2 Bounds on Error Expectation for Support Vector Machines

Vladimir Vapnik
AT&T Labs
Red Bank, NJ 07701
vlad@research.att.com

Olivier Chapelle
École Normale Supérieure de Lyon
69364 Lyon Cedex 07, France
ochapell@ens-lyon.fr

We introduce the concept of span of support vectors (SV) and show that the generalization ability of support vector machines (SVM) depends on this new geometrical concept. We prove that the value of the span is always smaller (and can be much smaller) than the diameter of the smallest sphere containing the support vectors, used in previous bounds (Vapnik, 1998). We also demonstate experimentally that the prediction of the test error given by the span is very accurate and has direct application in model selection (choice of the optimal parameters of the SVM)

2.1 Introduction

Recently, a new type of algorithm with a high level of performance called Support Vector Machines (SVM) has been introduced (Boser et al., 1992; Vapnik, 1995).

Usually, the good generalization ability of SVM is explained by the existence of a large margin : bounds on the error rate for a hyperplane that separates the data with some margin were obtained in (Bartlett and Shawe-Taylor, 1999; Shawe-Taylor et al., 1998). In Vapnik (1998), another type of bound was obtained which demonstrated that for the separable case the expectation of probability of error for hyperplanes passing through the origin depends on the expectation of R^2/ρ^2, where R is the maximal norm of support vectors and ρ is the margin.

In this paper we derive bounds on the expectation of error for SVM from the *leave-one-out* estimator, which is an unbiased estimate of the probability of test error. These bounds (which are tighter than the one defined in Vapnik (1998) and valid for hyperplanes not necessarily passing trough the origin) depend on a new concept called *the span of support vectors*.

The bounds obtained show that the generalization ability of SVM depends on more complex geometrical constructions than large margin.

To introduce the concept of the span of support vectors we have to describe the basics of SVM.

2.2 SVM for Pattern Recognition

We call the hyperplane

$$\mathbf{w}_0 \cdot \mathbf{x} + b_0 = 0$$

optimal hyperplane

optimal if it separates the training data

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_\ell, y_\ell), \quad \mathbf{x} \in \mathbb{R}^m,\ y \in \{-1, 1\}$$

and if the margin between the hyperplane and the closest training vector is maximal. This means that the optimal hyperplane has to satisfy the inequalities

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, ..., \ell$$

and has to minimize the functional

$$R(\mathbf{w}) = \mathbf{w} \cdot \mathbf{w}.$$

dual formulation

This quadratic optimization problem can be solved in the dual space of Lagrange multipliers. One constructs the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} - \sum_{i=1}^{\ell} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] \tag{2.1}$$

and finds its saddle point: the point that minimizes this functional with respect to $\mathbf{w}$ and b and maximizes it with respect to

$$\alpha_i \geq 0, \quad i = 1, ..., \ell. \tag{2.2}$$

Minimization over $\mathbf{w}$ defines the equation

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x}_i \tag{2.3}$$

and minimization over b defines the equation

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0. \tag{2.4}$$

Substituting (2.3) back into the Lagrangian (2.1) and taking into account (2.4), we obtain the functional

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j, \tag{2.5}$$

which we have to maximize with respect to parameters $\boldsymbol{\alpha}$ satisfying two constraints: equality constraint (2.4) and positivity constraints (2.2). The optimal solution $\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_\ell^0)$ specifies the coefficients for the optimal hyperplane

$$\mathbf{w}_0 = \sum_{i=1}^{\ell} \alpha_i^0 y_i \mathbf{x}_i.$$

decision function

Therefore the optimal hyperplane is

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i \mathbf{x}_i \cdot \mathbf{x} + b_0 = 0, \tag{2.6}$$

where b_0 is chosen to maximize the margin.

It is important to note that the optimal solution satisfies the Kuhn-Tucker conditions

$$\alpha_i^0 [y_i(\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) - 1] = 0.$$

From these conditions it follows that if the expansion of vector $\mathbf{w}_0$ uses vector $\mathbf{x}_i$ with non-zero weight α_i^0 then the following equality must hold

$$y_i(\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) = 1. \tag{2.7}$$

Vectors $\mathbf{x}_i$ that satisfy this equality are called *support vectors*.

margin

Note that the norm of vector $\mathbf{w}_0$ defines the margin ρ between optimal separating hyperplane and the support vectors

$$\rho = \frac{1}{||\mathbf{w}_0||}.$$

Therefore taking into account (2.4) and (2.7) we obtain

$$\frac{1}{\rho^2} = \mathbf{w}_0 \cdot \mathbf{w}_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 \mathbf{w}_0 \cdot \mathbf{x}_i = \sum_{i=1}^{\ell} y_i \alpha_i^0 (\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) = \sum_{i=1}^{\ell} \alpha_i^0 \tag{2.8}$$

where ρ is the margin for the optimal separating hyperplane.

non-separable case

In the non-separable case we introduce slack variables ξ_i and minimize the functional

$$R(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C \sum_{i=1}^{\ell} \xi_i$$

subject to constraints

$$y_i(\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) \geq 1 - \xi_i$$

$\xi_i \geq 0$.

When the constant C is sufficiently large and the data is separable, the solution of this optimization problem coincides with the one obtained for the separable case.

To solve this quadratic optimization problem for the non-separable case, we consider the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} - \sum_{i=1}^{\ell} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] + C \sum_{i=1}^{\ell} \xi_i - \sum_{i=1}^{\ell} \nu_i \xi_i,$$

which we minimize with respect to $\mathbf{w}$, b and ξ_i and maximize with respect to the Lagrange multipliers $\alpha_i \geq 0$ and $\nu_i \geq 0$.

The result of minimization over $\mathbf{w}$ and b leads to the conditions (2.3) and (2.4) and result of minimization over ξ_i gives the new condition

$$\alpha_i + \nu_i = C. \tag{2.9}$$

Taking into account that $\nu_i \geq 0$, we obtain

$$0 \leq \alpha_i \leq C. \tag{2.10}$$

Substituting (2.3) into the Lagrangian, we obtain that in order to find the optimal hyperplane, one has to maximize the functional (2.5), subject to constraints (2.4) and (2.10).

The box constraints (2.10) (instead of the positivity constraints (2.2)) entail the difference in the methods for constructing optimal hyperplanes in the non-separable case and in the separable case respectively.

For the non-separable case, the Kuhn - Tucker conditions

$$\alpha_i^0 [y_i(\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) - 1 + \xi_i] = 0$$

$$\nu_i \xi_i = 0 \tag{2.11}$$

must be satisfied. Vectors $\mathbf{x}_i$ that correspond to nonzero α_i^0 are referred as support vectors. For support vectors the equalities

$$y_i(\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) = 1 - \xi_i$$

hold. From conditions (2.11) and (2.9) it follows that if $\xi_i > 0$, then $\nu_i = 0$ and therefore $\alpha_i = C$.

category of a support vector

We will distinguish between two types of support vectors: support vectors for which $0 < \alpha_i^0 < C$ and support vectors for which $\alpha_i^0 = C$. To simplify notations we sort the support vectors such that the first n^* support vectors belong to the first category (with $0 < \alpha_i < C$) and the next $m = n - n^*$ support vectors belong to the second category (with $\alpha_i = C$).

When constructing SVMs one usually maps the input vectors $\mathbf{x} \in X$ into a high dimensional (even infinite dimensional) feature space $\phi(\mathbf{x}) \in \mathcal{F}$ where one constructs the optimal separating hyperplane. Note that both the optimal hyperplane (2.6) and the target functional (2.5) that has to be maximized to find

non linear SVM

the optimal hyperplane depend on the inner product between two vectors rather than on input vectors explicitly. Therefore one can use the general representation of inner product in order to calculate it. It is known that the inner product between two vectors $\phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2)$ has the following general representation

$$\phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2),$$

where $K(\mathbf{x}_1, \mathbf{x}_2)$ is a kernel function that satisfies the Mercer conditions (symmetric positive definite function). The form of kernel function $K(\mathbf{x}_1, \mathbf{x}_2)$ depends on the type of mapping of the input vectors $\phi(\mathbf{x})$. In order to construct the optimal hyperplane in feature space, it is sufficient to use a kernel function instead of inner product in expressions (2.5) and (2.6).

Further we consider bounds in the input space X. However all results are true for any mapping ϕ. To obtain the corresponding results in a feature space one uses the representation of the inner product in feature space $K(\mathbf{x}, \mathbf{x}_i)$ instead of the inner product $\mathbf{x} \cdot \mathbf{x}_i$.

2.3 The leave-one-out procedure

The bounds introduced in this paper are derived from the leave-one-out cross-validation estimate. This procedure is usually used to estimate the probability of test error of a learning algorithm.

leave-one-out procedure

Suppose that using training data of size ℓ one tries simultaneously to estimate a decision rule and evaluate the quality of this decision rule. Using training data, one constructs a decision rule. Then one uses the same training data to evaluate the quality of the obtained rule based on the *leave-one-out* procedure : one removes from the training data one element (say $(\mathbf{x}_p, y_p)$), constructs the decision rule on the basis of the remaining training data and then tests the removed element. In this fashion one tests all ℓ elements of the training data (using ℓ different decision rules). Let us denote the number of errors in the leave-one-out procedure by $\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)$.

Luntz and Brailovsky proved the following lemma :

Lemma 2.1 Luntz and Brailovsky (1969)
The leave-one-out procedure gives an almost unbiased estimate of the probability of test error

$$Ep_{error}^{\ell-1} = E\left(\frac{\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)}{\ell}\right),$$

where $p_{error}^{\ell-1}$ is the probability of test error for the machine trained on a sample of size $\ell - 1$.

"Almost" in the above lemma refers to the fact the probabilty of test error is for samples of size $\ell - 1$ insteal of ℓ.

Remark. For SVMs one needs to conduct the leave-one-out procedure only for support vectors: non support vectors will be recognized correctly since removing a

point which is not support vector does not change the decision function.

In section 2.5, we introduce upper bounds on the number of errors made by the leave-one-out procedure. For this purpose we need to introduce a new concept, called the *span of support vectors*.

2.4 Span of the set of Support Vectors

Let us first consider the separable case. Suppose that

$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$

is a set of support vectors and

$\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_n^0)$

is the vector of Lagrange multipliers for the optimal hyperplane.

For any fixed support vector $\mathbf{x}_p$ we define the set Λ_p as a constrained linear combinations of the points $\{\mathbf{x}_i\}_{i \neq p}$:

$$\Lambda_p = \left\{ \sum_{i=1,\ i\neq p}^{n} \lambda_i \mathbf{x}_i \ : \ \sum_{i=1,\ i\neq p}^{n} \lambda_i = 1, \text{ and } \forall i \neq p, \ \ \alpha_i^0 + y_i y_p \alpha_p^0 \lambda_i \geq 0 \right\}$$

Note that λ_i can be less than 0.

We also define the quantity S_p, which we call the *span* of the support vector $\mathbf{x}_p$
span as the distance between $\mathbf{x}_p$ and this set (see figure 2.1)

$$S_p^2 = d^2(\mathbf{x}_p, \Lambda_p) = \min_{\mathbf{x} \in \Lambda_p} (\mathbf{x}_p - \mathbf{x})^2, \tag{2.12}$$

As shown in figure 2.2, it can happen that $\mathbf{x}_p \in \Lambda_p$, which implies

$S_p = d(\mathbf{x}_p, \Lambda_p) = 0.$

Intuitively, for smaller $S_p = d(\mathbf{x}_p, \Lambda_p)$ the leave-one-out procedure is less likely to make an error on the vector $\mathbf{x}_p$. Indeed, we will prove (see lemma 2.3)) that if $S_p < 1/(D\alpha_p^0)$ (D is the diameter of the smallest sphere containing the training points), then the leave-one-out procedure classifies $\mathbf{x}_p$ correctly.

By setting $\lambda_p = -1$, we can rewrite S_p as :

$$S_p^2 = \min \left\{ \left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \right)^2 \ : \ \ \lambda_p = -1, \ \sum_{i=1}^{n} \lambda_i = 0, \ \alpha_i^0 + y_i y_p \alpha_p^0 \lambda_i \geq 0 \right\} \tag{2.13}$$

The maximal value of S_p is called the S-span

$S = \max\{d(\mathbf{x}_1, \Lambda_1), ..., d(\mathbf{x}_n, \Lambda_n)\} = \max_p S_p.$

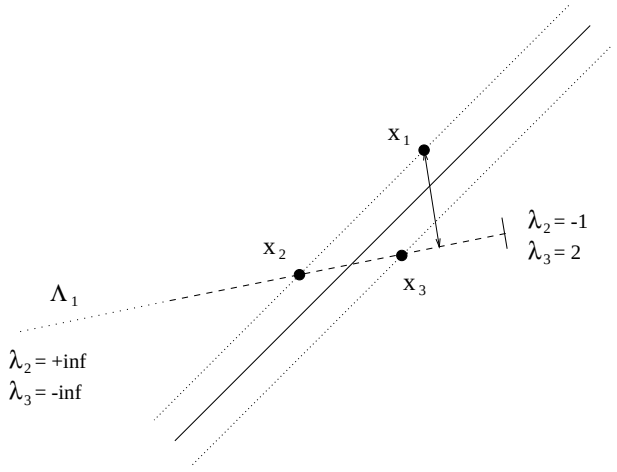


Figure 2.1 *Consider the 2D example above : 3 support vectors with* $\alpha_1 = \alpha_2 = \alpha_3/2$. *The set* Λ_1 *is the semi-opened dashed line :* $\Lambda_1 = \{\lambda_2\mathbf{x}_2 + \lambda_3\mathbf{x}_3,\ \lambda_2+\lambda_3 = 1,\ \lambda_2 \geq -1,\ \lambda_3 \leq 2\}$.

We will prove (cf lemma (2.2) below) that $S_p \leq D_{SV}$. Therefore,

$$S \leq D_{SV}. \tag{2.14}$$

Depending on $\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_n^0)$ the value of the span S can be much less than diameter D_{SV} of the support vectors. Indeed, in the example of figure 2, $d(\mathbf{x}_1, \Lambda_1) = 0$ and by symmetry, $d(\mathbf{x}_i, \Lambda_i) = 0$, for all i. Therefore in this example $S = 0$.

non-separable case

Now we generalize the span concept for the non-separable case. In the non-separable case we distinguish between two categories of support vectors: the support vectors for which

$$0 < \alpha_i < C \qquad i = 1, ..., n^*$$

and the support vectors for which

$$\alpha_j = C \qquad j = n^* + 1, ..., n.$$

We define the span of support vectors using support vectors of the first category.

That means we consider the value $S_p = d(\mathbf{x}_p, \Lambda_p)$ where

$$\Lambda_p = \left\{ \sum_{i=1,\ i\neq p}^{n^*} \lambda_i \mathbf{x}_i \ : \ \sum_{i=1,\ i\neq p}^{n^*} \lambda_i = 1,\ \forall i \neq p \ \ 0 \leq \alpha_i^0 + y_i y_p \alpha_p^0 \lambda_i \leq C \right\}$$

The differences in the definition of the span for the separable and the non-separable case are that in the non-separable case we ignore the support vectors of the second category and add an upper bound C in the constraints on λ_i.

Therefore in the non-separable case the value of the span of support vectors depends on the value of C.

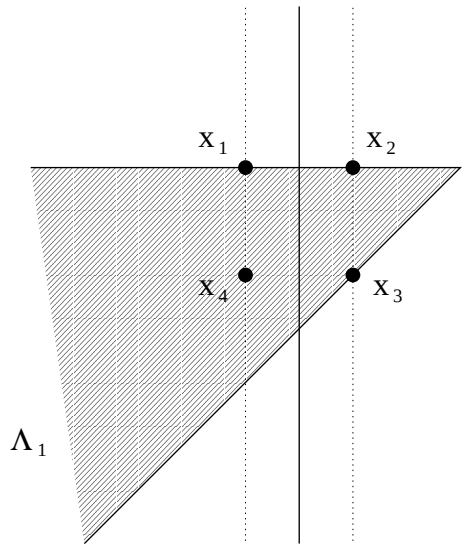


Figure 2.2 *In this example, we have* $\mathbf{x}_1 \in \Lambda_1$ *and therefore* $d(\mathbf{x}_1, \Lambda_1) = 0$. *The set* Λ_1 *has been computed using* $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$

It is not obvious that the set Λ_p is not empty. It is proven in the following lemma,

Lemma 2.2
Both in the separable and non-separable case, the set Λ_p is not empty. Moreover
bound on the span
$S_p = d(\mathbf{x}_p, \Lambda_p) \leq D_{SV}$

Proof is in Appendix

Remark : From lemma 2.2, we conclude (as in the separable case) that

$$S \leq D_{SV}, \tag{2.15}$$

where D_{SV} is the diameter of the smallest sphere containing the support vectors of the first category.

2.5 The Bounds

The generalization ability of SVMs can be explained by their capacity control. Indeed, the VC dimension of hyperplanes with margin ρ is less than $D^2/4\rho^2$, where D is the diameter of the smallest sphere containing the training points (Vapnik, 1995). This is the theoretical idea motivating the maximization of the margin.

This section presents new bounds on the generalization ability of SVMs. The major improvement lies in the fact that the bounds will depend on the span of the support vectors, which gives tighter bounds that ones depending on the diameter

of the training points.

Let us first introduce our fundamental result :

Lemma 2.3
If in the leave-one-out procedure a support vector $\mathbf{x}_p$ corresponding to $0 < \alpha_p < C$ is recognized incorrectly, then the inequality

$$\alpha_p^0 S_p \max(D, 1/\sqrt{C}) \geq 1$$

holds true.

See proof in Appendix

The previous lemma leads us to the following theorem for the separable case :

Theorem 2.1
Suppose that a SVM separates training data of size ℓ without error. Then the expectation of the probability of error $p_{error}^{\ell-1}$ for the SVM trained on the training data of size $\ell - 1$ has the bound

$$Ep_{error}^{\ell-1} \leq E\left(\frac{SD}{\ell\rho^2}\right),$$

where the values of span of support vectors S, diameter of the smallest sphere containing the training points D, and the margin ρ are considered for training sets of size ℓ.

Proof : Let us prove that the number of errors made by the leave-one-out procedure is bounded by $\frac{SD}{\rho^2}$. Taking the expectation and using lemma 2.1 will prove the theorem.

Consider a support vector $\mathbf{x}_p$ incorrectly classified by the leave-one-out procedure. Then lemma 2.3 gives $\alpha_p^0 S_p D \geq 1$ (we consider here the separable case and $C = \infty$) and

$$\alpha_p^0 \geq \frac{1}{SD}$$

holds true. Now let us sum the left and right hand sides of this inequality over all support vectors where the leave-one-out procedure commits an error

$$\frac{\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)}{SD} \leq \sum_* \alpha_i^0.$$

Here $\sum_*$ indicates that the sum is taken only over support vectors where the leave-one-out procedure makes an error. From this inequality we have

$$\frac{\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)}{\ell SD} \leq \frac{1}{\ell}\sum_{i=1}^{n} \alpha_i^0.$$

Therefore we have (using (2.8))

$$\frac{\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)}{\ell} \leq \frac{SD\sum_{i=1}^{n} \alpha_i^0}{\ell} = \frac{SD}{\ell\rho^2}.$$

Taking the expectation over both sides of the inequality and using the Luntz and Brailovsky Lemma we prove the theorem.

For the non-separable case the following theorem is true.

Theorem 2.2
The expectation of the probability of error $p_{error}^{\ell-1}$ for a SVM trained on the training data of size $\ell - 1$ has the bound

$$Ep_{error}^{\ell-1} \leq E\left(\frac{S\max(D, 1/\sqrt{C})\sum_{i=1}^{n^*}\alpha_i^0 + m}{\ell}\right),$$

where the sum is taken only over α_i corresponding to support vectors of the first category (for which $0 < \alpha_i < C$) and m is the number of support vectors of the second category (for which $\alpha_i = C$). The values of the span of support vectors S, diameter of the smallest sphere containing the training points D, and the Lagrange multipliers $\alpha^0 = (\alpha_1^0, ..., \alpha_n^0)$ are considered for training sets of size ℓ.

Proof : The proof of this theorem is similar to the proof of theorem 2.1. We consider all support vectors of the second category (corresponding to $\alpha_j = C$) as an error. For the first category of support vectors we estimate the number $\mathcal{L}^*(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)$ of errors in the leave-one-out procedure using the lemma 2.3 as in the proof of Theorem 2.1. We obtain

$$\begin{aligned}\frac{\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell)}{\ell} &\leq \frac{\mathcal{L}^*(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell) + m}{\ell} \\ &\leq \frac{S\max(D, 1/\sqrt{C})\sum^*\alpha_i + m}{\ell}\end{aligned}$$

Taking the expectation over both sides of the inequality and using the Luntz and Brailovsky Lemma we prove the theorem.

Note that in the case when $m = 0$ (separable case), the equality (2.8) holds true. In this case (provided that C is large enough) the bounds obtained in these two theorems coincide.

Note that in Theorems 2.1 and 2.2, it is possible using inequality (2.15) to bound the value of the span S by the diameter of the smallest sphere containing the support vectors D_{SV}. But, as pointed out by the experiments (see section 2.6), this would lead to looser bounds as the span can be much less than the diameter.

Extension

In the proof of lemma 2.3, it appears that the diameter of the training points D can be replaced by the span of the support vectors *after* the leave-one-out procedure. But since the set of support vectors after the leave-one-out procedure is unknown, we bounded this unknown span by D. Nevertheless this remark motivated us to analyze the case where the set of support vectors remains the same during the leave-one-out procedure.

In this situation, we are allowed to replace D by S in lemma 2.3 and more precisely, the following theorem is true.

Theorem 2.3
If the sets of support vectors of first and second categories remain the same during the leave-one-out procedure, then for any support vector $\mathbf{x}_p$, the following equality holds :

$$y_p(f^0(\mathbf{x}_p) - f^p(\mathbf{x}_p)) = \alpha_p^0 S_p^2$$

where f^0 and f^p are the decision function given by the SVM trained respectively on the whole training set and after the point $\mathbf{x}_p$ has been removed.

Proof is in Appendix

The assumption that the set of support vectors does not change during the leave-one-out procedure is not satisfied in most cases. Nevertheless, the proportion of points which violate this assumption is usually small compared to the number of support vectors. In this case Theorem 2.3 provides a good approximation of the result of the leave-one-out procedure, as pointed out by the experiments (see Section 2.6, figure 2.4).

Note that theorem 2.3 is stronger than lemma 2.3 for three reasons : the term $S_p \max(D, 1/\sqrt{C})$ becomes S_p^2, the inequality turns out to be an equality and the result is valid for any support vector.

The previous theorem enables us to compute the number of errors made by the leave-one-out procedure :

Corollary 2.1
Under the assumption of Theorem 2.3, the test error prediction given by the leave-one-out procedure is

span-rule

$$t_\ell = \frac{1}{\ell}\mathcal{L}(\mathbf{x}_1, y_1, ..., \mathbf{x}_\ell, y_\ell) = \frac{1}{\ell} Card\{p : \ \alpha_p^0 S_p^2 \geq y_p f^0(\mathbf{x}_p)\} \tag{2.16}$$

2.6 Experiments

The previous bounds on the generalization ability of Support Vector Machines involved the diameter of the smallest sphere enclosing the training points (Vapnik, 1995). We have shown (cf inequality (2.15)) that the span S is always smaller than this diameter, but to appreciate the gain, we conducted some experiments.

comparison span - diameter

First we compare the diameter of the smallest sphere enclosing the training points, the one enclosing the support vectors and the span of the support vectors using the postal database. This dataset consists of 7291 handwritten digits of size 16x16 with a test set of 2007 examples. Following Schölkopf et al. (1999), we split the training set in 23 subsets of 317 training examples. Our task is to separate digits 0 to 4 from 5 to 9. Error bars in figure 2.3 are standard deviations over the 23 trials. The diameters and the span in figure 2.3 are plotted for different values

of σ, the width of the RBF kernel we used :

$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}}$.

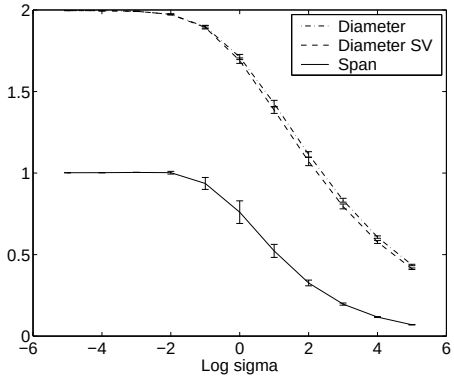


Figure 2.3 *Comparison of D, D_{SV} and S*

In this example, the span is up to 6 times smaller than the diameter.

Now we would like to use the span for predicting accurately the test error. This would enable us to perform efficient model selection, i.e. choosing the optimal values of parameters in SVMs (the width of the RBF kernel σ or the constant C, for instance).

Note that the span S is defined as a maximum $S = \max_p S_p$ and therefore taking into account the different values S_p should provide a more accurate estimation of the generalization error than the span S only. Therefore, we used the *span-rule* (2.16) in Corollary 1 to predict the test error.

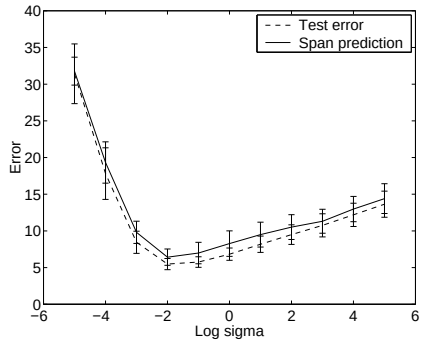
Our experiments have been carried out on two databases : a separable one, the postal database, described above and a noisy one, the breast-cancer database[1]. The latter has been split randomly 100 times into a training set containing 200 examples and a test set containing 77 examples.

model selection

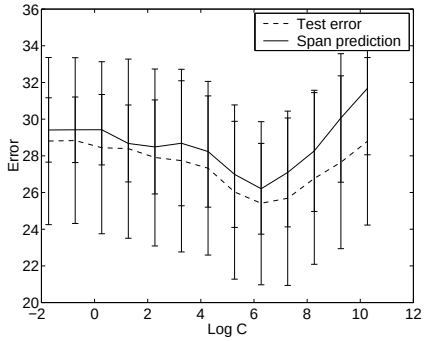
Figure 2.4a compares the test error and its prediction given by the span-rule (2.16) for different values of the width σ of the RBF kernel on the postal database. Figure 2.4b plots the same functions for different values of C on the breast-cancer database. The prediction is very accurate and the curves are almost identical.

The computation of the span-rule (2.16) involves computing the span S_p (2.13) for every support vector. Note, however, that we are interested in the inequality $S_p^2 \leq y_p f(\mathbf{x}_p)/\alpha_p^0$, rather than the exact value of the span S_p. Therefore, if while minimizing $S_p = d(\mathbf{x}_p, \Lambda_p)$ we find a point $\mathbf{x}^* \in \Lambda_p$ such that $d(\mathbf{x}_p, \mathbf{x}^*)^2 \leq$

1. Available from `http://horn.first.gmd.de/~raetsch/data/breast-cancer`



(a) choice of σ in the postal database



(b) choice of C in the breast-cancer database

Figure 2.4 *Test error and its prediction using the span-rule (2.16).*

$y_p f(\mathbf{x}_p)/\alpha_p^0$, we can stop the minimization because this point will be correctly classified by the leave-one-out procedure.

Figure 2.5 compares the time required to (a) train the SVM on the postal database, (b) compute the estimate of the leave-one-out procedure given by the span-rule (2.16) and (c) compute exactly the leave-one-out procedure. In order to have a fair comparison, we optimized the computation of the leave-one-out procedure in the following way : for every support vector $\mathbf{x}_p$, we take as starting point for the minimization (2.5) involved to compute f^p (the decision function after having removed the point $\mathbf{x}_p$), the solution given by f^0 on the whole training set. The reason is that f^0 and f^p are usually "close".

computation time

The results show that the time required to compute the span is not prohibitive and is very attractive compared to the leave-one-out procedure.

2.7 Conclusion

In this paper, we have shown that the generalization ability of support vector machines depends on a more complicated geometrical concept than the margin

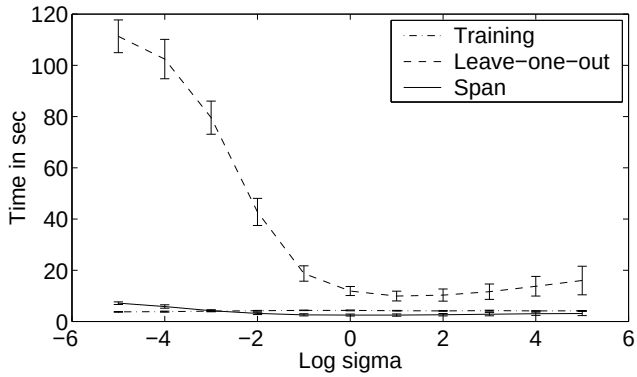


Figure 2.5 *Comparison of time required for SVM training, computation of span and leave-one-out on the postal database*

only. A direct application of the concept of span is the selection of the optimal parameters of the SVM since the span enables to get an accurate prediction of the test error.

Similary to chapter **??**, the concept of the span also lead to new learning algorithms involving the minimization of the number of errors made by the leave-one-out procedure.

Acknowledgments

The authors would like to thank Léon Bottou, Patrick Haffner and Yann Lecun and Sayan Mukherjee for very helpful discussions.

Appendix : proofs

Proof of lemma 2.2

We will prove this result for the non-separable case. The result is also valid for the separable case since it can be seen as a particular case of the non-separable one with C large enough.

Let us define Λ_p^+ as the subset of Λ_p with additional constraints $\lambda_i \geq 0$:

$$\Lambda_p^+ = \{ \sum_{i=1,\ i\neq p}^{n} \lambda_i \mathbf{x}_i \in \Lambda_p \ : \ \lambda_i \geq 0 \ i \neq p\}. \tag{2.17}$$

We shall prove that $\Lambda_p^+ \neq \emptyset$ by proving that a vector λ of the following form exists :

$$\lambda_j = 0, \quad j = n^* + 1, ..., n \tag{2.18}$$

$$\lambda_i = \mu \frac{C - \alpha_i^0}{\alpha_p^0}, \quad y_i = y_p, \quad i \neq p, \quad i = 1, ..., n^* \tag{2.19}$$

$$\lambda_i = \mu \frac{\alpha_i^0}{\alpha_p^0}, \quad y_i \neq y_p, \quad i = 1, ..., n^* \tag{2.20}$$

$$0 \leq \mu \leq 1 \tag{2.21}$$

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.22}$$

It is straightforward to check that if such a vector λ exists, then $\sum \lambda_i \mathbf{x}_i \in \Lambda_p^+$ and therefore $\Lambda_p^+ \neq \emptyset$. Since $\Lambda_p^+ \subset \Lambda_p$, we will have $\Lambda \neq \emptyset$.

Taking into account equations (2.19) and (2.20), we can rewrite constraint (2.22) as follows :

$$1 = \frac{\mu}{\alpha_p^0} \left(\sum_{\substack{i=1,\ i\neq p \\ y_i = y_p}}^{n^*} (C - \alpha_i^0) + \sum_{\substack{i=1 \\ y_i \neq y_p}}^{n^*} \alpha_i^0 \right) \tag{2.23}$$

We need to show that the value of μ given by equation (2.23) satisfies constraint (2.21).

For this purpose, let us define Δ as :

$$\Delta = \sum_{i/\ y_i = y_p}^{n^*} (C - \alpha_i^0) + \sum_{i/\ y_i \neq y_p}^{n^*} \alpha_i^0 \tag{2.24}$$

$$= -y_p \sum_{i=1}^{n^*} y_i \alpha_i^0 + \sum_{i/\ y_i = y_p}^{n^*} C \tag{2.25}$$

Now, note that

$$\sum_{i=1}^{n} y_i \alpha_i^0 = \sum_{i=1}^{n^*} y_i \alpha_i^0 + C \sum_{i=n^*+1}^{n} y_i = 0. \tag{2.26}$$

Combining equations (2.25) and (2.26) we get

$$\begin{aligned}\Delta &= C y_p \sum_{i=n^*+1}^{n} y_i + \sum_{i/\ y_i=y_p}^{n^*} C \\ &= Ck,\end{aligned}$$

where k is an integer.

Since equation (2.24) gives $\Delta > 0$, we have finally

$$\Delta \geq C. \tag{2.27}$$

Let us rewrite equation (2.23) as :

$$1 = \frac{\mu}{\alpha_p^0}(\Delta - (C - \alpha_p^0)).$$

We obtain

$$\mu = \frac{\alpha_p^0}{\Delta - (C - \alpha_p^0)} \tag{2.28}$$

Taking into account inequality (2.27), we finally get $0 \leq \mu \leq 1$. Thus, constraint (2.21) is fulfilled and Λ_p^+ is not empty.

Now note that the set Λ_p^+ is included in the convex hull of $\{\mathbf{x}_i\}_{i \neq p}$ and since $\Lambda_p^+ \neq \emptyset$, we obtain

$$d(\mathbf{x}_p, \Lambda_p^+) \leq D_{SV},$$

where D_{SV} is the diameter of the smallest ball containing the support vectors of the first category.

Since $\Lambda_p^+ \subset \Lambda_p$ we finally get

$$S_p = d(\mathbf{x}_p, \Lambda_p) \leq d(\mathbf{x}_p, \Lambda_p^+) \leq D_{SV}.$$

Proof of lemma 2.3

Let us first consider the separable case.

Suppose that our training set $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ is ordered such that the support vectors are the first n training points. The non-zero Lagrange multipliers associated with these support vectors are

$$\alpha_1^0, ..., \alpha_n^0 \tag{2.29}$$

In other words, the vector $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_n^0, 0, \ldots, 0)$ maximizes the functional

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.30}$$

subject to the constraints

$$\boldsymbol{\alpha} \geq 0, \tag{2.31}$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0. \tag{2.32}$$

Let us consider the result of the leave-one-out procedure on the support vector $\mathbf{x}_p$. This means that we maximized functional (2.30) subject to the constraints (2.31), (2.32) and the additional constraint

$$\alpha_p = 0, \tag{2.33}$$

and obtained the solution

$\boldsymbol{\alpha}^p = (\alpha_1^p, ..., \alpha_\ell^p)$.

Using this solution we construct the separating hyperplane

$\mathbf{w}_p \cdot \mathbf{x} + b_p = 0,$

where

$\mathbf{w}_p = \sum_{i=1}^{\ell} \alpha_i^p y_i \mathbf{x}_i.$

We would like to prove that if this hyperplane classifies the vector $\mathbf{x}_p$ incorrectly:

$$y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p) < 0 \tag{2.34}$$

then

$\alpha_p^0 \geq \frac{1}{S_p D}.$

Since $\boldsymbol{\alpha}^p$ maximizes (2.30) under constraints (2.31), (2.32) and (2.33), the following inequality holds true

$$W(\boldsymbol{\alpha}^p) \geq W(\boldsymbol{\alpha}^0 - \boldsymbol{\delta}), \tag{2.35}$$

where the vector $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$ satisfies the following conditions

$\delta_p = \alpha_p^0,$

$\alpha_0 - \delta \geq 0,$

$\sum_{i=1}^{n} \delta_i y_i = 0.$

$$\delta_i = 0, \quad i > n \tag{2.36}$$

From inequality (2.35) we obtain

$$W(\boldsymbol{\alpha}^0) - W(\boldsymbol{\alpha}^p) \leq W(\boldsymbol{\alpha}^0) - W(\boldsymbol{\alpha}^0 - \boldsymbol{\delta}). \tag{2.37}$$

Since $\boldsymbol{\alpha}^0$ maximizes (2.30) under the constraints (2.31) and (2.32), the following inequality holds true

$$W(\boldsymbol{\alpha}^0) \geq W(\boldsymbol{\alpha}^p + \boldsymbol{\gamma}), \tag{2.38}$$

where $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_\ell)$ is a vector satisfying the constraints

$$\boldsymbol{\alpha}_p + \boldsymbol{\gamma} \geq 0,$$

$$\sum_{i=1}^{\ell} \gamma_i y_i = 0.$$

$$\alpha_i^p = 0 \Longrightarrow \gamma_i = 0, \quad i \neq p \tag{2.39}$$

From (2.37) and (2.38), we have

$$W(\boldsymbol{\alpha}^p + \boldsymbol{\gamma}) - W(\boldsymbol{\alpha}^p) \leq W(\boldsymbol{\alpha}_0) - W(\boldsymbol{\alpha}_0 - \boldsymbol{\delta}) \tag{2.40}$$

Let us calculate both the left hand side,I_1, and the right hand side, I_2 of inequality (2.40).

$$\begin{aligned}
I_1 &= W(\boldsymbol{\alpha}^p + \boldsymbol{\gamma}) - W(\boldsymbol{\alpha}^p) \\
&= \sum_{i=1}^{\ell} (\alpha_i^p + \gamma_i) - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^p + \gamma_i)(\alpha_j^p + \gamma_j) y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\
&\quad - \sum_{i=1}^{\ell} \alpha_i^p + \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i^p \alpha_j^p y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\
&= \sum_{i=1}^{\ell} \gamma_i - \sum_{i,j}^{\ell} \gamma_i \alpha_j^p y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \frac{1}{2} \sum_{i,j}^{\ell} y_i y_j \gamma_i \gamma_j \mathbf{x}_i \cdot \mathbf{x}_j \\
&= \sum_{i=1}^{\ell} \gamma_i (1 - y_i \mathbf{w}_p \cdot \mathbf{x}_i) - \frac{1}{2} \sum_{i,j}^{\ell} y_i y_j \gamma_i \gamma_j (\mathbf{x}_i, \mathbf{x}_j)
\end{aligned}$$

Taking into account that

$$\sum_{i=1}^{\ell} \gamma_i y_i = 0$$

we can rewrite expression

$$I_1 = \sum_{i \neq p}^{\ell} \gamma_i [1 - y_i (\mathbf{w}_p \cdot \mathbf{x}_i + b_p)] + \gamma_p [1 - y_p (\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] - \frac{1}{2} \sum_{i,j}^{n} y_i y_j \gamma_i \gamma_j \mathbf{x}_i \cdot \mathbf{x}_j.$$

Since for $i \neq p$ the condition (2.39) means that either $\gamma_i = 0$ or $\mathbf{x}_i$ is a support

vector of the hyperplane $\mathbf{w}_p$, the following equality holds

$\gamma_i[y_i(\mathbf{w}_p \cdot \mathbf{x}_i + b_p) - 1] = 0.$

We obtain

$$I_1 = \gamma_p[1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] - \frac{1}{2}\sum_{i,j}^{n} y_i y_j \gamma_i \gamma_j \mathbf{x}_i \cdot \mathbf{x}_j.$$

Now let us define vector γ as follows:

$\gamma_p = \gamma_k = a,$

$\gamma_i = 0 \quad i \notin \{k, p\},$

where a is some constant and k such that $y_p \neq y_k$ and $\alpha_k^p > 0$. For this vector we obtain

$$\begin{aligned} I_1 &= a[1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] - \frac{a^2}{2}||\mathbf{x}_p - \mathbf{x}_k||^2 \\ &\geq a[1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] - \frac{a^2}{2}D^2. \end{aligned} \tag{2.41}$$

Let us choose the value a to maximize this expression

$$a = \frac{1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)}{D^2}.$$

Putting this expression back into (2.41) we obtain

$$I_1 \geq \frac{1}{2}\frac{(1 - y_p[(\mathbf{x}_p, \mathbf{w}_p) + b_p])^2}{D^2}.$$

Since, according to our assumption, the leave-one-out procedure commits an error at the point $\mathbf{x}_p$ (that is, the inequality (2.34) is valid), we obtain

$$I_1 \geq \frac{1}{2D^2}. \tag{2.42}$$

Now we estimate the right hand side of inequality (2.40)

$$I_2 = W(\boldsymbol{\alpha}^0) - W(\boldsymbol{\alpha}^0 - \boldsymbol{\delta}).$$

We choose $\delta_i = -y_i y_p \alpha_p^0 \lambda_i$, where λ is the vector that defines the value of $d(\mathbf{x}_p, \Lambda_p)$ in equation (2.13).

We have

$$\begin{aligned} I_2 &= W(\boldsymbol{\alpha}^0) - W(\boldsymbol{\alpha}^0 - \boldsymbol{\delta}) \\ &= \sum_{i=1}^{n} \alpha_i^0 - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i^0 \alpha_j^0 y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_{i=1}^{n}(\alpha_i^0 + y_i y_p \alpha_p^0 \lambda_i) \\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^0 + y_i y_p \alpha_p^0 \lambda_i)(\alpha_j^0 + y_j y_p \alpha_p^0 \lambda_j) y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \end{aligned}$$

$$= -y_p\alpha_p^0 \sum_{i=1}^{n} y_i\lambda_i + y_p\alpha_p^0 \sum_{i,j=1}^{n} \alpha_i^0\lambda_j y_i \mathbf{x}_i \cdot \mathbf{x}_j + \frac{1}{2}(\alpha_p^0)^2(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i)^2. \tag{2.43}$$

Since

$$\sum_{i=1}^{n} \lambda_i = 0,$$

and $\mathbf{x}_i$ is a support vector, we have

$$I_2 = y_p\alpha_p^0 \sum_{i=1}^{n} \lambda_i y_i[y_i(\mathbf{w}_0 \cdot \mathbf{x}_i + b_0) - 1] + \frac{(\alpha_p^0)^2}{2}(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i)^2 = \frac{(\alpha_p^0)^2}{2} S_p^2. \tag{2.44}$$

Combining (2.40), (2.42) and (2.44) we obtain

$$\alpha_p^0 S_p D \geq 1.$$

Consider now the non-separable case. The sketch of the proof is the same. There are only two differences :

First, the vector $\boldsymbol{\gamma}$ needs to satisfy

$$\boldsymbol{\alpha}_p + \boldsymbol{\gamma} \leq C.$$

A very similar proof to the one of lemma 2.2 gives us the existence of $\boldsymbol{\gamma}$.

The other difference lies in the choice of a in equation (2.41). The value of a which maximizes equation (2.41) is

$$a^* = \frac{1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)}{D^2}.$$

But we need to fulfill the condition $a \leq C$. Thus, if $a^* > C$, we replace a by C in equation (2.41) and we get :

$$\begin{aligned} I_1 &\geq C[1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] - \frac{C^2}{2} D^2 \\ &= CD^2(a^* - \frac{C}{2}) \\ &\geq CD^2 \frac{a^*}{2} \\ &= \frac{C}{2}[1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] \\ &\geq \frac{C}{2} \end{aligned}$$

The last inequality comes from (2.34).

Finally, we have

$$I_1 \geq \frac{1}{2} \min\left(C, \frac{1}{D^2}\right).$$

By combining this last inequality with (2.40) and (2.44) we prove the lemma.

Proof of theorem 2.3

The proof follows the proof of lemma 2.3. Under the assumption that the set of support vectors remain the same during the leave-one-out procedure, we can take

$$\boldsymbol{\delta} = \boldsymbol{\gamma} = \boldsymbol{\alpha}^0 - \boldsymbol{\alpha}^p$$

as $\boldsymbol{\alpha}^0 - \boldsymbol{\alpha}^P$ is a vector satisfying simultaneously the set of constraints (2.36) and (2.39).

Then inequality (2.40) becomes an equality :

$$I_1 = W(\boldsymbol{\alpha}^0) - W(\boldsymbol{\alpha}^p) = I_2 \tag{2.45}$$

From inequality (2.37), it follows that

$$I_2 \leq I_2^* = W(\boldsymbol{\alpha}^0) - W(\boldsymbol{\alpha}^0 - \boldsymbol{\delta}^*), \tag{2.46}$$

where $\delta_i^* = -y_i y_p \alpha_p^0 \lambda_i$ and λ is given by the definition of the span S_p (cf equation (2.13)).

The computation of I_2 and I_2^* is similar to the one involved in the proof of lemma 2.3 (cf equation (2.44))

$$I_2^* = \frac{(\alpha_p^0)^2}{2} S_p^2 - \alpha_p^0 [y_p(\mathbf{w}_0 \cdot \mathbf{x}_p + b_0) - 1]$$

$$I_2 = \frac{(\alpha_p^0)^2}{2} (\sum_i \lambda_i^* \mathbf{x}_i)^2 - \alpha_p^0 [y_p(\mathbf{w}_0 \cdot \mathbf{x}_p + b_0) - 1],$$

where

$$\lambda_i^* = y_i \frac{\alpha_i^p - \alpha_i^0}{\alpha_p^0}$$

From (2.46), we get $(\sum_i \lambda_i^* \mathbf{x}_i)^2 \leq S_p^2$.

Now note that $\sum_{i \neq p} \lambda_i^* \mathbf{x}_i \in \Lambda_p$ and by definition of S_p, $(\sum_i \lambda_i^* \mathbf{x}_i)^2 \geq S_p^2$. Finally, we have

$$(\sum_i \lambda_i^* \mathbf{x}_i)^2 = S_p^2. \tag{2.47}$$

The computation of I_1 gives (cf equation (2.41))

$$I_1 = \alpha_p^0 [1 - y_p(\mathbf{w}_p \cdot \mathbf{x}_p + b_p)] - \frac{(\alpha_p^0)^2}{2} (\sum_i \lambda_i^* \mathbf{x}_i)^2$$

Putting the values of I_1 and I_2 back in equation (2.45), we get

$$(\alpha_p^0)^2 (\sum_i \lambda_i^* \mathbf{x}_i)^2 = \alpha_p^0 y_p [f^0(\mathbf{x}_p) - f^p(\mathbf{x}_p)]$$

and the theorem is proven by dividing by α_p^0 and taking into account equation (2.47) :

$$\alpha_p^0 S_p^2 = y_p [f^0(\mathbf{x}_p) - f^p(\mathbf{x}_p)]$$

References

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.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition (in russian). *Technicheskaya Kibernetica*, 3, 1969.

B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Generalization bounds via eigenvalues of the Gram matrix. Submitted to COLT99, 1999.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 1998. To appear. Also: NeuroCOLT Technical Report NC-TR-96-053, 1996, ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports.

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

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.