma < 1$. This would correspond to ignoring training points whose margin is already at least γ and measuring the error of points with smaller margin against a target output of $\pm\gamma$.

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 4.13 *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 $k = \left\lceil \mathrm{fat}_{\mathcal{F}}(\gamma/16) + 64\tilde{D}^2 \right\rceil$ and $\tilde{D} = 2(\sqrt{D'(S,f,\gamma)(b-a)} + \eta)/\gamma$, provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: The corollary follows by observing that

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

5 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 $e(\mathcal{F})$, $e(f)(x) = |f(x) - t(x)|$, $e(\mathcal{F}) = \{e(f) | f \in \mathcal{F}\}$.

For a training point $(x,y) \in X \times \Re$ we define $d((x,y),f,\gamma) = \max\{0, |f(x)-y| - (\theta-\gamma)\}$. 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 $g_f = -\sum_{(x,y)\in S} d((x,y),f,\gamma)\delta_x$.

Proposition 5.1 *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 $\mathcal{N}(\gamma^-, \mathcal{F}, \mathbf{x}) \leq \mathcal{B}(m,\gamma)$, 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 $k = \lceil \log_2 \mathcal{B}(2m,\gamma/4) + \log_2 \mathcal{A}(2m,\gamma/4) \rceil$, and $A \geq \sup\{\langle g, \delta_x\rangle | g \in \mathcal{G}, x \in X\}$, 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 4.11 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, the result follows from bounding the covering numbers

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

For a training set S, we define $D(S,f,\gamma) = \sqrt{\sum_{(x,y)\in S} d((x,y),f,\gamma)^2}$. 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 5.2 *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 $\theta \in \Re$ 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 $\theta \geq \gamma > 0$ 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 $k = \left\lceil \mathrm{fat}_{\mathcal{F}}(\gamma/16) + 64\tilde{D}^2 \right\rceil$ and $\tilde{D} = 2(D(S,f,\gamma)+\eta)/\gamma$, provided $m \geq 2/\epsilon$ and there is no discrete probability on misclassified training points.

Proof: The proof follows the same pattern as that of Theorem 4.12, 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. $\blacksquare$

Note that we obtain a generalization bound for standard least squares regression by taking $\gamma = \theta$ in Theorem 5.2. In this case $D(S,f,\theta)$ 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.

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

[5] 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.

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

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

[8] 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.

[9] 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.

[10] B. D. Ripley, Pattern Recognition and Neural Networks, Cambridge: Cambridge University Press, 1996.

[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).

[13] John Shawe-Taylor and Nello Cristianini, Margin Distribution Bounds on Generalization, to appear in the Proceedings of EuroCOLT'99, 1999.

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

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

[16] 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.

Appendix A

Proof of Lemma 4.8 : 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

$$|\pi_\gamma(f+g)(x) - \pi_\gamma(f_i + g_j)(x)| \leq \gamma. \tag{1}$$

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 (1) 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 (1). 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. ■