Adaptive Margin Support Vector Machines for Classification

Ralf Herbrich*,†, Jason Weston*

* Department of Computer Science, Royal Holloway, University of London
† Department of Computer Science, Technical University of Berlin
{ralfh,jasonw}@dcs.rhbnc.ac.uk

Abstract

In this paper we propose a new learning algorithm for classification learning based on the Support Vector Machine (SVM) approach. Existing approaches for constructing SVMs [12] are based on minimization of a regularized margin loss where the margin is treated *equivalently* for each training pattern. We propose a reformulation of the minimization problem such that *adaptive margins* for each training pattern are utilized, which we call the Adaptive Margin (AM–) SVM. We give bounds on the generalization error of AM–SVMs which justify their robustness against outliers, and show experimentally that the generalization error of AM–SVMs is comparable to classical SVMs on benchmark datasets from the UCI repository.

1 Introduction

Recently, the study of classification learning has shown that algorithms which learn a *real–valued* function for classification can control their generalization error by making use of a quantity known as the *margin*. Based on these results, Support Vector Machines which *directly* control the margin have been proven to be successful in classification learning [4, 12, 10]. Moreover, it turned out to be favourable to formulate the decision functions in terms of a symmetric, positive definite, and square integrable function $k(\cdot,\cdot)$ referred to as a *kernel*. The class of decision functions — also known as *kernel classifiers* [11, 3] — is then given by[1]

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) \qquad \boldsymbol{\alpha} \geq \mathbf{0}\,. \tag{1}$$

Whilst the algorithms proposed so far are restricted to a *fixed margin* at each training pattern $(\mathbf{x}_i, y_i)$, we show that *adaptive* margins can successfully be used. Moreover, it turns out that adaptive margins effectively control the complexity of the model. The paper is structured as follows: In Section 2 we present the algorithm for Adaptive Margin Support Vector Machines (AM–SVMs) and reveal their relation to classical SVMs. In the following section we give bounds on the generalization error of AM–SVMs which justify the use of adaptive margins as a regularizer. In Section 4 results of a comparison of AM–SVMs with classical SVMs on benchmark datasets from the UCI repository are presented. Finally, in Section 5 we summarize the paper and discuss further directions.

2 Adaptive Margin SVMs

In a classification task one's ultimate goal is to find a function f that minimizes the *expected* risk functional

$$R(f) := \mathbf{E}_P L(f(\mathbf{x}), y)\,, \tag{2}$$

where we assume a distribution $P(\mathbf{x}, y)$. Here, the loss function $L(\cdot,\cdot)$ is assumed to be given. It is well known that this problem cannot be

[1] Although this class of functions is dependent on the training set, the restrictions put on $k(\cdot,\cdot)$ automatically ensure that the influence of each *new* basis function $k(\mathbf{x}_i,\cdot)$ decreases rapidly for increasing training set sizes ℓ. Thus we can assume the existence of a *fixed* feature space (see also [1]).

solved directly because $P(\mathbf{x}, y)$ is generally unknown. Instead, we are given an i.i.d. training set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\} \subset \mathcal{X} \times \{-1, +1\}$ and try to find some suitable f_{emp} based thereon. Minimization of the empirical risk

$$R_{\text{emp}}(f) := \mathbf{E}_S L(f(\mathbf{x}_i), y_i) \tag{3}$$

is an ill–posed problem [12] and thus may lead to solutions with a high expected risk $R(f_{\text{emp}})$. An approach to overcome this difficulty (also known as regularization) is to put further restrictions on the functions f. This can be achieved by adding a regularizer $Q(f)$ which effectively restricts the choice of models. Hence for some fixed $\lambda \geq 0$ learning aims at minimizing

$$R_{\text{reg}}(f) := R_{\text{emp}}(f) + \lambda Q(f)\,. \tag{4}$$

It was shown elsewhere [8] that tight bounds on $R(f_{\text{reg}})$ can be obtained making use of the real value returned by f_{reg}. Using the soft margin loss

$$L(f(\mathbf{x}), y) = \max(1 - yf(\mathbf{x}), 0) \tag{5}$$

introduced in [11] we can derive the following algorithms.

Quadratic (QP–) SVMs Using the quadratic regularization functional

$$Q_{\text{QP}}(f) = \|\mathbf{w}\|_2^2\,,$$

we obtain the following class of SVMs

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{\ell} \xi_i + \lambda \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} & y_i f(\mathbf{x}_i) \geq 1 - \xi_i \\ & \alpha_i, \xi_i \geq 0. \end{array}$$

Here we used

$$\mathbf{w} = \sum_{j=1}^{\ell} \alpha_j y_j \phi(\mathbf{x})\,, \tag{6}$$

where $\phi(\cdot)$ maps into a feature space $\mathcal{F}$ such that $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}} = k(\mathbf{x}, \mathbf{x}')$. It is known that $Q_{\text{QP}}(f)$ controls the VC–Dimension of the induced loss–function class $\{L(f(\cdot), \cdot)\}$ [11, 8]. This choice of regularizer favours flat functions in feature space.

Linear (LP–) SVMs Using the linear regularization functional

$$Q_{\text{LP}}(f) = \sum \alpha_i \tag{7}$$

results in the class of linear SVMs, i.e.

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{\ell} \xi_i + \lambda \sum_{i=1}^{\ell} \alpha_i \\ \text{subject to} & y_i f(\mathbf{x}_i) \geq 1 - \xi_i \\ & \alpha_i, \xi_i \geq 0 \end{array}$$

Recently it was shown that $Q_{\text{LP}}(f)$ can also be used to control the VC–Dimension of $\{L(f(\cdot), \cdot)\}$ [10]. In contrast to the quadratic regularizer, $Q_{\text{LP}}(f)$ favours non–smooth functions by strong penalizing of basis functions $\phi_j(\cdot)$ with a small eigenvalue [10].

Adaptive Margin (AM–) SVMs In both types of learning algorithm the margin error ξ_i at point $(\mathbf{x}_i, y_i)$ and the regularization on α_i are additive and therefore independently treated. To make the size of the margin at each training point a controlling variable we propose the following learning algorithm (AM–SVM)

$$\begin{array}{lrcl} \text{minimize} & \sum_{i=1}^{\ell} \xi_i & & \\ & y_i f(\mathbf{x}_i) & \geq & 1 - \xi_i + \lambda \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) \\ \text{subject to} & \xi_i & \geq & 0 \\ & \alpha_i & \geq & 0 \end{array}$$

This algorithm can be viewed in the following way (see Figure 1): Suppose the data lives on the surface of a hypersphere in $\mathcal{F}$. Then $k(\mathbf{x}_i, \mathbf{x}_j)$ is the cosine of the angle between $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$. As soon as a point $\phi(\mathbf{x}_k)$ is an outlier (the cosine of the angles to points in its class are small and to points in the other class are large) α_k in Equation (1) has to be large in order to classify $\phi(\mathbf{x}_k)$ correctly. Whilst SVMs use the same margin for such an outlier, they attempt to classify $\phi(\mathbf{x}_k)$ correctly. In AM–SVMs the margin is automatically increased to $1 + \lambda \alpha_k k(\mathbf{x}_i, x_i)$ for $\phi(\mathbf{x}_k)$ and thus less attempt is made to change the decision function. Moreover, it becomes clear that in AM–SVMs the points $\phi(\mathbf{x}_k)$ which are representatives of clusters (centers) in feature space $\mathcal{F}$, i.e. those which have large values of the cosine of the angles to points from its class, will have nonzero α_k. It is worthwhile to study the influence of λ:

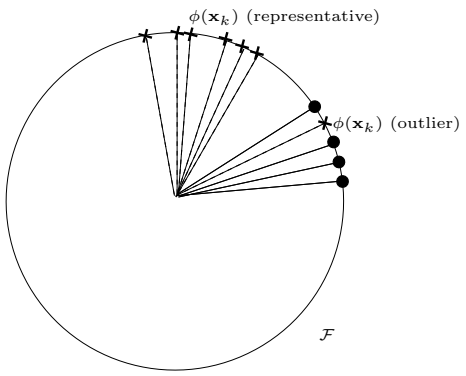

Figure 1: Adaptation of margins at each training pattern depending on the distance $k(\mathbf{x}_i, \mathbf{x}_j)$ in feature space $\mathcal{F}$. Note that $k(\mathbf{x}_i, \mathbf{x}_j)$ is large if the enclosed angle between data points is small. See the text for explanation.

- If $\lambda = 0$ no adaptation of the margins is performed. This is in correspondence to Equation (4) which implies $R_{\text{emp}}(f) = R_{\text{reg}}(f)$.
- If $\lambda \to \infty$ the margin at each point tends to infinity $(1 + \lambda\alpha_i k(\mathbf{x}_i, \mathbf{x}_i))$ and thus setting all α's to an equal and small value is the solution of AM–SVMs. This corresponds to paying *no* attention to $R_{\text{emp}}(f)$ and is equivalent to kernel density estimation on each class (Parzen windows) [5].
- If $\lambda = 1$ the resulting algorithm is equivalent to Leave–One–Out SVMs [13] motivated by the following bound on the leave–one–out error of QP–SVMs [3].

  Theorem 1. *For any training set $S = \{x_i, y_i\}_{i=1}^{\ell}$ with $y_i \in \{-1, +1\}$, using a kernel–classifier given by Equation* (1) *and SVMs with L_2 norm for learning, the leave-one-out error estimate of the classifier is bounded by*

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\theta\left(-y_i\sum_{j\neq i}y_j\alpha_j k(\mathbf{x}_i,\mathbf{x}_j)\right), \qquad (8)$$

  where $\theta(\cdot)$ is the step function.

  Note that Theorem 1 is not valid for AM–SVMs with $\lambda = 1$.

3 Theoretical Analysis

To obtain margin distributions for Adaptive Margin Machines we apply the following theorem to be found in [9]:

Theorem 2. *Consider a fixed but unknown probability distribution on the input space $\mathcal{X}$ with support in the ball of radius R about the origin. Then with probability $1-\delta$ over randomly drawn training sets S of size ℓ for all $\gamma > 0$ such that $d((\mathbf{x}, y), \mathbf{w}, \gamma) = 0$, for some $(\mathbf{x}, y) \in S$, the generalization of a linear classifier $\mathbf{w}$ on $\mathcal{X}$ satisfying $\|\mathbf{w}\|_{\mathcal{F}} \leq 1$ is bounded by*

$$\frac{2}{\ell}\left(\kappa\log\left(\frac{8e\ell}{\kappa}\right)\log(32\ell) + \mathcal{O}\left(\log\left(\frac{\ell\log(\ell)}{\delta}\right)\right)\right),$$

where

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

$$\begin{aligned} D = D(S, \mathbf{w}, \gamma) &= \sqrt{\sum_{i=1}^{\ell} d_i^2} \\ d_i = d((\mathbf{x}_i, y), \mathbf{w}, \gamma) &= \max\{0, \gamma - y\langle \mathbf{w}, \mathbf{x}_i\rangle_{\mathcal{F}}\} \end{aligned}$$

and provided $\ell \geq \max\{2/\epsilon, 6\}$ and $\kappa \leq e\ell$.

Applying the bound to AM–SVMs we obtain the following theorem.

Theorem 3. *Consider a fixed but unknown probability distribution on the input space $\mathcal{X}$ with support in the ball of radius R about the origin. Then with probability $1-\delta$ over randomly drawn training set S of size ℓ for $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\boldsymbol{\xi} \geq 0$ which are feasible solutions of AM–SVMs such that $d((\mathbf{x}, y), \mathbf{w}, 1) = 0$, for some $(\mathbf{x}, y) \in S$, the generalization error $R(f)$ is bounded by*

$$\frac{2}{\ell}\left(\kappa\log\left(\frac{8e\ell}{\kappa}\right)\log(32\ell) + \mathcal{O}\left(\log\left(\frac{\ell\log(\ell)}{\delta}\right)\right)\right),$$

where

$$\begin{aligned} \kappa &\leq \left\lfloor 65[(WR + 3D)^2] \right\rfloor, \\ D &= \sqrt{\sum_{i=1}^{\ell}[\max\{0, \xi_i - \lambda\alpha_i k(\mathbf{x}_i, \mathbf{x}_i)\}]^2}, \\ W^2 &= \sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \end{aligned}$$

provided $\ell \geq \max\{2/R(f), 6\}$ *and* $\kappa \leq e\ell$.

Proof. Firstly, AM–SVMs learn linear classifiers $f(x) = \langle \mathbf{w}, \phi(x) \rangle_{\mathcal{F}}$ where $\mathbf{w}$ is defined by Equation (6). We wish to redefine the measure of margin error $d((\mathbf{x}, y), \mathbf{w}, \gamma) = \gamma - y_i f(\mathbf{x}_i)$ in Theorem 2 in terms of ξ_i and $\lambda\alpha_i k(\mathbf{x}_i, \mathbf{x}_i)$ to capture the adaptive margin of a training point $\mathbf{x}_i$. Then we know from the assumption of a feasible solution $\boldsymbol{\alpha}, \boldsymbol{\xi}$ that

$$\begin{aligned} \max\{0, \gamma - y_i f(\mathbf{x}_i))\} \quad &\leq \\ \max\{0, \gamma - 1 + \xi_i - \lambda\alpha_i k(\mathbf{x}_i, \mathbf{x}_i)\} \quad &. \end{aligned}$$

In order to apply Theorem 2 for *any* vector $\mathbf{w}$ we have to divide γ, D, and $\boldsymbol{\alpha}$ by $W = \|\mathbf{w}\|_{\mathcal{F}}$. This gives

$$\kappa = \left\lfloor \frac{65[(R + \frac{1}{W}D)^2 + 2.25\frac{1}{W}RD]}{\gamma^2} W^2 \right\rfloor .$$

This allows us to fix $\gamma = 1$ without loss of generality. Making use of

$$[(R + \frac{1}{W}D)^2 + 2.25\frac{1}{W}RD]W^2 \leq [(WR + 3D)^2] ,$$

the theorem is proven. □

From the theorem, one can gain the following insights. Our goal to minimize generalization error is achieved by minimizing κ, the minimum of which is a trade off between minimizing W (the margin) and D (the loss with adaptive margin). We require a small value of both but small values of one term typically give large values of the other. By minimizing $\sum_{i=1}^{\ell} \xi_i$ AM–SVMs effectively control the trade-off between the two terms through the parameter λ. For small values of λ, D is small and W can take any value as it is not minimized (it can be forced to very large values). For large λ the boosted margin in D acts a regularizer, penalizing large values of α. This results in small values of W (a smooth function) but large values of D (large training error). This bound motivates the objective function of AM–SVMs which at first appear to only minimize error and have no regularization. In fact, as we have seen, the regularization comes from the adaptive margin in the constraints controlled by λ.

4 Experiments

Artificial Data We first describe some two dimensional examples to illustrate how the new regularization technique works. We generated a two class problem in $\mathbb{R}^2$ (represented by crosses and dots). We trained an AM–SVM using RBF–kernels ($\sigma = 0.5$) with $\lambda = 1, 2, 5, 10$ (see Figure 2). As can be seen increasing λ allows AM–SVM to widen the margin for points far away from the decision surface. Consequently, the algorithm is more robust to outliers which results in very smooth decision functions. In Figure 3 we used the same dataset and trained ν LP–SVMs [1]. ν LP–SVMs are obtained by reparameterizing Equation (2) where ν upper–bounds the number of margin errors (see [1]). Varying $\lambda = 0.0, 0.1, 0.2, 0.5$ shows that *margin* errors are sacrificed in order to lower the complexity of the decision function f measured in the one–norm (see Equation (7)). As we have already mentioned this leads to non–smooth functions. Furthermore it should be noted that the outlier (dot) on the far left side leads to very rugged decision functions. Similar conclusions can be drawn for ν QP–SVMs [7] (see Figure 4) though the decision functions are smoother. Thus, AM–SVMs turn out to be more robust than classical SVM.

	AB	AB_R	SVM	AM–SVM
Banana	12.3	10.9	11.5	10.6
B. Cancer	30.4	26.5	26.0	26.3
Diabetes	26.5	23.9	23.5	23.4
Heart	20.3	16.6	16.0	16.1
Thyroid	4.4	4.4	4.8	5.0
Titanic	22.6	22.6	22.4	22.7

Table 1: Comparison of percentage test error of AdaBoost (AB), Regularized AdaBoost (AB_R), classical (QP–) Support Vector Machines (SVM) and Adaptive Margin Support Vector Machines with fixed $\lambda = 1$ (AM–SVM) on 6 datasets.

Benchmark Datasets We conducted computer simulations using datasets from the UCI, DELVE and STATLOG benchmark repositories, following the same experimental setup as in [6]. Briefly, the setup is as follows: the performance of

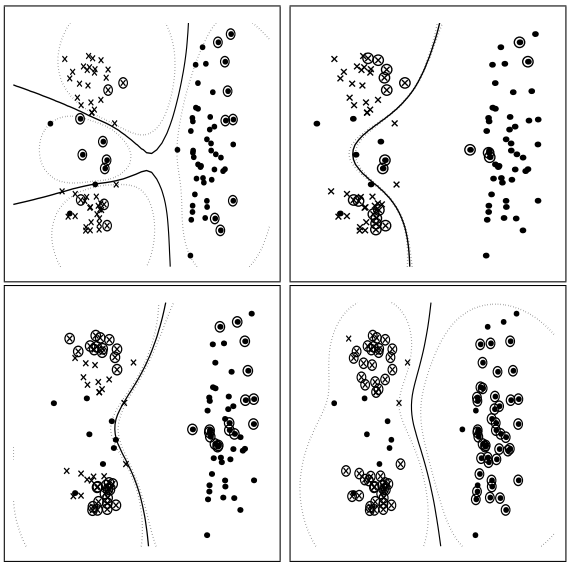

Figure 2: Decision functions (solid lines) obtained by AM–SVMs with different choices of the regularization parameter λ. The dashed line represents the minimal margin over all training points. **(a)** $\lambda = 1$ is equivalent to setting the diagonal elements of the kernel matrix to zero. **(b)** $\lambda = 2$, **(c)** $\lambda = 5$, and **(d)** $\lambda = 10$ widens the amount to which margin errors at each point are accepted and thus results in very flat functions. Note, that less attention is paid to the outlier (dot) at the left hand side.

a classifier is measured by its average error over one hundred partitions of the datasets into training and testing sets. Free parameter(s) in the learning algorithm are chosen as the median value of the best model chosen by cross validation of the first five training datasets. For our comparison we fixed the parameter $\lambda = 1$ for AM–SVMs.

Table 1 compares percentage test error of AM–SVMs to AdaBoost (AB), Regularized AdaBoost (AB_R) and SVMs which are all known to be excellent classifiers[2]. AM–SVMs (even with fixed $\lambda = 1$) were very competitive with SVMs and AB_R (for which the best parameters were found by cross-validation). This indicates tuning of λ could give even better performance for AM–SVMs. AdaBoost, which has no regularisation parameter, is outperformed by the other three algorithms.

[2]The results for AB, AB_R and SVMs were taken from [6].

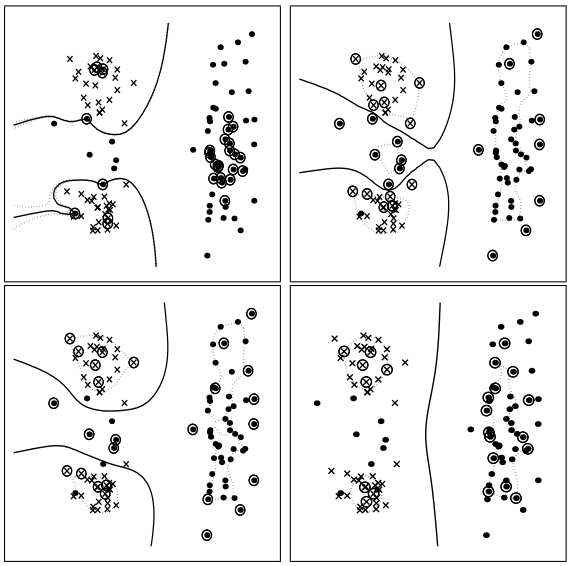

Figure 3: Decision functions (solid lines) obtained by ν LP–SVMs with different choices of the assumed noise level ν. The dashed line represents the margin. **(a)** $\nu = 0.0$ leads to very non–smooth and overfitted decision functions. **(b)** $\nu = 0.1$, **(c)** $\nu = 0.2$, and **(d)** $\nu = 0.5$ smooth the decision function.

5 Discussion

In this paper we presented a new learning algorithm for kernel classifiers. This approach pushed the idea of capacity control via margin maximization to its limit by allowing adapting margins at each training pattern. We have shown experimentally that this reformulation results in an algorithm which is very robust against outliers. Nevertheless, our algorithm has a parameter λ which needs to be optimized for a given learning problem. Further investigations will be made in the derivation of bounds on the leave–one–out error of this algorithm which allows for efficient model order selection. To gain more insight into the role of the parameter λ it seems worthwhile to cast the algorithm in a regularization framework (see Section 2). Finally, we want to note that penalization of the diagonal of the kernel matrix is a well known technique in regression known as ridge regression [2]. Hence, penalizing the diagonal of the kernel matrix results in orthogonal data vectors in feature space which is a commonly used technique of

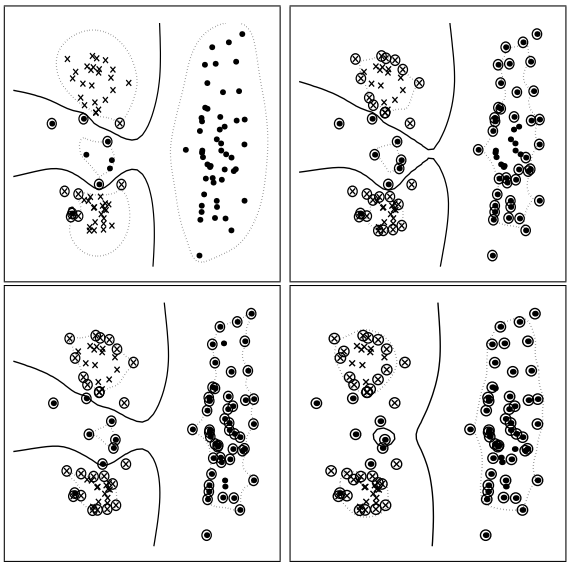

Figure 4: Decision functions (solid lines) obtained by ν QP–SVMs with different choices of the assumed noise level ν. The dashed line represents the margin. **(a)** $\nu = 0.0$ leads to an overfitted decision functions (note the captured outlier in the lower left region). **(b)** $\nu = 0.1$, **(c)** $\nu = 0.2$, and **(d)** $\nu = 0.5$ allow for much flatter functions though regularizing differently to AM–SVMs.

regularization.

Acknowledgments We are indebted to both John Shawe–Taylor and Vladimir Vapnik for their help. The authors would also like to thank Alex Gammerman, Tom Melluish, and Craig Saunders for discussions. Ralf Herbrich would like to thank the Department of Computer Science at Royal Holloway for the warm hospitality during his research stay. Jason Weston is supported by the ESPRC through grant GR/L35812.

References

[1] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K. Robert-Müller, K. Obermayer, and B. Williamson. Classification on proximity data with LP–machines. In *Proceedings of ICANN 99*, 1999. accepted for publication.

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

[3] T. S. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*. Morgan Kaufmann, 1999.

[4] L. Mason and P. Bartlett. Direct optimization of margins. In *Advances in Neural Information Processing Systems*, page in press, San Mateo, CA, 1998. Morgan Kaufmann.

[5] E. Parzen. On estimation of probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[6] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. Technical report, Royal Holloway, University of London, 1998. TR–98–21.

[7] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. Technical report, Royal Holloway, University of London, 1998. CSD–TR–98–31.

[8] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data–dependent hierarchies. Technical report, Royal Holloway, University of London, 1996. NC–TR–1996–053.

[9] J. Shawe-Taylor and N. Cristianini. Margin distribution bounds on generalization. Technical report, Royal Holloway, University of London, 1998. NC2–TR–1998–020.

[10] A. J. Smola. *Learning with Kernels*. PhD thesis, Technical University Berlin, Berlin, Germany, 1998.

[11] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer–Verlag, New York, 1995.

[12] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[13] J. Weston. Leave–one–out support vector machines. In *Proccedings of the International Joint Conference on Artifical Intelligence*, Sweden, 1999.