2 Margin Distribution and Soft Margin

John Shawe-Taylor
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
j.shawe-taylor@dcs.rhbnc.ac.uk
http://www.cs.rhbnc.ac.uk/people/staff/shawe-taylor.shtml

Nello Cristianini
Department of Engineering Mathematics, University of Bristol
Queen's Building, University Walk, Bristol BS8 1TR, UK
nello.cristianini@bristol.ac.uk
http://zeus.bris.ac.uk/~ennc/nello.html

Typical bounds on generalization of Support Vector Machines are based on the minimum distance between training examples and the separating hyperplane. There has been some debate as to whether a more robust function of the margin distribution could provide generalization bounds. Freund and Schapire (1998) 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 to give 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. Furthermore, we show that 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 (1995). Finally, we discuss the relations of this approach with other techniques, such as regularization and shrinkage methods[1].

1. Parts of this work have appeared in Shawe-Taylor and Cristianini (1999b,a)

2.1 Introduction

The presence of noise in the data introduces a trade-off in every learning problem: complex hypotheses can be very accurate on the training set, but have worse predictive power than simpler and slightly inaccurate hypotheses. Hence the right balance between accuracy and simplicity of a hypothesis needs to be sought and this is usually attained by minimizing a cost function formed of two parts, one describing the complexity of the hypothesis, the other measuring its training error. In the case of linear functions this leads to an additional difficulty as the problem of minimising the number of training errors is computationally infeasible if we parametrize the problem in terms of the dimension of the inputs (Arora et al., 1997). We avoid this apparent impasse by bounding the generalization in terms of a different function of the training set performance, namely one based on the distribution of margin values, but not directly involving training error. We will show in this paper that minimising this new criterion can be performed efficiently.

When considering large margin classifiers, where the complexity of a hypothesis is measured by its margin with respect to the data, the presence of noise can lead to further problems, for example datasets may be non-separable, and hence their margin would be negative, making application of the non-agnostic result impossible. Moreover solutions found by maximizing the margin are not stable with respect to the training points – slight modifications in the training set can significantly change the hypothesis – a brittleness which makes the maximal margin solution somehow undesirable. These problems have led to the technique of the "soft-margin", a procedure aimed at extending the large margin algorithms to the noisy case by permitting a trade-off between accuracy and margin.

non-separable data

Despite successes in extending this style of analysis to the agnostic case (Bartlett, 1998) (see (**??**) in this book) and applying it to neural networks (Bartlett, 1998), boosting algorithms (Schapire et al., 1998) and Bayesian algorithms (Cristianini et al., 1998), there has been concern that the measure of the distribution of margin values attained by the training set is largely ignored in a bound in terms of 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.

Though more robust algorithms have been introduced, the problem of robust bounds has remained open until recently. Freund and Schapire (1998) showed that for on-line learning a measure of the margin distribution can be used to give mistake bounds for a perceptron algorithm, and a bound on the expected error. Following a similar technique, in this paper we provide theoretical *pac* bounds on generalization using a more general function of the margin distribution achieved on the training set; we show that this technique can be viewed as a change of kernel and that algorithms arising from the approach correspond exactly to those originally proposed by Cortes and Vapnik (1995) as techniques for agnostic learning. Finally, we will show that the algorithms obtained in this way are intimately related to

margin distribution

certain techniques, usually derived in the framework of regularization or of Bayesian analysis and hence this work can be used to provide a learning-theoretic justification for such techniques.

Note that this style of analysis can also be used to transfer other hard margin results into a soft margin setting, and furthermore it can be extended to cover the nonlinear and regression cases (Shawe-Taylor and Cristianini, 1998).

2.2 Margin Distribution Bound on Generalization

We consider learning from examples 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 $\mathrm{Er}_z(f)$ we denote the number of classification errors of the function f 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.

fat shattering dimension

Definition 2.1
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 relevance of the fat shattering dimension and margin for learning is illustrated in the following theorem which bounds the generalization error in terms of the fat shattering dimension of the underlying function class measured at a scale proportional to the margin.

Theorem 2.1
(Shawe-Taylor et al., 1998) Consider a real valued function class $\mathcal{H}$ having fat-shattering dimension bounded above by the function $\mathrm{fat} : \mathbb{R} \to \mathbb{N}$ which is continuous from the right. Fix $\theta \in \mathbb{R}$. Then with probability at least $1 - \delta$ a learner who correctly classifies m independently generated examples S with $h = T_\theta(f) \in T_\theta(\mathcal{H})$ such that $\gamma = \min_i y_i(f(x_i) - \theta) > 0$ will have the 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 $k = \mathrm{fat}(\gamma/8) \leq em$.

The first bound on the fat shattering dimension of bounded linear functions in a finite dimensional space was obtained by Shawe-Taylor et al. (1998). Gurvits (1997) generalised this to infinite dimensional Banach spaces. We will quote an improved

version of this bound for inner product spaces which is contained in (Bartlett and Shawe-Taylor, 1999) (slightly adapted here for an arbitrary bound on the norm of the linear operators).

Theorem 2.2 Fat shattering of linear functions
(Bartlett and Shawe-Taylor, 1999) 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

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

We first summarise results from Shawe-Taylor and Cristianini (1999b). Let X be an inner product space. We define the following inner product space derived from X.

Definition 2.2
Let $L_f(X)$ be the set of real valued functions f on X with countable support $\mathrm{supp}(f)$ (that is functions in $L_f(X)$ are non-zero for only countably many points) for which the sum of the squared values

$$\|f\|^2 = \sum_{x \in \mathrm{supp}(f)} f(x)^2$$

converges. We define the inner product of two functions $f, g \in L_f(X)$, by

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

Note that the sum which defines the inner product can be shown to converge by using the Cauchy-Schwartz inequality on the difference of partial sums and hence showing that the partial sums form a Cauchy sequence. Clearly the space is closed under addition and multiplication by scalars.

map to a separation space

Now for any fixed $\Delta > 0$ we define an embedding of X into the inner product 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. Embedding the input space X into $X \times L_f(X)$ maps the training data into a space where it can be separated by a large margin classifier and hence we can apply Theorem 2.1. The cost of performing this separation appears in the norm of the linear operator acting in $L_f(X)$ which forces the required margin. The following definition specifies the amount by which a training point has to be adjusted to reach the desired margin γ.

For a linear classifier $(\mathbf{u}, b)$ on X and margin $\gamma \in \mathbb{R}$ we define

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

This quantity is the amount by which $(\mathbf{u}, b)$ fails to reach the margin γ on the point (x, y) or 0 if its margin is larger than γ. For a misclassified point (x, y) we will have $d((x, y), (\mathbf{u}, b), \gamma) > \gamma$, and so misclassification is viewed as a worse margin error, but is not distinguished into a separate category. We now augment $(\mathbf{u}, b)$ to the

linear functional

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

in the space $X \times L_f(X)$. The action of the additional component is exactly enough to ensure that those training points that failed to reach margin γ in the input space now do so in the augmented space. The cost of the additional component is in its effect of increasing the square of the norm of the linear functional by $D(S,(\mathbf{u},b),\gamma)^2/\Delta^2$, where

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

At the same time the norm of the training points has been increased by the additional component $\Delta\delta_x$. Taking both these adjustments into account and verifying that the off-training set performance of the augmented classifier matches exactly the original linear function gives the following theorem as a consequence of Theorems 2.1 and 2.2.

Theorem 2.3

bound for a fixed map

(Shawe-Taylor and Cristianini, 1999b) Fix $\Delta > 0$, $b \in \mathbb{R}$. 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

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

where

$$h = \left\lfloor \frac{64.5(R^2+\Delta^2)(1+D(S,(\mathbf{u},b),\gamma)^2/\Delta^2)}{\gamma^2} \right\rfloor,$$

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

Note that unlike Theorem 2.1 the theorem does not require that the linear classifier $(\mathbf{u},b)$ correctly classifies the training data. Misclassified points will contribute more to the quantity $D(S,(\mathbf{u},b),\gamma)$, but will not change the structure of the result. This contrasts with their effect on Theorem 2.1 where resorting to the agnostic version introduces a square root into the expression for the generalization error.

In practice we wish to choose the parameter Δ in response to the data in order to minimize the resulting bound. In order to obtain a bound which holds for different values of Δ it will be necessary to apply the Theorem 2.3 several times for a finite subset of values. Note that the minimum of the expression for h (ignoring the constant and suppressing the denominator γ^2) is $(R+D)^2$ attained when $\Delta = \sqrt{RD}$. The discrete set of values must be chosen to ensure that we can get

a good approximation to this optimal value. The solution is to choose a geometric sequence of values – see Shawe-Taylor and Cristianini (1999b) for details.

Theorem 2.4

bound for optimal map

(Shawe-Taylor and Cristianini, 1999b) Fix $b \in \mathbb{R}$. 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

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

where

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

for $D = D(S,(\mathbf{u},b),\gamma)$, and provided $m \geq \max\{2/\epsilon, 6\}$, $h \leq em$ and there is no discrete probability on misclassified training points.

As discussed above the bound can be used for classifiers that misclassify some training points. The effect of misclassified points will only be felt in the value of D. Such points do not change the form of the expression. This is in contrast with traditional agnostic bounds which involve the square root of the ratio of the fat shattering dimension and sample size (see for example expression (**??**) in this book). If a point is an extreme outlier, it is possible that its effect on D might be such that the bound will be worse than that obtained using the agnostic approach (where the 'size' of misclassification is irrelevant). However, it is likely that in usual situations the bound given here will be significantly tighter than the standard agnostic one. The other advantage of the new bound will be discussed in the next section where we show that in contrast to the computational difficulty of minimizing the number of misclassifications, there exists an efficient algorithm for optimizing the value of h given in Theorem 2.4.

2.3 An Explanation for the Soft Margin Algorithm

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 choosing the hyperplane and value of γ, which result in a margin distribution in the original space for which the bound of Theorem 2.4 is minimal. It turns out that not only is the answer yes, but also that such an algorithm already exists.

separation kernels

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

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

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

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

and the Lagrange multipliers α_i would be obtained by solving the Quadratic Programming problem of minimizing in the positive quadrant the dual objective function:

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

This is exacly the dual QP problem that one would obtain by solving the soft margin problem in one of the cases stated in the appendix of Cortes and Vapnik (1995):

soft margin

$$\begin{aligned} \text{minimize}: \quad & \frac{1}{2} \langle \mathbf{u}, \mathbf{u} \rangle + C \sum \xi_i^2 \\ \text{subjectto}: \quad & y_j [\langle \mathbf{u}, x_j \rangle - b] \geq 1 - \xi_j \\ & \xi_i \geq 0 \end{aligned}$$

The solution they obtain is:

$$L = \sum \alpha_i - \sum y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \frac{1}{4C} \sum \alpha_i^2$$

which makes clear how the trade off parameter C in their formulation is related to the kernel parameter Δ.

2.4 Related Techniques

Another way of looking at this technique is that optimizing the soft margin, or enlarging the margin distribution, is equivalent to replacing the covariance matrix K with the covariance K'

covariance of augmented data

$$K' = K + \lambda I$$

which has a heavier diagonal. Again, there is a simple relationship between the trade off parameter λ and the Δ and C of 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 approach has also been considered by Smola and Schölkopf (1998) for the regression case where they also introduce an upper bound on the size of the α's in order to improve robustness to outliers.

Figure 2.4 shows the results of experiments performed on the ionosphere data of the UCI repository (Merz and Murphy, 1998). The plot is of the generalization error for different values of the parameter λ.

equivalent techniques

This technique is well known in classical statistics, where it is sometimes called the "shrinkage method" (see Ripley (1996)). Basically, in Bayesian discrimination (see section **??**) it suggests replacing the empirical covariance function Σ with some function closer to the identity I, by choosing an element of the line joining them $(1-\lambda)\Sigma + \lambda I$. A redundant degree of freedom is then removed, leaving with the new covariance $\Sigma + \lambda I$. In the case of linear regression this technique, known as ridge regression, can be derived from assuming Gaussian noise on the target values. It was originally motivated by the trade off between bias and variance (Hoerl and Kennard, 1970) and leads to a form of weight decay. This approach is equivalent to a form of regularization in the sense of Tikhonov. The theory of ill-posed problems was developed by Tikhonov in the context of solving inverse problems (Tikhonov and Arsenin, 1977). Smola and Schölkopf (1998) derived ridge regression using dual variables and for example C. Saunders (1998) have applied this to benchmark problems. It is well known that one can perform regularization by replacing the covariance matrix $X^T X$ with $X^T X + \lambda I$, and learning machines based on Gaussian Processes implicitly exploit this fact in addition to the choice of kernel.

Another explanation proposed for the same technique 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 than the hard margin case, since the eigenvalues are all increased by λ. The so-called box constraint algorithm which minimises the 1-norm of the slack variables is not directly comparable with the 2-norm case considered here.

Remark 2.1
Note that

$$R\sqrt{\sum \xi_i^2} = RD = \Delta^2 = \lambda = \frac{1}{4C}$$

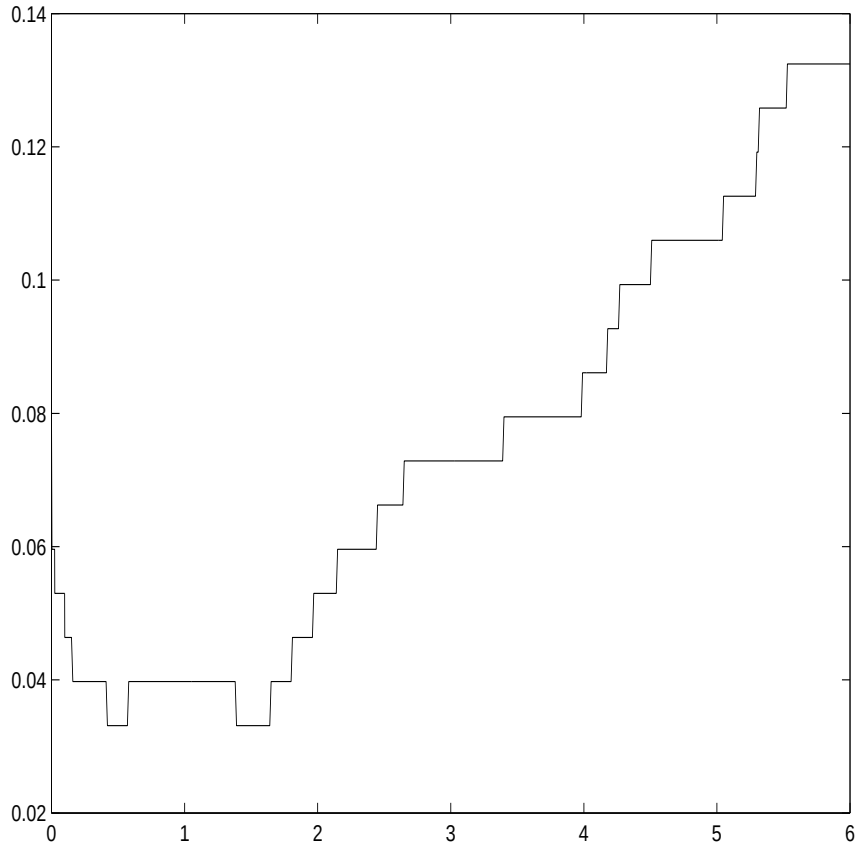

Figure 2.1 Generalization error as a function of λ, in a hard margin problem with augmented covariance $K' = K + \lambda I$, for *ionosphere* data.

so a choice of γ in the margin distribution bound controls the parameter C in the soft margin setting, and the trade-off parameter λ in the regularization setting. A reasonable choice of γ can be one that minimizes some VC bound on the capacity, for example maximising the margin in the augmented space, or controlling other parameters (margin; eigenvalues; radius; etc). Note also that this formulation also makes intuitive sense: a small γ corresponds to a small λ and to a large C: little noise is assumed, and so there is little need for regularization; vice versa a large γ corresponds to a large λ and a small C, which corresponds to assuming a high level of noise. Similar reasoning leads to similar relations in the regression case.

2.5 Conclusion

The analysis we have presented provides a principled way to deal with noisy data in large margin classifiers, and justifies the like the soft margin algorithm as originally proposed by Cortes and Vapnik. We have proved that one such algorithm exactly minimizes the bound on generalization provided by our margin distribution analysis, and is equivalent to using an augmented version of the kernel. Many techniques developed for the hard margin case can then be extended to the soft-margin case, as long as the quantities they use can be measured in terms of the modified kernel (margin, radius of the ball, eigenvalues).

The algorithms obtained in this way are strongly related to regularization techniques, and other methods developed in different frameworks in order to deal with noise. Computationally, the algorithm can be more stable and better conditioned than the standard maximal margin approach.

Finally, the same proof technique can also be used to produce analogous bounds for nonlinear functions in the classification case, and for the linear and nonlinear regression case with different losses, as reported in the full paper (Shawe-Taylor and Cristianini, 1998).

Acknowledgements

This work was supported by the European Commission under the Working Group Nr. 27150 (NeuroCOLT2) and by the UK EPSRC funding council. The authors would like to thank Colin Campbell and Bernhard Schölkopf for useful discussions. They would also like to thank useful comments from an anonymous referee that helped to refine Definition 2.2.

References

S. Arora, L. Babai, J. Stern, and Z. Sweedyk. Hardness of approximate optima in lattices, codes, and linear systems. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 43–54, Cambridge, MA, 1999. MIT Press.

P. 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 Transactions on Information Theory*, 44(2):525–536, 1998.

V. Vovk C. Saunders, A. Gammermann. Ridge regression learning algorithm in dual variables. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, San Francisco, CA, 1998. Morgan Kaufmann.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273 – 297, 1995.

N. Cristianini, J. Shawe-Taylor, and P. Sykacek. Bayesian classifiers are large margin hyperplanes in a hilbert space. In J. Shavlik, editor, *Machine Learning: Proceedings of the Fifteenth International Conference*, San Francisco, CA, 1998. Morgan Kaufmann.

Y. Freund and R.E. Schapire. Large margin classification using the perceptron algorithm. In J. Shavlik, editor, *Machine Learning: Proceedings of the Fifteenth International Conference*, San Francisco, CA, 1998. Morgan Kaufmann.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in banach spaces. In *Proceedings of Algorithm Learning Theory, ALT-97*, pages 352–363. Springer Verlag, 1997.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

C. J. Merz and P. M. Murphy. UCI repository of machine learning databases, 1998. [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University

Press, Cambridge, 1996.

R. Schapire, Y. Freund, P. Bartlett, and W. Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1998. (To appear. An earlier version appeared in: D.H. Fisher, Jr. (ed.), Proceedings ICML97, Morgan Kaufmann.).

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

J. Shawe-Taylor and N. Cristianini. Further results on the margin distribution. In *Proceedings of the Twelth Annual Conference on Computational Learning Theory, COLT'99*, 1999a.

J. Shawe-Taylor and N. Cristianini. Margin distribution bounds on generalization. In *Proceedings of the European Conference on Computational Learning Theory, EuroCOLT'99*, pages 263–273, 1999b.

J. Shawe-Taylor and Nello Cristianini. Robust bounds on generalization from the margin distribution. NeuroCOLT Technical Report NC-TR-1998-020, ESPRIT NeuroCOLT2 Working Group, http://www.neurocolt.com, 1998.

A. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211 – 231, 1998.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. W. H. Winston, Washington, D.C., 1977.