Radius-margin Bound on the Leave-one-out Error of Multi-class SVMs

Yannick Darcy, Yann Guermeur

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Radius-margin Bound on the Leave-one-out Error of Multi-class SVMs

Yannick Darcy — Yann Guermeur

N° 5780

December 13, 2005

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--5780--FR+ENG

Radius-margin Bound on the Leave-one-out Error of Multi-class SVMs

Yannick Darcy* , Yann Guermeur†

Thème BIO — Systèmes biologiques
Projet MODBIO

Rapport de recherche n° 5780 — December 13, 2005 — 27 pages

Abstract: Using a support vector machine requires to set two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, one can use various procedures based on cross-validation. Obviously, the major drawback of such procedures rests in their time requirements. To overcome this difficulty, several upper bounds on the leave-one-out error of pattern recognition support vector machines have been derived. In this report, we demonstrate a direct extension of one of these bounds, called the radius-margin bound, to the case of the standard multi-class SVM.

Key-words: Leave-one-out error, SVMs, M-SVMs, model selection.

* UMR 7503-UHP
† UMR 7503-CNRS

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes

Résumé : L'utilisation d'une machine à vecteurs support requiert la détermination des valeurs de deux types d'hyper-paramètres : le paramètre de "marge molle" C et les paramètres du noyau. Pour effectuer cette tâche de sélection de modèle, différentes procédures fondées sur la validation croisée sont actuellement disponibles. Leur défaut premier réside dans le temps de calcul qu'elles nécessitent. Afin de surmonter cette difficulté, différentes bornes supérieures sur l'erreur "leave-one-out" des SVM calculant des dichotomies ont été proposées. Dans ce rapport, nous établissons une extension directe de l'une de ces bornes, nommée borne "rayon-marge", au cas de la SVM multi-classe standard.

Mots-clés : Erreur "leave-one-out", SVM, M-SVM, sélection de modèle.

1 Introduction

Using a support vector machine requires to set two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, several approaches are available ([7, 8]). A simple solution consists in applying a cross-validation procedure. This, however, can appear highly time consuming, especially if the procedure is a leave-one-out one. This is the reason why, in recent years, a number of upper bounds on the leave-one-out error of bi-class SVMs have been proposed in literature (see for instance [11, 12, 9]). Unfortunately, so far, little has been done to extend them to the multi-class case. A notable exception is the work Wang and co-workers [13]. The authors propose two extensions of Chapelle's "radius-margin bound" [10, 11, 3]. However, to preserve the appealing properties of the original (bi-class) result, they deviate significantly from a direct extension. As they put it themselves: "However, a direct extension of this bound to the multi-class scenario is not viable because it is rooted in the theoretical results of bi-class SVMs."

In this report, we introduce a direct extension of the radius-margin bound to the multi-class case. This leads us to introduce original concepts, such as the one of margins for multi-class SVMs. It appears once more that the statistical theory of multi-category discriminant analysis cannot be built as a trivial generalization of the theory developed for the computation of dichotomies. Our bound, when applied in the bi-class case, should be sharper than the original one, although this is obtained at a significant additional cost.

The organization of this document is as follows. Section 2 introduces the bi-class SVMs, summarizing their architecture and training algorithm. A detailed proof of the original bound is then presented in Section 3. In Section 4, the standard multi-class SVM is introduced. The extension of the RM Theorem to this machine is the subject of Section 5.

2 Bi-class SVMs

2.1 Architecture and algorithm

We first summarize the main points of the training algorithm of pattern recognition SVMs ([1, 4]). A SVM is characterized by a kernel $\kappa \rightarrow \mathcal{X}^2$ (one projection operator of which we call Φ) and a soft margin constant C. Suppose that it is trained on a subset $s_m = \{(x_i, y_i)\}$, $(1 \leq i \leq m)$ of $\mathcal{X} \times \{-1, 1\}$. The algorithm constructs an *optimal hyperplane*, that is to say an hyperplane maximizing the "margin". The "margin" is the smallest Euclidean distance of a point of the training set to the hyperplane, the equation of which is $\langle w^0, \Phi(x)\rangle + b^0 = 0$. Moreover, this optimal hyperplane is expressed in its canonical form ([10], p.412) i.e.

$$\min_x \|\langle w, \Phi(x)\rangle + b\| = 1.$$

The hyperplane is thus the solution of the following Quadratic Programming (QP) problem[1]:

Problem 1

$$\min_{w,b} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \right\}$$

$$s.t. \begin{cases} y_i(\langle w, \Phi(x_i)\rangle + b) \geq 1 - \xi_i, & (1 \leq i \leq m) \\ \xi_i \geq 0, & (1 \leq i \leq m) \end{cases} .$$

The couple (w^0, b^0) is the solution of this optimization problem. The Lagrangian function associated with Problem 1 is given by:

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \left(y_i(\langle w, \Phi(x_i)\rangle + b) - 1 + \xi_i\right) - \sum_{i=1}^{m} \beta_i \xi_i. \quad (1)$$

Setting the gradient of (1) with respect to w equal to the null vector gives the expression of w as a function of the Lagrange multipliers α_i which is:

$$w = \sum_{i=1}^{m} y_i \alpha_i \Phi(x_i). \quad (2)$$

Setting the gradient of (1) with respect to b equal to the null vector gives

$$\sum_{i=1}^{m} y_i \alpha_i = 0. \quad (3)$$

[1] The notions of mathematical programming used in this report can be found, for example, in [5].

Setting the gradient of (1) with respect to the slack variables ξ_i equal to the null vector gives

$$\alpha_i + \beta_i = C, \quad (1 \leq i \leq m). \tag{4}$$

Substituting (2) and using (3) and (4) in (1) leads to the Wolfe-dual formulation of Problem 1, the constraints of which are deduced from (3) and (4):

Problem 2

$$\max_{\alpha} \left\{ -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) + \sum_{i=1}^{m} \alpha_i \right\}$$

$$s.t. \begin{cases} 0 \leq \alpha_i \leq C, \quad (1 \leq i \leq m) \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases} .$$

The objective function of this latter problem can be reformulated as

$$J(\alpha) = -\frac{1}{2} \alpha^T H \alpha + 1^T \alpha$$

where the general term of the Hessian matrix H is given by $h_{i,j} = y_i y_j \kappa(x_i, x_j)$. Let $\alpha^0 = \left[\alpha_1^0, \ldots, \alpha_m^0\right]^T$ be the optimal solution of Problem 2. According to (2), the expression of w^0 is:

$$w^0 = \sum_{i=1}^{m} y_i \alpha_i^0 \Phi(x_i). \tag{5}$$

2.2 Some useful results

We introduce some results needed in the proofs of the key lemma and RM Theorem. First, note that the general term of vector $1 - H\alpha^0$ is simply

$$1 - \sum_{j=1}^{m} \alpha_j^0 y_i y_j \kappa(x_j, x_i) = 1 - y_i \langle w^0, \Phi(x_i) \rangle = y_i b^0 + \xi_i^0. \tag{6}$$

Proposition 1 *In the separable cas, we have:*

$$\alpha^{0T} H \alpha^0 = \|w^0\|^2 = \sum_{i=1}^{m} \alpha_i^0 = \frac{1}{\gamma^2} \tag{7}$$

with γ the margin.

Proof

- ${\alpha^0}^T H \alpha^0 = \|w^0\|^2$
  Actually, for any vector z of size m the components of which are z_i, we have:

$$z^T H z = \sum_{i=1}^{m} \sum_{j=1}^{m} z_i h_{ij} z_j = \sum_{i=1}^{m} \sum_{j=1}^{m} z_i z_j y_i y_j \kappa(x_i, x_j). \tag{8}$$

  Since $\kappa(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$,

$$z^T H z = \left\| \sum_{i=1}^{m} z_i y_i \Phi(x_i) \right\|^2. \tag{9}$$

  Hence, when z is an optimal solution of Problem 2, this directly leads to the expression of $\|w^0\|^2$.

- $\|w^0\|^2 = \sum_{i=1}^{m} \alpha_i^0$
  Since $\forall i \;\; \xi_i = 0$, the Kuhn-Tucker conditions imply that

$$\alpha_i^0 \left(1 - y_i(\langle w^0, \Phi(x_i) \rangle + b^0)\right) = 0.$$

  Thus

$$\sum_{i=1}^{m} \alpha_i^0 \left(1 - y_i(\langle w^0, \Phi(x_i) \rangle + b^0)\right) = 0.$$

  This is equivalent to:

$$\sum_{i=1}^{m} \alpha_i^0 - b^0 \sum_{i=1}^{m} \alpha_i^0 y_i - \langle w^0, \sum_{i=1}^{m} \alpha_i^0 y_i \Phi(x_i) \rangle = 0.$$

  Simplifying with the use of (3) and (5), we get:

$$\sum_{i=1}^{m} \alpha_i^0 - \|w^0\|^2 = 0.$$

- $\|w^0\|^2 = \frac{1}{\gamma^2}$
  This springs directly from the fact that the hyperplane is expressed in its canonical form.

■

3 Bi-class radius-margin bound

3.1 Key lemma in the bi-class case

In this section we introduce the key lemma which is a basis for the proof of the RM Theorem.

Lemma 1 *Let us consider a pattern recognition SVM on a domain $\mathcal{X}$. This SVM is characterized by a kernel κ (a projection operator Φ) and a soft margin constant C. Suppose that it is trained on a subset $s_m = \{(x_i, y_i)\}$, $(1 \leq i \leq m)$ of $\mathcal{X} \times \{-1, 1\}$, the points of which it separates without error. Suppose further that all the corresponding dual variables α_i^0 satisfy $\alpha_i^0 < C$. Consider now the same machine, trained on $s_m \setminus \{(x_p, y_p)\}$. If it makes an error on (x_p, y_p), then the inequality*

$$\alpha_p^0 \geq \frac{1}{\mathcal{D}_m^2} \tag{10}$$

holds, where $\mathcal{D}_m$ is the diameter of the smallest sphere containing the projections of the support vectors of the initial machine in the feature space.

Proof First, note that support vectors are points x_i verifying:

$$0 < \alpha_i^0 \tag{11}$$

which, given the hypotheses, implies that $0 < \alpha_i^0 < C$.
The hypotheses of Lemma 1 also implicitly imply the fact that $\alpha_p^0 \neq 0$, of which we will make use several times. Let (w^p, b^p) be the couple characterizing the optimal hyperplane when the machine is trained on $s_m \setminus \{(x_p, y_p)\}$. Assuming that this machine makes an error on the example (x_p, y_p) means that

$$y_i(\langle w^p, \Phi(x_i)\rangle + b^p) < 0.$$

Let $\alpha^p = [\alpha_1^p, \ldots, \alpha_{p-1}^p, 0, \alpha_{p+1}^p \ldots, \alpha_m^p]^T$ be the vector of $[0, C)^m$ the components of which are the dual variables of the second SVM, with a 0 in position p. This representation is used to characterize directly the second SVM with respect to the first one. Indeed, α^p is an optimal solution of Problem 2 under the additional constraint $\alpha_p^p = 0$. Let us define two more vectors in $\mathbb{R}_+^m$, $\lambda^p = [\lambda_i^p]$ and $\mu^p = [\mu_i^p]$. λ^p is constructed in such a way that the vector $\alpha^0 - \alpha_p^0 \lambda^p$ is a feasible solution of Problem 2 under the additional constraint that $\alpha_p^0 - \alpha_p^0 \lambda_p^p = 0$, i.e. $\alpha^0 - \alpha_p^0 \lambda^p$ satisfies the same constraints as α^p. We have thus:

$$\forall i \neq p, \;\; \alpha_i^0 - \alpha_p^0 \lambda_i^p \geq 0 \Longleftrightarrow \lambda_i^p \leq \frac{\alpha_i^0}{\alpha_p^0}$$

$$\forall i \neq p, \;\; \lambda_i^p \geq 0 \Longrightarrow \alpha_i^0 - \alpha_p^0 \lambda_i^p \leq C$$

$$\alpha_p^0 - \alpha_p^0 \lambda_p^p = 0 \Longleftrightarrow \lambda_p^p = 1.$$

Using (3), we have:

$$\sum_{i=1}^{m} y_i \left(\alpha_i^0 - \alpha_p^0 \lambda_i^p\right) = 0 \Longleftrightarrow \alpha_p^0 \sum_{i=1}^{m} y_i \lambda_i^p = 0 \Longleftrightarrow \sum_{i=1}^{m} y_i \lambda_i^p = 0.$$

Vector λ^p satisfies the following properties:

$$\begin{cases} \lambda_p^p = 1 \\ \forall i, \;\; 0 \leq \lambda_i^p \leq \frac{\alpha_i^0}{\alpha_p^0} \\ \sum_{i=1}^{m} y_i \lambda_i^p = 0 \end{cases} . \tag{12}$$

In addition, we assign nonzero values only to some coordinates of λ^p (apart from λ_p^p) that correspond to points the class of which is the opposite of the class of x_p. Note that for any point x_i that is not a support vector, we have $\alpha_i^0 = 0$ and consequently $\lambda_i^p = 0$. Our set of points for which $\lambda_i^p \neq 0$ is thus a subset of the set of support vectors. To sum up, vector λ^p satisfies the following properties:

$$\begin{cases} \begin{cases} y_i = y_p \begin{cases} i = p, \;\; \lambda_i^p = 1 \\ i \neq p, \;\; \lambda_i^p = 0 \end{cases} \\ y_i \neq y_p, \;\; 0 \leq \lambda_i^p \leq \frac{\alpha_i^0}{\alpha_p^0} \end{cases} \\ \sum_{i=1}^{m} y_i \lambda_i^p = 0 \end{cases} . \tag{13}$$

Quite similarly to λ^p, μ^p is specified so that $\alpha^p + c\mu^p$ satisfies the same constraints as α^0, where c is a scalar belonging to $(0, C]$. Furthermore, we impose that $\mu_p^p = 1$. First, note that, as required,

$$\forall i \neq p, \;\; \mu_i^p \geq 0 \Longrightarrow \alpha_i^p + c\mu_i^p \geq 0$$

and

$$\alpha_p^p + c\mu_p^p = c \Longrightarrow 0 \leq \alpha_p^p + c\mu_p^p \leq C.$$

The remaining constraints are expressed as:

$$\sum_{i=1}^{m} y_i \left(\alpha_i^p + c\mu_i^p\right) = 0 \Longleftrightarrow \sum_{i=1}^{m} y_i \mu_i^p = 0,$$

$$\forall i \neq p, \;\; \alpha_i^p + c\mu_i^p \leq C \Longleftrightarrow \mu_i^p \leq \frac{C - \alpha_i^p}{c}.$$

Note that since all the α_i^0 are strictly inferior to C, changing C for a larger value C' does not change the optimal solution of Problem 2. As a consequence, without loss of generality we can make the hypothesis that C is arbitrarily large. The last constraint thus vanishes. Performing as in the case of λ^p, we assign nonzero values only to some coordinates of μ^p (apart from μ_p^p) that correspond to support vectors of (w^0, b^0) the class of which is the

opposite of the class of x_p. The reason for this choice will appear at the end of the proof. Finally, vector μ^p satisfies the following properties:

$$\begin{cases} \begin{cases} y_i = y_p \begin{cases} i = p, \;\; \mu_i^p = 1 \\ i \neq p, \;\; \mu_i^p = 0 \end{cases} \\ y_i \neq y_p \begin{cases} \alpha_i^0 = 0, \;\; \mu_i^p = 0 \\ \alpha_i^0 > 0, \;\; \mu_i^p \geq 0 \end{cases} \end{cases} \\ \sum_{i=1}^m y_i \mu_i^p = 0 \end{cases} . \tag{14}$$

By construction of vectors λ^p and μ^p, we have $J(\alpha^0 - \alpha_p^0 \lambda^p) \leq J(\alpha^p)$ and $J(\alpha^p + c\mu^p) \leq J(\alpha^0)$, and by way of consequence,

$$J(\alpha^0) - J(\alpha^0 - \alpha_p^0 \lambda^p) \geq J(\alpha^0) - J(\alpha^p) \geq J(\alpha^p + c\mu^p) - J(\alpha^p). \tag{15}$$

Since

$$J(\alpha^0 - \alpha_p^0 \lambda^p) = -\frac{1}{2}(\alpha^0 - \alpha_p^0 \lambda^p)^T H (\alpha^0 - \alpha_p^0 \lambda^p) + 1^T (\alpha^0 - \alpha_p^0 \lambda^p),$$

the expression of the first term is

$$J(\alpha^0) - J(\alpha^0 - \alpha_p^0 \lambda^p) = \frac{\left(\alpha_p^0\right)^2}{2} {\lambda^p}^T H \lambda^p + \alpha_p^0 {\lambda^p}^T \left(1 - H\alpha^0\right).$$

But recalling (6) and (13) this can be simplified:

$${\lambda^p}^T \left(1 - H\alpha^0\right) = b^0 \sum_{i=1}^m y_i \lambda_i^p + \sum_{i=1}^m \lambda_i^p \xi_i^0 = \sum_{i=1}^m \lambda_i^p \xi_i^0. \tag{16}$$

At this point, we make use of the hypothesis that all the dual variables α_i^0 are strictly inferior to C. According to the Kuhn-Tucker conditions, this implies that all the slack variables ξ_i^0 are equal to 0. Thus, the expression of $J(\alpha^0) - J(\alpha^0 - \alpha_p^0 \lambda^p)$ is simply (recalling (9))

$$J(\alpha^0) - J(\alpha^0 - \alpha_p^0 \lambda^p) = \frac{\left(\alpha_p^0\right)^2}{2} {\lambda^p}^T H \lambda^p = \frac{1}{2} (\alpha_p^0)^2 \left\| \sum_{i=1}^m \lambda_i^p y_i \Phi(x_i) \right\|^2. \tag{17}$$

We now turn to the right-hand side of (15), directly simplifying it thanks to (9).

$$J(\alpha^p + c\mu^p) - J(\alpha^p) = c{\mu^p}^T \left(1 - H\alpha^p\right) - \frac{c^2}{2} \left\| \sum_{i=1}^m \mu_i^p y_i \Phi(x_i) \right\|^2.$$

Proceeding as in the case of the left-hand side of (15), we establish that

$${\mu^p}^T \left(1 - H\alpha^p\right) = b_p \sum_{i=1}^m y_i \mu_i^p + \sum_{i=1}^m \mu_i^p \xi_i^p = \sum_{i=1}^m \mu_i^p \xi_i^p.$$

If the initial SVM, characterized by the couple (w^0, b^0), has nothing but null slack variables, then the sole slack variable of the SVM (w^p, b^p) which can be positive is ξ_p^p. Thus as in (16),

$$\mu^{pT}\left(1 - H\alpha^p\right) = \xi_p^p = 1 - y_p\left(\langle w^p, \Phi(x_p)\rangle + b^p\right)$$

and by way of consequence

$$J(\alpha^p + c\mu^p) - J(\alpha^p) = c\left[1 - y_p\left(\langle w^p, \Phi(x_p)\rangle + b^p\right)\right] - \frac{c^2}{2}\left\|\sum_{i=1}^{m} \mu_i^p y_i \Phi(x_i)\right\|^2. \qquad (18)$$

Combining (17) and (18) finally gives:

$$\frac{1}{2}(\alpha_p^0)^2\left\|\sum_{i=1}^{m} \lambda_i^p y_i \Phi(x_i)\right\|^2 \geq c\left[1 - y_p\left(\langle w^p, \Phi(x_p)\rangle + b^p\right)\right] - \frac{c^2}{2}\left\|\sum_{i=1}^{m} \mu_i^p y_i \Phi(x_i)\right\|^2. \qquad (19)$$

We want to find the best lower bound on the left-hand side of this inequality (or more precisely α_p^0), thus we look for the maximum of the right-hand side with respect to c. On $\mathbb{R}_+^*$ this maximum is obtain for:

$$c^* = \frac{1 - y_p\left(\langle w_p, \Phi(x_p)\rangle + b_p\right)}{\left\|\sum_{i=1}^{m} \mu_i^p y_i \Phi(x_i)\right\|^2}.$$

To understand why we maximize on $\mathbb{R}_+^*$ rather than on $(0, C]$, remember that without loss of generality we can make the hypothesis that C is arbitrarily large, which exonerates us from checking that $c^* < C$. By substitution in (19), $(\alpha_p^0)^2$ can thus be bounded from below as follows:

$$(\alpha_p^0)^2 \geq \frac{\left[1 - y_p\left(\langle w_p, \Phi(x_p)\rangle + b_p\right)\right]^2}{\left\|\sum_{i=1}^{m} \lambda_i^p y_i \Phi(x_i)\right\|^2 \left\|\sum_{i=1}^{m} \mu_i^p y_i \Phi(x_i)\right\|^2}.$$

Since by hypothesis the SVM (w^p, b^p) misclassifies x_p, $\left[1 - y_p\left(\langle w^p, \Phi(x_p)\rangle + b^p\right)\right]^2 > 1$. This provides us with a simpler lower bound on $(\alpha_p^0)^2$, namely

$$(\alpha_p^0)^2 \geq \frac{1}{\left\|\sum_{i=1}^{m} \lambda_i^p y_i \Phi(x_i)\right\|^2 \left\|\sum_{i=1}^{m} \mu_i^p y_i \Phi(x_i)\right\|^2}. \qquad (20)$$

Recalling the constraints assigning nonzero values only to some coordinates of λ_p and μ_p that correspond to support vectors the class of which is the opposite of the class of x_p, the terms in the denominator of (20) appear as the difference between two vectors. One of these vectors (in both cases) is $\Phi(x_p)$, whereas the other one is a convex combination, in the feature space, of support vectors of the opposite class (here appears the interest of the additional constraints on μ^p). As a consequence, both vectors belong to the smallest ball

containing the support vectors of (w^0, b^0). By application of the triangle inequality, we thus obtain:

$$\left\| \sum_{i=1}^{m} \lambda_i^p y_i \Phi(x_i) \right\|^2 \leq \mathcal{D}_m^2, \quad \left\| \sum_{i=1}^{m} \mu_i^p y_i \Phi(x_i) \right\|^2 \leq \mathcal{D}_m^2. \tag{21}$$

Thanks to (21), it is possible to substitute both $\|\sum_{i=1}^m \lambda_i^p y_i \Phi(x_i)\|^2$ and $\|\sum_{i=1}^m \mu_i^p y_i \Phi(x_i)\|^2$ with $\mathcal{D}_m^2$ in (20), which concludes the proof. ■

3.2 RM theorem

Theorem 1 *Let us consider a pattern recognition SVM on a domain $\mathcal{X}$. This SVM is characterized by a kernel κ (a projection operator of which we call Φ) and a soft margin constant C. Suppose that it is trained on a subset $s_m = \{(x_i, y_i)\}$, $(1 \leq i \leq m)$ of $\mathcal{X} \times \{-1, 1\}$, the points of which it separates without error. Suppose further that all the corresponding dual variables α_i^0 satisfy $\alpha_i^0 < C$. Let $\mathcal{L}_m$ be the number of errors resulting from applying a leave-one-out procedure to this machine, $\mathcal{D}_m$ the diameter of the smallest ball containing the support vectors of the initial machine and γ its margin. We have:*

$$\mathcal{L}_m \leq \frac{\mathcal{D}_m^2}{\gamma^2}. \tag{22}$$

Proof Actually, the key lemma exhibits a non-trivial bound on α_i^0 when the machine trained on the set without (x_i, y_i) makes an error on this point. Thus, the sum of the α_i^0 is constituted of the terms we can bound thanks to the key lemma ($\alpha_i^0 \geq \frac{1}{\mathcal{D}_m^2}$) and the others with the trivial bound ($\alpha_i^0 \geq 0$). We have thus:

$$\sum_{i=1}^{m} \alpha_i^0 \geq \frac{\mathcal{L}_m}{\mathcal{D}_m^2},$$

and consequently:

$$\mathcal{L}_m \leq \mathcal{D}_m^2 \sum_{i=1}^{m} \alpha_i^0.$$

Since we are in the separable case, we can use the right-hand side equality of (7) and conclude:

$$\mathcal{L}_m \leq \frac{\mathcal{D}_m^2}{\gamma^2}.$$

■

4 Multi-class SVMs

There are several types of multi-class SVMs (see [6] for a survey). We will focus here on the machine introduced independently and under different forms in [14, 10, 2]. For the sake of simplicity, we will only refer to it as "the" M-SVM in the sequel.

4.1 Architecture and algorithm

We first summarize the main points of the training algorithm of this machine. Let us consider a M-SVM on a domain $\mathcal{X}$. It is characterized by a kernel κ (a projection operator of which is Φ) and a soft margin constant C. Let $\mathcal{C} = \{C_k\}$, $(1 \leq k \leq Q)$ be the set of categories. Suppose that the machine is trained on a set $s_m = \{(x_i, C(x_i))\}$, $(1 \leq i \leq m)$ of m couples in $\mathcal{X} \times \mathcal{C}$. The algorithm constructs a set of optimal hyperplanes (w_k, b_k), $(1 \leq k \leq Q)$, by solving a QP problem. Let $W = \{w_1, \ldots, w_k, \ldots, w_Q\}$ and $b = \{b_1, \ldots, b_k, \ldots, b_Q\}$. The model corresponds to the class of functions h:

$$\forall x \in \mathcal{X}, \quad h(x) = \begin{bmatrix} \langle w_1, \Phi(x) \rangle \\ \vdots \\ \langle w_k, \Phi(x) \rangle \\ \vdots \\ \langle w_Q, \Phi(x) \rangle \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_k \\ \vdots \\ b_Q \end{bmatrix}. \tag{23}$$

This QP problem is the following one:

Problem 3

$$\min_{W,b} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1, k \neq C(x_i)}^{Q} \xi_{ik} \right\}$$

$$s.t. \begin{cases} \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}, \quad (1 \leq i \leq m), (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, \quad (1 \leq i \leq m), (1 \leq k \neq C(x_i) \leq Q) \end{cases}.$$

Let the couple (W^0, b^0) denote the solution of this optimization problem. The Lagrangian function associated with Problem 3 is given by:

$$L(W, b, \xi, \alpha, \beta) =$$

$$\frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} - \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \left\{ \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)} - b_k - 1 + \xi_{ik} \right\}$$

$$- \sum_{i=1}^{m} \sum_{k=1}^{Q} \beta_{ik} \xi_{ik}. \tag{24}$$

Setting the gradient of (24) with respect to w_k equal to the null vector gives the expression of w_k as a function of the α_{ik}, which is:

$$w_k = \sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha_{il} \Phi(x_i) - \sum_{i=1}^{m} \alpha_{ik} \Phi(x_i). \tag{25}$$

Setting the gradient of (24) with respect to b_k equal to the null vector gives

$$\sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0. \tag{26}$$

Setting the gradient of (24) with respect to ξ_k equal to the null vector gives

$$\alpha_{ik} + \beta_{ik} = C, \;\; (1 \leq i \leq m), (1 \leq k \leq Q), k \neq C(x_i). \tag{27}$$

Substituting (25) and using (26) and (27) in (24) leads to the Wolfe-dual formulation of Problem 3, the constraints of which are deduced from (26) and (27):

Problem 4

$$\max_{\alpha} \{J(\alpha)\}$$

$$s.t. \begin{cases} 0 \leq \alpha_{ik} < C, \;\; (1 \leq i \leq m), (1 \leq k \leq Q), k \neq C(x_i) \\ \sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0, \;\; (1 \leq k \leq Q-1) \end{cases}$$

where:

$$J(\alpha) =$$

$$-\frac{1}{2} \left\{ \sum_{i \simeq j} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik} \alpha_{jl} \kappa(x_i, x_j) - 2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \alpha_{jC(x_i)} \kappa(x_i, x_j) + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \alpha_{jk} \kappa(x_i, x_j) \right\}$$

$$+ \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}.$$

with $i \simeq j$ meaning that x_i and x_j belong to the same category.

In the formulation of Problem 4, we have implicitly introduced the dummy variables $\alpha_{iC(x_i)} = 0$, $(1 \leq i \leq m)$. In what follows, they will be used again wherever they make notation simpler. The objective function can be expressed as

$$J(\alpha) = -\frac{1}{2} \alpha^T H \alpha + 1^T_{(Q-1)m} \alpha$$

where the general term of the Hessian matrix H can be found in the technical documentation of the M-SVM software (`http://www.loria.fr/~guermeur/technical_doc/index.html`).

Let $\alpha^0 = \left[\alpha^0_{11}, \ldots, \alpha^0_{1Q}, \ldots, \alpha^0_{i1}, \ldots, \alpha^0_{iQ}, \ldots, \alpha^0_{mQ}\right]^T$ be the optimal solution of Problem 4. According to (25), the expression of w^0_k is then:

$$w^0_k = \sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha^0_{il} \Phi(x_i) - \sum_{i=1}^{m} \alpha^0_{ik} \Phi(x_i). \tag{28}$$

Definition 1 (Multi-class margin) *γ_{kl}, the* margin *between categories C_k and C_l, is defined as in the bi-class case as the smallest distance of a point either in C_k or C_l to the hyperplane separating those categories.*

For $1 \leq k < l \leq Q$, let δ_{kl} be:

$$\delta_{kl} = \min \left[\left\{ \min_{x_i \in C_k} \langle w_k - w_l, \Phi(x_i) \rangle + b_k - b_l - 1, \min_{x_j \in C_l} \langle w_l - w_k, \Phi(x_i) \rangle + b_l - b_k - 1 \right\} \right]. \tag{29}$$

We have:

$$\gamma_{kl} = \frac{1 + \delta_{kl}}{\|w_k - w_l\|}. \tag{30}$$

Note that we know the values of the δ_{kl} as soon as (w^0, b^0) is known.

4.2 Some useful results

We introduce some results needed in the proof of the extensions of the key lemma and the RM Theorem. Note that the general term of vector $1 - H\alpha^0$ is simply $\frac{\partial}{\partial \alpha_{ik}} J(\alpha^0)$. It is thus equal to

$$1 - \left\{ \sum_{x_j \in C(x_i)} \sum_{l=1}^{Q} \alpha^0_{jl} \kappa(x_j, x_i) - \sum_{j=1}^{m} \alpha^0_{jC(x_i)} \kappa(x_j, x_i) - \sum_{x_j \in C_k} \sum_{l=1}^{Q} \alpha^0_{jl} \kappa(x_j, x_i) + \sum_{j=1}^{m} \alpha^0_{jk} \kappa(x_j, x_i) \right\}.$$

Keeping in mind the expression of the vectors defining the hyperplanes (25), this can be rewritten as

$$1 - \langle w^0_{C(x_i)} - w^0_k, \Phi(x_i) \rangle = b^0_{C(x_i)} - b^0_k + \xi^0_{ik}. \tag{31}$$

Let us express the relationship between $\sum_{k<l} \|w_k - w_l\|^2$ and $\sum_{k=1}^{Q} \|w_k\|^2$:

$$\sum_{k<l} \|w_k - w_l\|^2 = (Q-1) \sum_{k=1}^{Q} \|w_k\|^2 - 2 \sum_{k<l} \langle w_k, w_l \rangle. \tag{32}$$

Keeping in mind that $\sum_{k=1}^{Q} w_k = 0$, it springs that

$$\|\sum_{k=1}^{Q} w_k\|^2 = 0 = \sum_{k=1}^{Q} \sum_{l=1}^{Q} \langle w_k, w_l \rangle = \sum_{k=1}^{Q} \|w_k\|^2 + 2 \sum_{k<l} \langle w_k, w_l \rangle.$$

Thus,

$$\sum_{k=1}^{Q} \|w_k\|^2 = -2 \sum_{k<l} \langle w_k, w_l \rangle$$

which, by substitution in (32) leads to

$$\sum_{k<l} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2. \tag{33}$$

Proposition 2 *We have:*

$$\alpha^{0T} H \alpha^0 = \sum_{k=1}^{Q} \|w_k^0\|^2 = \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 = \frac{1}{Q} \sum_{k<l} \frac{(1+\delta_{kl})^2}{\gamma_{kl}^2}. \tag{34}$$

Proof

- $\alpha^{0T} H \alpha^0 = \sum_{k=1}^{Q} \|w_k^0\|^2$
  Given the equality between the primal and the dual objective functions at the optimum, we have:
  $$-\frac{1}{2} \alpha^{0T} H \alpha^0 + 1_{(Q-1)m}^T \alpha = \frac{1}{2} \sum_{k=1}^{Q} \|w_k^0\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik}^0.$$
  Combining it with (31), recalling that the ξ_{ik}^0 are null,

$$-\frac{1}{2} \alpha^{0T} H \alpha^0 = \frac{1}{2} \sum_{k=1}^{Q} \|w_k^0\|^2 - \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \left\{ \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)}^0 - b_k^0 \right\}. \tag{35}$$

  We have:
  $$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 (b_{C(x_i)}^0 - b_k^0) = \sum_{l=1}^{Q} \left\{ \sum_{x_i \in C_l} \sum_{k=1}^{Q} \alpha_{ik}^0 - \sum_{i=1}^{m} \alpha_{il}^0 \right\} b_l,$$
  which, due to (26), implies

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 (b_{C(x_i)}^0 - b_k^0) = 0. \tag{36}$$

  (35) thus simplifies into:

$$-\frac{1}{2} \alpha^{0T} H \alpha^0 = \frac{1}{2} \sum_{k=1}^{Q} \|w_k^0\|^2 - \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \langle w_{C(x_i)}^0 - w_k^0, \Phi(x_i) \rangle. \tag{37}$$

Given (25),

$$\|w_k^0\|^2 =$$

$$\langle \sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha_{il}^0 \Phi(x_i) - \sum_{i=1}^{m} \alpha_{ik}^0 \Phi(x_i), \sum_{x_j \in C_k} \sum_{n=1}^{Q} \alpha_{jn}^0 \Phi(x_j) - \sum_{j=1}^{m} \alpha_{jk}^0 \Phi(x_j) \rangle =$$

$$\sum_{x_i \in C_k} \sum_{x_j \in C_k} \sum_{l=1}^{Q} \sum_{n=1}^{Q} \alpha_{ik}^0 \alpha_{jn}^0 \kappa(x_i, x_j) - 2 \sum_{x_i \in C_k} \sum_{j=1}^{m} \sum_{l=1}^{Q} \alpha_{il}^0 \alpha_{jk}^0 \kappa(x_i, x_j) + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{ik}^0 \alpha_{jk}^0 \kappa(x_i, x_j).$$

Summing over the index k thus gives:

$$\sum_{k=1}^{Q} \|w_k^0\|^2 =$$

$$\sum_{i \simeq j} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik}^0 \alpha_{jl}^0 \kappa(x_i, x_j) - 2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \alpha_{jC(x_i)}^0 \kappa(x_i, x_j) + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \alpha_{jk}^0 \kappa(x_i, x_j). \tag{38}$$

Still using (25), we find

$$\alpha_{ik}^0 \langle w_{C(x_i)}^0 - w_k^0, \Phi(x_i) \rangle =$$

$$\sum_{x_j \in C(x_i)} \sum_{l=1}^{Q} \alpha_{ik}^0 \alpha_{jl}^0 \kappa(x_i, x_j) - \sum_{j=1}^{m} \alpha_{ik}^0 \alpha_{jC(x_i)}^0 \kappa(x_i, x_j)$$

$$- \sum_{x_j \in C_k} \sum_{l=1}^{Q} \alpha_{ik}^0 \alpha_{jl}^0 \kappa(x_i, x_j) + \sum_{j=1}^{m} \alpha_{ik}^0 \alpha_{jk}^0 \kappa(x_i, x_j).$$

Summing over the index k gives:

$$\sum_{k=1}^{Q} \alpha_{ik}^0 \langle w_{C(x_i)}^0 - w_k^0, \Phi(x_i) \rangle =$$

$$\sum_{x_j \in C(x_i)} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik}^0 \alpha_{jl}^0 \kappa(x_i, x_j) - \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \alpha_{jC(x_i)}^0 \kappa(x_i, x_j)$$

$$- \sum_{j=1}^{m} \sum_{l=1}^{Q} \alpha_{iC(x_j)}^0 \alpha_{jl}^0 \kappa(x_i, x_j) + \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \alpha_{jk}^0 \kappa(x_i, x_j).$$

Summing over the index i gives:

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \langle w_{C(x_i)}^0 - w_k^0, \Phi(x_i) \rangle =$$

$$\sum_{i\simeq j}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\alpha_{ik}^0\alpha_{jl}^0\kappa(x_i,x_j)-\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0\alpha_{jC(x_i)}^0\kappa(x_i,x_j)$$

$$-\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{Q}\alpha_{iC(x_j)}^0\alpha_{jl}^0\kappa(x_i,x_j)+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0\alpha_{jk}^0\kappa(x_i,x_j)=$$

$$\sum_{i\simeq j}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\alpha_{ik}^0\alpha_{jl}^0\kappa(x_i,x_j)-2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0\alpha_{jC(x_i)}^0\kappa(x_i,x_j)+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0\alpha_{jk}^0\kappa(x_i,x_j). \tag{39}$$

From (38) and (39), it springs that:

$$\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0(\langle w_{C(x_i)}^0-w_k^0,\Phi(x_i)\rangle)=\sum_{k=1}^{Q}\|w_k^0\|^2, \tag{40}$$

which, by substitution in (37), provides us with the announced result.

- $\sum_{k=1}^{Q}\|w_k\|^2=\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}$.
  Since by hypothesis all the ξ_{ik} are null, one of the Kuhn-Tucker conditions is:

$$\alpha_{ik}^0\left\{\langle w_{C(x_i)}^0-w_k^0,\Phi(x_i)\rangle+b_{C(x_i)}^0-b_k^0-1\right\}=0,$$

  and thus:

$$\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0\left\{\langle w_{C(x_i)}^0-w_k^0,\Phi(x_i)\rangle+b_{C(x_i)}^0-b_k^0-1\right\}=0.$$

  Recalling (36) and (40), this directly leads to:

$$\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0\left\{\langle w_{C(x_i)}^0-w_k^0,\Phi(x_i)\rangle+b_{C(x_i)}^0-b_k^0-1\right\}=$$

$$\sum_{k=1}^{Q}\|w_k^0\|^2-\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^0=0.$$

- $\sum_{k=1}^{Q}\|w_k^0\|^2=\frac{1}{Q}\sum_{k<l}\frac{(1+\delta_{kl})^2}{\gamma_{kl}^2}$
  This is deduced from combining (30) and (33).

■

5 Multi-class radius-margin bound

5.1 Key lemma in the multi-class case

We now deal with the extension of the key lemma in the multi-class context.

Lemma 2 *Let us consider a multi-class SVM on a domain $\mathcal{X}$. Let $\mathcal{C} = \{C_k\}$, $(1 \leq k \leq Q)$ be the set of categories. This M-SVM is characterized by a kernel κ (a projection operator Φ) and a soft margin constant C. Suppose that it is trained on a set $s_m = \{(x_i, C(x_i))\}$, $(1 \leq i \leq m)$ of m couples in $\mathcal{X} \times \mathcal{C}$, which it discriminates without error. Suppose further that all the corresponding dual variables α^0_{ik} satisfy $\alpha^0_{ik} < C$. Consider now the same machine, trained on $s_m \setminus \{(x_p, C(x_p))\}$ (the corresponding dual variables are α^p_{ik}). If it makes an error on $(x_p, C(x_p))$, classifying this example in category C_n, then the inequality*

$$\alpha^0_{pn} \geq \frac{1}{K\mathcal{D}_m^2} \tag{41}$$

holds, where $\mathcal{D}_m$ is the diameter of the smallest sphere containing the projections of the support vectors in the feature space. K is defined, using the support vectors, as:

$$K = \sqrt{\max_{i,k \,:\, \alpha^0_{ik} > 0} (K_{\lambda,i,k}.K_{\mu,i,k})}, \tag{42}$$

where $\lambda = [\lambda_{ik}] \in \mathbb{R}_+^{Qm}$, $\lambda_{iC(x_i)} = 0$, $(1 \leq i \leq m)$, and $K_{\lambda,i,k}$ is the value of the objective function at the optimal solution of the QP problem:

Problem 5

$$\min_{\lambda} J_{ik}(\lambda)$$

$$s.t. \begin{cases} \forall l, \;\; \lambda_{il} = \frac{\alpha^0_{jl}}{\alpha^0_{ik}} \\ \forall j \neq i, \forall l, \;\; 0 \leq \lambda_{jl} \leq \frac{\alpha^0_{jl}}{\alpha^0_{ik}} \\ \sum_{x_j \in C_l} \sum_{o=1}^{Q} \lambda_{jo} - \sum_{q=1}^{m} \lambda_{ql} = 0, \;\; (1 \leq l \leq Q-1) \end{cases}$$

with

$$J_{ik}(\lambda) = \sum_{l=1}^{Q} \left(\sum_{j=1}^{m} \lambda_{jl} \right)^2 .$$

$\mu = [\mu_{ik}] \in \mathbb{R}_+^{mQ}$, $\mu_{iC(x_i)} = 0$, $(1 \leq i \leq m)$, and $K_{\mu,i,k}$ is the value of the objective function at the optimal solution of the QP problem:

Problem 6

$$\min_{\mu} J_{ik}(\mu)$$

$$\begin{cases} \begin{cases} l = k, \ \mu_{il} = 1 \\ l \neq k \ \mu_{il} = 0 \end{cases} \\ \forall j \neq i, \forall l, \ \mu_{jl} \geq 0 \\ \sum_{x_j \in C_l} \sum_{o=1}^{Q} \mu_{jo} - \sum_{q=1}^{m} \mu_{ql} = 0, \ \ (1 \leq l \leq Q-1) \end{cases}$$

J_{ik} is defined as for λ:

$$J_{ik}(\mu) = \sum_{l=1}^{Q} \left(\sum_{j=1}^{m} \mu_{jl} \right)^2 .$$

Proof Let (W^p, b^p) be the couple characterizing the hyperplanes when the machine is trained on $s_m \setminus \{(x_p, C(x_p))\}$. Let

$$\alpha^p = [\alpha^p_{11}, \ldots, \alpha^p_{(p-1)Q}, 0, \ldots, 0, \alpha^p_{(p+1)1}, \ldots, \alpha^p_{mQ}]^T$$

be the vector of $[0, C)^{mQ}$ the components of which are the dual variables of the second M-SVM, with $\alpha^p_{pk} = 0$, $(1 \leq k \leq Q)$. This representation is used to characterize directly the second M-SVM with respect to the first one. Indeed, α^p is an optimal solution of Problem 4 under the additional constraint $\alpha^p_{pk} = 0$, $(1 \leq k \leq Q)$. Let us define two more vectors in $\mathbb{R}_+^{mQ}$, $\lambda^p = [\lambda^p_{ik}]$ and $\mu^p = [\mu^p_{ik}]$ $(1 \leq i \leq m), (1 \leq k \leq Q)$, with $\lambda^p_{iC(x_i)} = \mu^p_{iC(x_i)} = 0$, $(1 \leq i \leq m)$. λ^p satisfies additional properties so that the vector $\alpha^0 - \alpha^0_{pn}\lambda^p$ is a feasible solution of Problem 4 under the additional constraint that $\alpha^0_{pk} - \alpha^0_{pn}\lambda^p_{pk} = 0$, $(1 \leq k \leq Q)$, i.e. $\alpha^0 - \alpha^0_{pn}\lambda^p$ satisfies the same constraints as α^p. We have thus:

$$\forall i \neq p, \forall k \neq C(x_i), \ \ \alpha^0_{ik} - \alpha^0_{pn}\lambda^p_{ik} \geq 0 \Longleftrightarrow \lambda^p_{ik} \leq \frac{\alpha^0_{ik}}{\alpha^0_{pn}}$$

$$\forall i \neq p, \forall k \neq C(x_i), \ \ \lambda^p_{ik} \geq 0 \Longrightarrow \alpha^0_{ik} - \alpha^0_{pn}\lambda^p_{ik} \leq C$$

$$\forall k, \ \ \alpha^0_{pk} - \alpha^0_{pn}\lambda^p_{pk} = 0 \Longleftrightarrow \lambda^p_{pk} = \frac{\alpha^0_{pk}}{\alpha^0_{pn}}$$

$$\sum_{x_i \in C_k} \sum_{l=1}^{Q} \left(\alpha^0_{il} - \alpha^0_{pn}\lambda^p_{il}\right) - \sum_{i=1}^{m} \left(\alpha^0_{ik} - \alpha^0_{pn}\lambda^p_{ik}\right) = 0 \Longleftrightarrow$$

$$\sum_{x_i \in C_k} \sum_{l=1}^{Q} \lambda^p_{il} - \sum_{i=1}^{m} \lambda^p_{ik} = 0.$$

To sum up, vector λ^p satisfies the following properties:

$$\begin{cases} \forall k, \ \ \lambda^p_{pk} = \frac{\alpha^0_{pk}}{\alpha^0_{pn}} \\ \forall i \neq p, \forall k, \ \ 0 \leq \lambda^p_{ik} \leq \frac{\alpha^0_{ik}}{\alpha^0_{pn}} \\ \sum_{x_i \in C_k} \sum_{l=1}^{Q} \lambda^p_{il} - \sum_{i=1}^{m} \lambda^p_{ik} = 0, \ \ (1 \leq k \leq Q-1) \end{cases} . \tag{43}$$

μ^p is specified so that $\alpha^p + c\mu^p$ satisfies the same constraints as α^0, where c is a scalar belonging to $(0, C]$. Furthermore, we impose that $\mu^p_{pk} = 0$ for k different from n and $\mu^p_{pn} = 1$.

$$\forall i \neq p, \forall k \neq C(x_i), \;\; \mu^p_{ik} \geq 0 \Longrightarrow \alpha^p_{ik} + c\mu^p_{ik} \geq 0$$

$$\forall i \neq p, \forall k \neq C(x_i), \;\; \alpha^p_{ik} + c\mu^p_{ik} \leq C \Longleftrightarrow \mu^p_{ik} \leq \frac{C - \alpha^p_{ik}}{c}$$

$$\sum_{x_i \in C_k} \sum_{l=1}^{Q} (\alpha^p_{il} + c\mu^p_{il}) - \sum_{i=1}^{m} (\alpha^p_{ik} + c\mu^p_{ik}) = 0 \Longleftrightarrow \sum_{x_i \in C_k} \sum_{l=1}^{Q} \mu^p_{il} - \sum_{i=1}^{m} \mu^p_{ik} = 0.$$

Remember that since we can choose an arbitrary large C, the constraint $\mu^p_{ik} \leq \frac{C - \alpha^p_{ik}}{c}$ vanishes. To sum up, vector μ^p satisfies the following properties:

$$\begin{cases} \begin{cases} k = n, \;\; \mu^p_{ik} = 1 \\ k \neq n \;\; \mu^p_{ik} = 0 \end{cases} \\ \forall i \neq p, \forall k, \;\; \mu^p_{ik} \geq 0 \\ \sum_{x_i \in C_k} \sum_{l=1}^{Q} \mu^p_{il} - \sum_{i=1}^{m} \mu^p_{ik} = 0, \;\; (1 \leq k \leq Q-1) \end{cases} . \tag{44}$$

By construction of vectors λ^p and μ^p, we have $J(\alpha^0 - \alpha^0_{pn}\lambda^p) \leq J(\alpha^p)$ and $J(\alpha^p + c\mu^p) \leq J(\alpha^0)$, and by way of consequence,

$$J(\alpha^0) - J(\alpha^0 - \alpha^0_{pn}\lambda^p) \geq J(\alpha^0) - J(\alpha^p) \geq J(\alpha^p + c\mu^p) - J(\alpha^p). \tag{45}$$

Since

$$J(\alpha^0 - \alpha^0_{pn}\lambda^p) = -\frac{1}{2}(\alpha^0 - \alpha^0_{pn}\lambda^p)^T H(\alpha^0 - \alpha^0_{pn}\lambda^p) + 1^T(\alpha^0 - \alpha^0_{pn}\lambda^p)$$

the expression of the first term is

$$J(\alpha^0) - J(\alpha^0 - \alpha^0_{pn}\lambda^p) = \frac{\left(\alpha^0_{pn}\right)^2}{2} \lambda^{pT} H \lambda^p + \alpha^0_{pn} \lambda^{pT} \left(1 - H\alpha^0\right).$$

But recalling (31) we can simplify:

$$\lambda^{pT} \left(1 - H\alpha^0\right) = \sum_{i=1}^{m} \sum_{k \neq C(x_i)} \left(b_{C(x_i)} - b_k\right) \lambda^p_{ik} + \sum_{i=1}^{m} \sum_{k \neq C(x_i)} \lambda^p_{ik} \xi^0_{ik}.$$

But

$$\sum_{i=1}^{m} \sum_{k \neq C(x_i)} \left(b_{C(x_i)} - b_k\right) \lambda^p_{ik} = \sum_{l=1}^{Q} \left(\sum_{x_i \in C_l} \sum_{k \neq l} \lambda^p_{ik} - \sum_{x_i \notin C_l} \lambda^p_{il} \right) b_l = 0,$$

thus

$$\lambda^{pT} \left(1 - H\alpha^0\right) = \sum_{i=1}^{m} \sum_{k \neq C(x_i)} \lambda^p_{ik} \xi^0_{ik}.$$

At this point, we make use of the hypothesis that all the dual variables α^0_{ik} are strictly inferior to C. According to the Kuhn-Tucker conditions, this implies that all the slack variables ξ^0_{ik} are equal to 0. Thus, the expression of $J(\alpha^0) - J(\alpha^0 - \alpha^0_{pn}\lambda^p)$ is simply

$$J(\alpha^0) - J(\alpha^0 - \alpha^0_{pn}\lambda^p) = \frac{\left(\alpha^0_{pn}\right)^2}{2}\lambda^{pT} H \lambda^p.$$

Using (34), we find

$$J(\alpha^0) - J(\alpha^0 - \alpha^0_{pn}\lambda^p) = \frac{\left(\alpha^0_{pn}\right)^2}{2}\sum_{k=1}^{Q}\left\| \sum_{x_i \in C_k}\sum_{l=1}^{Q}\lambda^p_{il}\Phi(x_i) - \sum_{i=1}^{m}\lambda^p_{ik}\Phi(x_i)\right\|^2. \tag{46}$$

We turn now to the right-hand side of (45).

$$J(\alpha^p + c\mu^p) - J(\alpha^p) = c\mu^{pT}(1 - H\alpha^p) - \frac{c^2}{2}\sum_{k=1}^{Q}\left\| \sum_{x_i \in C_k}\sum_{l=1}^{Q}\mu^p_{il}\Phi(x_i) - \sum_{i=1}^{m}\mu^p_{ik}\Phi(x_i)\right\|^2.$$

Proceeding as in the case of the left-hand side of (45), we establish that

$$\mu^{pT}(1 - H\alpha^p) = \sum_{i=1}^{m}\sum_{k \neq C(x_i)}\mu^p_{ik}\xi^p_{ik}.$$

If the initial SVM, characterized by the couple (W^0, b^0), has nothing but null slack variables, then the sole slack variables of the SVM (W^p, b^p) which can be positive are thoses associated with x_p, i.e. the variables ξ^p_{pk}, $(1 \leq k \neq C(x_p) \leq Q)$. Thus,

$$\mu^{pT}(1 - H\alpha^p) = \sum_{k \neq C(x_p)}\mu^p_{pk}\xi^p_{pk}.$$

Since $\mu^p_{pk} = 0$ for $k \neq n$ and $\mu^p_{pn} = 1$, this simplifies into:

$$\mu^{pT}(1 - H\alpha^p) = \xi^p_{pn} = 1 - \langle w_{C(x_p)} - w_n, \Phi(x_p)\rangle - b_{C(x_p)} + b_n$$

and by way of consequence

$$J(\alpha^p + c\mu^p) - J(\alpha^p) = c\left[1 - \langle w_{C(x_p)} - w_n, \Phi(x_p)\rangle - b_{C(x_p)} + b_n\right] -$$

$$\frac{c^2}{2}\sum_{k=1}^{Q}\left\| \sum_{x_i \in C_k}\sum_{l=1}^{Q}\mu^p_{il}\Phi(x_i) - \sum_{i=1}^{m}\mu^p_{ik}\Phi(x_i)\right\|^2. \tag{47}$$

Combining (46) and (47) finally gives:

$$\frac{\left(\alpha_{pn}^0\right)^2}{2}\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\lambda_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\lambda_{ik}^p\Phi(x_i)\right\|^2 \geq$$

$$c\left[1-\langle w_{C(x_p)}-w_n,\Phi(x_p)\rangle-b_{C(x_p)}+b_n\right]-\frac{c^2}{2}\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\mu_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\mu_{ik}^p\Phi(x_i)\right\|^2. \tag{48}$$

The value of the scalar c maximizing the right-hand side of (48) is:

$$c^*=\frac{\left[1-\langle w_{C(x_p)}-w_n,\Phi(x_p)\rangle-b_{C(x_p)}+b_n\right]}{\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\mu_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\mu_{ik}^p\Phi(x_i)\right\|^2}.$$

By substitution in (48), this means that $(\alpha_{pn}^0)^2$ can be bounded from below as follows:

$$\left(\alpha_{pn}^0\right)^2 \geq$$

$$\frac{\left[1-\langle w_{C(x_p)}-w_n,\Phi(x_p)\rangle-b_{C(x_p)}+b_n\right]^2}{\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\lambda_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\lambda_{ik}^p\Phi(x_i)\right\|^2\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\mu_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\mu_{ik}^p\Phi(x_i)\right\|^2}.$$

Since by hypothesis, the SVM (W^p,b^p) makes an error on x_p, classifying it in the category C_n,

$$\langle w_{C(x_p)}-w_n,\Phi(x_p)\rangle+b_{C(x_p)}+b_n<0$$

and thus

$$\left[1-\langle w_{C(x_p)}-w_n,\Phi(x_p)\rangle-b_{C(x_p)}+b_n\right]^2>1.$$

This provides us with a simpler lower bound on $(\alpha_{pn}^0)^2$, namely

$$\left(\alpha_{pn}^0\right)^2 \geq$$

$$\frac{1}{\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\lambda_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\lambda_{ik}^p\Phi(x_i)\right\|^2\sum_{k=1}^{Q}\left\|\sum_{x_i\in C_k}\sum_{l=1}^{Q}\mu_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\mu_{ik}^p\Phi(x_i)\right\|^2}. \tag{49}$$

Let $K_{\lambda,p,n,k}$ be:

$$K_{\lambda,p,n,k}=\sum_{i=1}^{m}\lambda_{ik}^p.$$

We have:

$$\left\| \sum_{x_i \in C_k} \sum_{l=1}^{Q} \lambda_{il}^p \Phi(x_i) - \sum_{i=1}^{m} \lambda_{ik}^p \Phi(x_i) \right\|^2 = (K_{\lambda,p,n,k})^2 \left\| Cv_1(\Phi(x_i)) - Cv_2(\Phi(x_i)) \right\|^2$$

where $Cv_1(\Phi(x_i))$ and $Cv_2(\Phi(x_i))$, as in the bi-class case, are two convex combinations the difference of which we can bound by $\mathcal{D}_m$. Then,

$$\sum_{k=1}^{Q} \left\| \sum_{x_i \in C_k} \sum_{l=1}^{Q} \lambda_{il}^p \Phi(x_i) - \sum_{i=1}^{m} \lambda_{ik}^p \Phi(x_i) \right\|^2 \leq \mathcal{D}_m^2 \sum_{k=1}^{Q} (K_{\lambda,p,n,k})^2.$$

Since the same reasoning can be done with vector μ^p in place of the vector λ^p, with $K_{\mu,p,n,k}$ replacing $K_{\lambda,p,n,k}$, this leads to

$$\left(\alpha_{pn}^0\right)^2 \geq \frac{1}{\mathcal{D}_m^4 \sum_{k=1}^{Q} (K_{\lambda,p,n,k})^2 \sum_{k=1}^{Q} (K_{\mu,p,n,k})^2}.$$

By construction, λ^p and μ^p satisfy the constraints of Problems 5 and 6 respectively. As a consequence, they can be chosen so that $\sum_{k=1}^{Q} (K_{\lambda,p,n,k})^2 = K_{\lambda,p,n}$ and $\sum_{k=1}^{Q} (K_{\mu,p,n})^2 = K_{\mu,p,n}$. By definition of constant K, we then get:

$$\sum_{k=1}^{Q} (K_{\lambda,p,n,k})^2 \sum_{k=1}^{Q} (K_{\mu,p,n,k})^2 \leq K^2$$

and finally:

$$\left(\alpha_{pn}^0\right)^2 \geq \frac{1}{\mathcal{D}_m^4 K^2}.$$

Taking the square root concludes the proof of the lemma. ■

5.2 MC-RM theorem

Theorem 2 *Let us consider a multi-class SVM on a domain $\mathcal{X}$. Let $\mathcal{C} = \{C_k\}$, $(1 \leq k \leq Q)$ be the set of categories. This M-SVM is characterized by a kernel κ (a projection operator Φ) and a soft margin constant C. Suppose that it is trained on a set $s_m = \{(x_i, C(x_i))\}$, $(1 \leq i \leq m)$ of m couples in $\mathcal{X} \times \mathcal{C}$, which it discriminates without error. Suppose further that all the corresponding dual variables α_{ik}^0 satisfy $\alpha_{ik}^0 < C$. Let $\mathcal{L}_m$ be the number of errors resulting from applying a leave-one-out procedure to this machine, $\mathcal{D}_m$ the diameter of the smallest ball containing the support vectors of the initial machine and $(\gamma_{kl})_{1 \leq k < l \leq Q}$ its bi-class margins according to Definition 1. Let the δ_{kl} be given by (29) and K be the constant in the key lemma (equation (42)). The following upper bound holds true:*

$$\mathcal{L}_m \leq \frac{K \mathcal{D}_m^2}{Q} \sum_{k<l} \frac{(1+\delta_{kl})^2}{\gamma_{kl}^2}. \tag{50}$$

Proof Actually, the key lemma exhibits a non-trivial bound on α_{in}^0 when the machine trained on the set without (x_i, y_i) makes an error classifying it in C_n. Thus, the sum of the α_{ik}^0 is constituted of the terms we can bound thanks to the key lemma ($\alpha_{ik}^0 \geq \frac{1}{K\mathcal{D}_m^2}$) and the others with the trivial bound ($\alpha_{ik}^0 \geq 0$). We have thus:

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \geq \frac{\mathcal{L}_m}{K\mathcal{D}_m^2}$$

and consequently:

$$\mathcal{L}_m \leq K\mathcal{D}_m^2 \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0.$$

Recalling Proposition 2, we conclude:

$$\mathcal{L}_m \leq K\mathcal{D}_m^2 \frac{1}{Q} \sum_{k<l} \frac{(1+\delta_{kl})^2}{\gamma_{kl}^2}.$$

■

One could find strange the presence of the $\frac{1}{Q}$ coefficient in the right-hand side of (50). Indeed, at first sight, it seems that this formula is not equivalent to the one of Theorem 1 when $Q = 2$. This is however the case, and the difference springs from the fact that the penalty term in the primal objective function is $\sum_{k=1}^{Q} \|w_k\|^2$ instead of $\sum_{k<l} \|w_k - w_l\|^2$, the use of which would make the relationship between the bi-class and the multi-class cases more straightforward.

6 Conclusions and future work

In this report, we have established a direct extension of the radius-margin bound to the case of the standard multi-class SVM. Obviously, its practical interest primarily depends on the possibility to compute the constant K at a reasonable cost. This could be the case, since this computation does not involve the computation of the kernel κ on the points of the training set. If so, then the bi-class variant of the bound could even prove superior to Chapelle's one (since it should be tighter by construction). Our result can be directly compared with those proposed in [13], both from the point of view of sharpness and cpu time requirements. This is the subject of an ongoing study.

Acknowledgements

The work of the authors is supported by the ACI "Masses de Données".

Contents

References

[1] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[2] E.J. Bredensteiner and K.P. Bennett. Multicategory Classification by Support Vector Machines. *Computational Optimization and Applications*, 12(1/3):53–79, 1999.

[3] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.

[4] C. Cortes and V.N. Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[5] R. Fletcher. *Practical Methods of Optimization.* Wiley, 1987.

[6] Y. Guermeur, A. Elisseeff, and D. Zelus. A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers. *Applied Stochastic Models in Business and Industry*, 21(2):199–214, 2005.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Springer Series in Statistics. Springer, 2001.

[8] P. Massart. Concentrations inequalities and model selection. In *Ecole d'Eté de Probabilité de Saint-Flour XXXIII*, LNM. Springer-Verlag, 2003.

[9] M. Opper and O. Winther. Gaussian processes and SVM: Mean field and leave-one-out. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 311–326. The MIT Press, 2000.

[10] V.N. Vapnik. *Statistical learning theory.* John Wiley & Sons, Inc., N.Y., 1998.

[11] V.N. Vapnik and 0. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000.

[12] G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross-validation for support vector machines: another way to look at margin-like quantities. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–309. The MIT Press, 2000.

[13] L. Wang, P. Xue, and K.L. Chan. Generalized Radius-Margin Bounds for Model Selection in Multi-class SVMs. Technical report, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, 2005.

[14] J. Weston and C. Watkins. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399