A Quadratic Loss Multi-Class SVM

Emmanuel Monfrini, Yann Guermeur

A Quadratic Loss Multi-Class SVM

Emmanuel Monfrini — Yann Guermeur

April 30, 2008

A Quadratic Loss Multi-Class SVM

Emmanuel Monfrini* , Yann Guermeur†

— April 30, 2008 — 24 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, the method of choice is cross-validation. Its leave-one-out variant is known to produce an estimator of the generalization error which is almost unbiased. Its major drawback rests in its time requirement. To overcome this difficulty, several upper bounds on the leave-one-out error of the pattern recognition SVM have been derived. Among those bounds, the most popular one is probably the radius-margin bound. It applies to the hard margin pattern recognition SVM, and by extension to the 2-norm SVM. In this report, we introduce a quadratic loss M-SVM, the M-SVM^2, as a direct extension of the 2-norm SVM to the multi-class case. For this machine, a generalized radius-margin bound is then established.

Key-words: M-SVMs, model selection, leave-one-out error, radius-margin bound.

* UMR 7503-UHP
† UMR 7503-CNRS

Une SVM multi-classe à coût quadratique

Résumé : La mise en œuvre d'une machine à vecteurs support requiert la détermination des valeurs de deux types d'hyper-paramètres : le paramètre de "marge douce" C et les paramètres du noyau. Pour effectuer cette tâche de sélection de modèle, la méthode de choix est la validation croisée. Sa variante "leave-one-out" est connue pour fournir un estimateur de l'erreur en généralisation presque sans biais. Son défaut premier réside dans le temps de calcul qu'elle nécessite. Afin de surmonter cette difficulté, plusieurs majorants de l'erreur "leave-one-out" de la SVM calculant des dichotomies ont été proposés. La plus populaire de ces bornes supérieures est probablement la borne "rayon-marge". Elle s'applique à la version à marge dure de la machine, et par extension à la variante dite "de norne 2". Ce rapport introduit une M-SVM "à coût quadratique", la M-SVM^2, comme une extension directe de la SVM de norne 2 au cas multi-classe. Pour cette machine, une borne "rayon-marge" généralisée est ensuite établie.

Mots-clés : M-SVM, sélection de modèle, erreur "leave-one-out", borne "rayon-marge".

1 Introduction

Using a support vector machine (SVM) [2, 4] 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 (see for instance [9, 12]). The solution of choice consists in applying a cross-validation procedure. Among those procedures, the leave-one-out one appears especially attractive, since it is known to produce an estimator of the generalization error which is almost unbiased [11]. The seamy side of things is that it is highly time consuming. This is the reason why, in recent years, a number of upper bounds on the leave-one-out error of pattern recognition SVMs have been proposed in literature (see [3] for a survey). Among those bounds, the tightest one is the span bound [16]. However, the results of Chapelle and co-workers presented in [3] show that another bound, the radius-margin one [15], achieves equivalent performance for model selection while being far simpler to compute. This is the reason why it is currently the most popular bound. It applies to the hard margin machine and, by extension, to the 2-norm SVM (see for instance Chapter 7 in [13]).

In this report, a multi-class extension of the 2-norm SVM is introduced. This machine, named M-SVM^2, is a quadratic loss multi-class SVM, i.e., a multi-class SVM (M-SVM) in which the ℓ_1-norm on the vector of slack variables has been replaced with a quadratic form. The standard M-SVM on which it is based is the one of Lee, Lin and Wahba [10]. As the 2-norm SVM, its training algorithm is equivalent to the training algorithm of a hard margin machine obtained by a simple change of kernel. We then establish a generalized radius-margin bound on the leave-one-out error of the hard margin version of the M-SVM of Lee, Lin and Wahba.

The organization of this paper is as follows. Section 2 presents the multi-class SVMs, by describing their common architecture and the general form taken by their different training algorithms. It focuses on the M-SVM of Lee, Lin and Wahba. In Section 3, the M-SVM^2 is introduced as a particular case of quadratic loss M-SVM. Its connection with the hard margin version of the M-SVM of Lee, Lin and Wahba is highlighted, as well as the fact that it constitutes a multi-class generalization of the 2-norm SVM. Section 4 is devoted to the formulation and proof of the corresponding multi-class radius-margin bound. At last, we draw conclusions and outline our ongoing research in Section 5.

2 Multi-Class SVMs

2.1 Formalization of the learning problem

We are interested here in multi-class pattern recognition problems. Formally, we consider the case of Q-category classification problems with $3 \leq Q < \infty$, but our results extend to the case of dichotomies. Each object is represented by its description $x \in \mathcal{X}$ and the set $\mathcal{Y}$ of the categories y can be identified with the set of indexes of the categories: $[\![1, Q]\!]$. We assume that the link between objects and categories can be described by an unknown probability measure P on the product space $\mathcal{X} \times \mathcal{Y}$. The aim of the learning problem consists in selecting in a set $\mathcal{G}$ of functions $g = (g_k)_{1 \leq k \leq Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$ a function classifying data in an optimal way. The criterion of optimality must be specified. The function g assigns $x \in \mathcal{X}$ to the category l if and only if $g_l(x) > \max_{k \neq l} g_k(x)$. In case of ex æquo, x is assigned to a dummy category denoted by $*$. Let f be the decision function (from $\mathcal{X}$ into $\mathcal{Y} \bigcup \{*\}$) associated with g. With these definitions at hand, the objective function to be minimized is the probability of error $P\left(f\left(X\right) \neq Y\right)$. The optimization process, called *training*, is based on empirical data. More precisely, we assume that there exists a random pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, distributed according to P, and we are provided with a m-sample $D_m = ((X_i, Y_i))_{1 \leq i \leq m}$ of independent copies of (X, Y).

There are two questions raised by such problems: how to properly choose the class of functions $\mathcal{G}$ and how to determine the best candidate g^* in this class, using only D_m. This report addresses the first question, named *model selection*, in the particular case when the model considered is a M-SVM. The second question, named *function selection*, is addressed for instance in [8].

2.2 Architecture and training algorithms

M-SVMs, like all the SVMs, belong to the family of kernel machines. As such, they operate on a class of functions induced by a positive semidefinite (Mercer) kernel. This calls for the formulation of some definitions and propositions.

Definition 1 (Positive semidefinite kernel) *A* positive semidefinite kernel κ *on the set* $\mathcal{X}$ *is a continuous and symmetric function* $\kappa : \mathcal{X}^2 \rightarrow \mathbb{R}$ *verifying:*

$$\forall n \in \mathbb{N}^*, \ \forall \left(x_i\right)_{1 \leq i \leq n} \in \mathcal{X}^n, \ \forall \left(a_i\right)_{1 \leq i \leq n} \in \mathbb{R}^n, \ \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa\left(x_i, x_j\right) \geq 0.$$

Definition 2 (Reproducing kernel Hilbert space [1]) *Let* $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ *be a Hilbert space of functions on* $\mathcal{X}$ *(*$\mathbf{H} \subset \mathbb{R}^{\mathcal{X}}$*). A function* $\kappa : \mathcal{X}^2 \rightarrow \mathbb{R}$ *is a* reproducing kernel *of* $\mathbf{H}$ *if and only if:*

1. $\forall x \in \mathcal{X}, \ \kappa_x = \kappa\left(x, \cdot\right) \in \mathbf{H}$*;*

2. $\forall x \in \mathcal{X}, \forall h \in \mathbf{H}, \ \langle h, \kappa_x \rangle_{\mathbf{H}} = h(x)$ *(reproducing property).*

A Hilbert space of functions which possesses a reproducing kernel is called a reproducing kernel Hilbert space *(RKHS).*

Proposition 1 *Let $(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})$ be a RKHS of functions on $\mathcal{X}$ with reproducing kernel κ. Then, there exists a map Φ from $\mathcal{X}$ into a Hilbert space $\left(E_{\Phi(\mathcal{X})}, \langle \cdot, \cdot \rangle\right)$ such that:*

$$\forall \left(x, x'\right) \in \mathcal{X}^2, \ \kappa\left(x, x'\right) = \langle \Phi\left(x\right), \Phi\left(x'\right) \rangle. \tag{1}$$

Φ is called a feature map *and $E_{\Phi(\mathcal{X})}$ a* feature space*.*

The connection between positive semidefinite kernels and RKHS is the following.

Proposition 2 *If κ is a positive semidefinite kernel on $\mathcal{X}$, then there exists a RKHS $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ of functions on $\mathcal{X}$ such that κ is a reproducing kernel of $\mathbf{H}$.*

Let κ be a positive semidefinite kernel on $\mathcal{X}$ and let $(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})$ be the RKHS spanned by κ. Let $\bar{\mathcal{H}} = (\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})^Q$ and let $\mathcal{H} = ((\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}) + \{1\})^Q$. By construction, $\mathcal{H}$ is the class of vector-valued functions $h = (h_k)_{1 \leq k \leq Q}$ on $\mathcal{X}$ such that

$$h(\cdot) = \left(\sum_{i=1}^{m_k} \beta_{ik} \kappa\left(x_{ik}, \cdot\right) + b_k \right)_{1 \leq k \leq Q}$$

where the x_{ik} are elements of $\mathcal{X}$, as well as the limits of these functions when the sets $\{x_{ik} : 1 \leq i \leq m_k\}$ become dense in $\mathcal{X}$ in the norm induced by the dot product (see for instance [17]). Due to Equation 1, $\mathcal{H}$ can be seen as a multivariate affine model on $\Phi\left(\mathcal{X}\right)$. Functions h can then be rewritten as:

$$h(\cdot) = \left(\langle w_k, \cdot \rangle + b_k \right)_{1 \leq k \leq Q}$$

where the vectors w_k are elements of $E_{\Phi(\mathcal{X})}$. They are thus described by the pair $(\mathbf{w}, \mathbf{b})$ with $\mathbf{w} = (w_k)_{1 \leq k \leq Q} \in E_{\Phi(\mathcal{X})}^Q$ and $\mathbf{b} = (b_k)_{1 \leq k \leq Q} \in \mathbb{R}^Q$. As a consequence, $\bar{\mathcal{H}}$ can be seen as a multivariate linear model on $\Phi\left(\mathcal{X}\right)$, endowed with a norm $\|.\|_{\bar{\mathcal{H}}}$ given by:

$$\forall \bar{h} \in \bar{\mathcal{H}}, \ \left\| \bar{h} \right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2} = \|\mathbf{w}\|,$$

where $\|w_k\| = \sqrt{\langle w_k, w_k \rangle}$. With these definitions and propositions at hand, a generic definition of the M-SVMs can be formulated as follows.

Definition 3 (M-SVM, Definition 42 in [8]) *Let $\left((x_i, y_i)\right)_{1 \leq i \leq m} \in \left(\mathcal{X} \times [\![1, Q]\!]\right)^m$ and $\lambda \in \mathbb{R}_+^*$. A Q-*category M-SVM *is a large margin discriminant model obtained by minimizing over the hyperplane $\sum_{k=1}^{Q} h_k = 0$ of $\mathcal{H}$ a penalized risk $J_{M\text{-}SVM}$ of the form:*

$$J_{M\text{-}SVM}\left(h\right) = \sum_{i=1}^{m} \ell_{M\text{-}SVM}\left(y_i, h\left(x_i\right)\right) + \lambda \left\| \bar{h} \right\|_{\bar{\mathcal{H}}}^2$$

where the data fit component involves a loss function $\ell_{M\text{-}SVM}$ which is convex.

Three main models of M-SVMs can be found in literature. The oldest one is the model of Weston and Watkins [19], which corresponds to the loss function ℓ_{WW} given by:

$$\ell_{\mathrm{WW}}(y, h(x)) = \sum_{k \neq y} \left(1 - h_y(x) + h_k(x)\right)_+ ,$$

where the *hinge loss* function $(\cdot)_+$ is the function $\max(0, \cdot)$. The second one is due to Crammer and Singer [5] and corresponds to the loss function ℓ_{CS} given by:

$$\ell_{\mathrm{CS}}(y, \bar{h}(x)) = \left(1 - \bar{h}_y(x) + \max_{k \neq y} \bar{h}_k(x)\right)_+ .$$

The most recent model is the one of Lee, Lin and Wahba [10] which corresponds to the loss function ℓ_{LLW} given by:

$$\ell_{\mathrm{LLW}}\left(y, h(x)\right) = \sum_{k \neq y} \left(h_k(x) + \frac{1}{Q-1}\right)_+ . \tag{2}$$

Among the three models, the M-SVM of Lee, Lin and Wahba is the only one that implements asymptotically the Bayes decision rule. It is *Fisher consistent* [20, 14].

2.3 The M-SVM of Lee, Lin and Wahba

The substitution in Definition 3 of $\ell_{\text{M-SVM}}$ with the expression of the loss function ℓ_{LLW} given by Equation 2 provides us with the expressions of the quadratic programming (QP) problems corresponding to the training algorithms of the hard margin and soft margin versions of the M-SVM of Lee, Lin and Wahba.

Problem 1 (Hard margin M-SVM)

$$\min_{\mathbf{w}, \mathbf{b}} J_{HM}\left(\mathbf{w}, \mathbf{b}\right)$$

$$s.t. \begin{cases} \langle w_k, \Phi(x_i) \rangle + b_k \leq -\frac{1}{Q-1}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{k=1}^{Q} w_k = 0 \\ \sum_{k=1}^{Q} b_k = 0 \end{cases}$$

where

$$J_{HM}\left(\mathbf{w}, \mathbf{b}\right) = \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 .$$

Problem 2 (Soft margin M-SVM)

$$\min_{\mathbf{w}, \mathbf{b}} J_{SM}\left(\mathbf{w}, \mathbf{b}\right)$$

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

where

$$
J_{SM}(\mathbf{w}, \mathbf{b}) = \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k \neq y_i} \xi_{ik}.
$$

In Problem 2, the ξ_{ik} are *slack variables* introduced in order to relax the constraints of correct classification. The coefficient C, which characterizes the trade-off between prediction accuracy on the training set and smoothness of the solution, can be expressed in terms of the regularization coefficient λ as follows: $C = (2\lambda)^{-1}$. It is called the *soft margin parameter*. Instead of directly solving Problems 1 and 2, one usually solves their Wolfe dual [6]. We now derive the dual problem of Problem 1. Giving the details of the implementation of the Lagrangian duality will provide us with partial results which will prove useful in the sequel.

Let $\alpha = (\alpha_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q} \in \mathbb{R}_+^{Qm}$ be the vector of Lagrange multipliers associated with the constraints of good classification. It is for convenience of notation that this vector is expressed with double subscript and that the dummy variables α_{iy_i}, all equal to 0, are introduced. Let $\delta \in E_{\Phi(\mathcal{X})}$ be the Lagrange multiplier associated with the constraint $\sum_{k=1}^{Q} w_k = 0$ and $\beta \in \mathbb{R}$ the Lagrange multiplier associated with the constraint $\sum_{k=1}^{Q} b_k = 0$. The Lagrangian function of Problem 1 is given by:

$$
L(\mathbf{w}, \mathbf{b}, \alpha, \beta, \delta) =
$$

$$
\frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 - \langle \delta, \sum_{k=1}^{Q} w_k \rangle - \beta \sum_{k=1}^{Q} b_k + \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \left(\langle w_k, \Phi(x_i) \rangle + b_k + \frac{1}{Q-1} \right). \quad (3)
$$

Setting the gradient of the Lagrangian function with respect to w_k equal to the null vector provides us with Q alternative expressions for the optimal value of vector δ:

$$
\delta^* = w_k^* + \sum_{i=1}^{m} \alpha_{ik}^* \Phi(x_i), \quad (1 \leq k \leq Q). \quad (4)
$$

Since by hypothesis, $\sum_{k=1}^{Q} w_k^* = 0$, summing over the index k provides us with the expression of δ^* as a function of dual variables only:

$$
\delta^* = \frac{1}{Q} \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \Phi(x_i). \quad (5)
$$

By substitution into (4), we get the expression of the vectors w_k at the optimum:

$$w_k^* = \frac{1}{Q}\sum_{i=1}^{m}\sum_{l=1}^{Q}\alpha_{il}^*\Phi(x_i) - \sum_{i=1}^{m}\alpha_{ik}^*\Phi(x_i), \quad (1 \leq k \leq Q)$$

which can also be written as

$$w_k^* = \sum_{i=1}^{m}\sum_{l=1}^{Q}\alpha_{il}^*\left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_i), \quad (1 \leq k \leq Q) \tag{6}$$

where δ is the Kronecker symbol.

Let us now set the gradient of (3) with respect to $\mathbf{b}$ equal to the null vector. It comes:

$$\beta^* = \sum_{i=1}^{m}\alpha_{ik}^*, \quad (1 \leq k \leq Q)$$

and thus

$$\sum_{i=1}^{m}\sum_{l=1}^{Q}\alpha_{il}^*\left(\frac{1}{Q} - \delta_{k,l}\right) = 0, \quad (1 \leq k \leq Q).$$

Given the constraint $\sum_{k=1}^{Q} b_k = 0$, this implies that:

$$\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^* b_k^* = \beta^* \sum_{k=1}^{Q} b_k^* = 0. \tag{7}$$

By application of (6),

$$\begin{aligned}\sum_{k=1}^{Q}\|w_k^*\|^2 &= \sum_{k=1}^{Q}\langle\sum_{i=1}^{m}\sum_{l=1}^{Q}\alpha_{il}^*\left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_i), \sum_{j=1}^{m}\sum_{n=1}^{Q}\alpha_{jn}^*\left(\frac{1}{Q} - \delta_{k,n}\right)\Phi(x_j)\rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{Q}\sum_{n=1}^{Q}\alpha_{il}^*\alpha_{jn}^*\langle\Phi(x_i), \Phi(x_j)\rangle\sum_{k=1}^{Q}\left(\frac{1}{Q} - \delta_{k,l}\right)\left(\frac{1}{Q} - \delta_{k,n}\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{Q}\sum_{n=1}^{Q}\alpha_{il}^*\alpha_{jn}^*\left(\delta_{l,n} - \frac{1}{Q}\right)\kappa(x_i, x_j). \end{aligned} \tag{8}$$

Still by application of (6),

$$\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^*\langle w_k^*, \Phi(x_i)\rangle = \sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^*\langle\sum_{j=1}^{m}\sum_{l=1}^{Q}\alpha_{jl}^*\left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_j), \Phi(x_i)\rangle$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik}^* \alpha_{jl}^* \left(\frac{1}{Q} - \delta_{k,l} \right) \kappa(x_i, x_j). \tag{9}$$

Combining (8) and (9) gives:

$$\frac{1}{2} \sum_{k=1}^{Q} \|w_k^*\|^2 + \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \langle w_k^*, \Phi(x_i) \rangle = -\frac{1}{2} \sum_{k=1}^{Q} \|w_k^*\|^2$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik}^* \alpha_{jl}^* \left(\delta_{k,l} - \frac{1}{Q} \right) \kappa(x_i, x_j). \tag{10}$$

In what follows, we use the notation e_n to designate the vector of $\mathbb{R}^n$ such that all its components are equal to e. Let H be the matrix of $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ of general term:

$$h_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q} \right) \kappa(x_i, x_j).$$

With these notations at hand, reporting (7) and (10) in (3) provides us with the algebraic expression of the Lagrangian function at the optimum:

$$L(\alpha^*) = -\frac{1}{2} \alpha^{*T} H \alpha^* + \frac{1}{Q-1} 1_{Qm}^T \alpha^*.$$

This eventually provides us with the Wolfe dual formulation of Problem 1:

Problem 3 (Hard margin M-SVM, dual formulation)

$$\max_{\alpha} J_{LLW,d}(\alpha)$$

$$s.t. \begin{cases} \alpha_{ik} \geq 0, \quad (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{i=1}^{m} \sum_{l=1}^{Q} \alpha_{il} \left(\frac{1}{Q} - \delta_{k,l} \right) = 0, \quad (1 \leq k \leq Q) \end{cases}$$

where

$$J_{LLW,d}(\alpha) = -\frac{1}{2} \alpha^T H \alpha + \frac{1}{Q-1} 1_{Qm}^T \alpha,$$

with the general term of the Hessian matrix H being

$$h_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q} \right) \kappa(x_i, x_j).$$

Let the couple $(\mathbf{w}^0, \mathbf{b}^0)$ denote the optimal solution of Problem 1 and equivalently, let $\alpha^0 = \left(\alpha_{ik}^0\right)_{1 \leq i \leq m, 1 \leq k \leq Q} \in \mathbb{R}_+^{Qm}$ be the optimal solution of Problem 3. According to (6), the expression of w_k^0 is then:

$$w_k^0 = \sum_{i=1}^{m} \sum_{l=1}^{Q} \alpha_{il}^0 \left(\frac{1}{Q} - \delta_{k,l} \right) \Phi(x_i).$$

2.4 Geometrical margins

From a geometrical point of view, the algorithms described above tend to construct a set of hyperplanes $\{(w_k, b_k) : 1 \leq k \leq Q\}$ that maximize globally the C_Q^2 *margins* between the differents categories. If these margins are defined as in the bi-class case, their analytical expression is more complex.

Definition 4 (Geometrical margins, Definition 7 in [7]) *Let us consider a Q-category M-SVM (a function of $\mathcal{H}$) classifying the examples of its training set $\{(x_i, y_i) : 1 \leq i \leq m\}$ without error. γ_{kl}, its* margin between categories k and l, *is defined as the smallest distance of a point either in k or l to the hyperplane separating those categories. Let us denote*

$$d_{M\text{-}SVM} = \min_{1 \leq k < l \leq Q} \left\{ \min \left[\min_{i: y_i = k} \left(h_k(x_i) - h_l(x_i)\right), \min_{j: y_j = l} \left(h_l(x_j) - h_k(x_j)\right) \right] \right\}$$

and for $1 \leq k < l \leq Q$, let $d_{M\text{-}SVM,kl}$ be:

$$d_{M\text{-}SVM,kl} = \frac{1}{d_{M\text{-}SVM}} \min \left[\min_{i: y_i = k} \left(h_k(x_i) - h_l(x_i) - d_{M\text{-}SVM}\right), \min_{j: y_j = l} \left(h_l(x_j) - h_k(x_j) - d_{M\text{-}SVM}\right) \right].$$

Then we have:

$$\gamma_{kl} = d_{M\text{-}SVM} \frac{1 + d_{M\text{-}SVM,kl}}{\|w_k - w_l\|}.$$

Given the constraints of Problem 1, the expression of $d_{\text{M-SVM}}$ corresponding to the M-SVM of Lee, Lin and Wahba is:

$$d_{\text{LLW}} = \frac{Q}{Q-1}.$$

Remark 1 *The values of the parameters $d_{M\text{-}SVM,kl}$ (or $d_{LLW,kl}$ in the case of interest) are known as soon as the pair $\left(\mathbf{w}^0, \mathbf{b}^0\right)$ is known.*

The connection between the geometrical margins and the penalizer of $J_{\text{M-SVM}}$ is given by the following equation:

$$\sum_{k<l} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2, \tag{11}$$

the proof of which can for instance be found in Chapter 2 of [7]. We introduce now a result needed in the proof of the master theorem of this report.

Proposition 3 *For the hard margin M-SVM of Lee, Lin and Wahba, we have:*

$$\frac{Q}{(Q-1)^2} \sum_{k<l} \left(\frac{1 + d_{LLW,kl}}{\gamma_{kl}} \right)^2 = \sum_{k=1}^{Q} \|w_k^0\|^2 = {\alpha^0}^T H \alpha^0 = \frac{1}{Q-1} 1_{Qm}^T \alpha^0.$$

Proof

- $\frac{Q}{(Q-1)^2} \sum_{k<l} \left(\frac{1+d_{\mathrm{LLW},kl}}{\gamma_{kl}} \right)^2 = \sum_{k=1}^{Q} \|w_k^0\|^2$

  This equation is a direct consequence of Definition 4 and Equation 11.

- $\sum_{k=1}^{Q} \|w_k^0\|^2 = {\alpha^0}^T H \alpha^0$

  This is a direct consequence of Equation 10 and the definition of matrix H.

- ${\alpha^0}^T H \alpha^0 = \frac{1}{Q-1} 1_{Qm}^T \alpha^0$

  One of the Kuhn-Tucker optimality conditions is:

  $$\alpha_{ik}^0 \left(\langle w_k^0, \Phi(x_i) \rangle + b_k^0 + \frac{1}{Q-1} \right) = 0, \ \ (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q),$$

  and thus:

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

  By application of (7), this simplifies into

  $$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^0 \langle w_k^0, \Phi(x_i) \rangle + \frac{1}{Q-1} 1_{Qm}^T \alpha^0 = 0.$$

  Since

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

  is a direct consequence of (10), this concludes the proof.

■

3 The M-SVM^2

3.1 Quadratic loss multi-class SVMs: motivation and principle

The M-SVMs presented in Section 2.2 share a common feature with the standard pattern recognition SVM: the contribution of the slack variables to their objective functions is linear. Let ξ be the vector of these variables. In the cases of the M-SVMs of Weston and Watkins and Lee, Lin and Wahba, we have $\xi = (\xi_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$ with $(\xi_{iy_i})_{1 \leq i \leq m} = 0_m$, and in the case of the model of Crammer and Singer, it is simply $\xi = (\xi_i)_{1 \leq i \leq m}$. In both cases, the contribution to the objective function is $C\|\xi\|_1$.

In the bi-class case, there exists a variant of the standard SVM which is known as the 2-*norm SVM* since for this machine, the empirical contribution to the objective function is $C\|\xi\|_2^2$. Its main advantage, underlined for instance in the Chapter 7 of [13], is that its training algorithm can be expressed, after an appropriate change of kernel, as the training algorithm of a hard margin machine. As a consequence, its leave-one-out error can be upper bounded thanks to the radius-margin bound.

Unfortunately, a naive extension of the 2-norm SVM to the multi-class case, resulting from substituting in the objective function of either of the three M-SVMs $\|\xi\|_1$ with $\|\xi\|_2^2$, does not preserve this property. Section 2.4.1.4 of [7] gives detailed explanations about that point. The strategy that we propose to exhibit interesting multi-class generalizations of the 2-norm SVM consists in studying the class of *quadratic loss M-SVMs*, i.e., the class of extensions of the M-SVMs such that the contribution of the slack variables is a quadratic form:

$$C\xi^T M\xi = C\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\sum_{l=1}^{Q} m_{ik,jl}\xi_{ik}\xi_{jl}$$

where $M = (m_{ik,jl})_{1 \leq i,j \leq m, 1 \leq k,l \leq Q}$ is a symmetric positive semidefinite matrix.

3.2 The M-SVM^2 as a multi-class generalization of the 2-norm SVM

In this section, we establish that the idea introduced above provides us with a solution to the problem of interest when the M-SVM used is the one of Lee, Lin and Wahba and the general term of the matrix M is $m_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q}\right)\delta_{i,j}$. The corresponding machine, named M-SVM^2, generalizes the 2-norm SVM to an arbitrary (but finite) number of categories.

Problem 4 (M-SVM^2)

$$\min_{\mathbf{w},\mathbf{b}} J_{M\text{-}SVM^2}(\mathbf{w},\mathbf{b})$$

$$s.t. \begin{cases} \langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}, \quad (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{k=1}^{Q} w_k = 0 \\ \sum_{k=1}^{Q} b_k = 0 \end{cases}$$

where

$$J_{M\text{-}SVM^2}(\mathbf{w},\mathbf{b}) = \frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\left(\delta_{k,l}-\frac{1}{Q}\right)\delta_{i,j}\xi_{ik}\xi_{jl}.$$

Note that as in the bi-class case, it is useless to introduce nonnegativity constraints for the slack variables. The Lagrangian function associated with Problem 4 is thus

$$L\left(\mathbf{w},\mathbf{b},\xi,\alpha,\beta,\delta\right) =$$

$$\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\xi^T M\xi - \langle\delta,\sum_{k=1}^{Q}w_k\rangle - \beta\sum_{k=1}^{Q}b_k$$

$$+\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}\left(\langle w_k,\Phi(x_i)\rangle + b_k + \frac{1}{Q-1} - \xi_{ik}\right). \tag{12}$$

Setting the gradient of L with respect to ξ equal to the null vector gives

$$2CM\xi^* = \alpha^* \tag{13}$$

which has for immediate consequence that

$$C{\xi^*}^T M\xi^* - {\alpha^*}^T\xi^* = -C{\xi^*}^T M\xi^*. \tag{14}$$

Using the same reasoning that we used to derive the objective function of Problem 3 and (14), at the optimum, (12) simplifies into:

$$L\left(\xi^*,\alpha^*\right) = -\frac{1}{2}{\alpha^*}^T H\alpha^* - C{\xi^*}^T M\xi^* + \frac{1}{Q-1}1_{Qm}^T\alpha^*. \tag{15}$$

Besides, using (13),

$$\alpha_{in}^*\alpha_{ip}^* = 4C^2\sum_{k=1}^{Q}\left(\delta_{k,n}-\frac{1}{Q}\right)\xi_{ik}^*\sum_{l=1}^{Q}\left(\delta_{l,p}-\frac{1}{Q}\right)\xi_{il}^*$$

and thus

$$\alpha_{in}^*\alpha_{ip}^* = 4C^2\sum_{k=1}^{Q}\sum_{l=1}^{Q}\left(\delta_{k,n}\delta_{l,p} - (\delta_{k,n}+\delta_{l,p})\frac{1}{Q} + \frac{1}{Q^2}\right)\xi_{ik}^*\xi_{il}^*.$$

By a double summation over n and p, we have:

$$\sum_{n=1}^{Q}\sum_{p=1}^{Q}\alpha_{in}^*\alpha_{ip}^*\left(\delta_{n,p}-\frac{1}{Q}\right) = 4C^2\sum_{k=1}^{Q}\sum_{l=1}^{Q}\xi_{ik}^*\xi_{il}^*\sum_{n=1}^{Q}\sum_{p=1}^{Q}\left(\delta_{k,n}\delta_{l,p} - (\delta_{k,n}+\delta_{l,p})\frac{1}{Q} + \frac{1}{Q^2}\right)\left(\delta_{n,p}-\frac{1}{Q}\right).$$

Since

$$\sum_{n=1}^{Q}\sum_{p=1}^{Q}\left(\delta_{k,n}\delta_{l,p}-(\delta_{k,n}+\delta_{l,p})\frac{1}{Q}+\frac{1}{Q^2}\right)\left(\delta_{n,p}-\frac{1}{Q}\right)=\delta_{k,l}-\frac{1}{Q},$$

this simplifies into

$$\sum_{n=1}^{Q}\sum_{p=1}^{Q}\alpha^*_{in}\alpha^*_{ip}\left(\delta_{n,p}-\frac{1}{Q}\right)=4C^2\sum_{k=1}^{Q}\sum_{l=1}^{Q}\left(\delta_{k,l}-\frac{1}{Q}\right)\xi^*_{ik}\xi^*_{il}.$$

Finally, a double summation over i and j implies that

$$\alpha^{*T}M\alpha^*=4C^2\xi^{*T}M\xi^*.$$

A substitution into (15) provides us with:

$$L\left(\alpha^*\right)=-\frac{1}{2}\alpha^{*T}\left(H+\frac{1}{2C}M\right)\alpha^*+\frac{1}{Q-1}1_{Qm}^T\alpha^*.$$

As in the case of the hard margin version of the M-SVM of Lee, Lin and Wahba, setting the gradient of (12) with respect to $\mathbf{b}$ equal to the null vector gives:

$$\sum_{i=1}^{m}\sum_{l=1}^{Q}\alpha^*_{il}\left(\frac{1}{Q}-\delta_{k,l}\right)=0,\quad(1\leq k\leq Q).$$

Putting things together, we obtain the following expression for the dual problem of Problem 4:

Problem 5 (M-SVM^2, dual formulation)

$$\max_{\alpha} J_{M\text{-}SVM^2,d}(\alpha)$$

$$s.t.\begin{cases}\alpha_{ik}\geq 0,\quad(1\leq i\leq m),(1\leq k\neq y_i\leq Q)\\ \sum_{i=1}^{m}\sum_{l=1}^{Q}\alpha_{il}\left(\frac{1}{Q}-\delta_{k,l}\right)=0,\quad(1\leq k\leq Q)\end{cases}$$

where

$$J_{M\text{-}SVM^2,d}(\alpha)=-\frac{1}{2}\alpha^T\left(H+\frac{1}{2C}M\right)\alpha+\frac{1}{Q-1}1_{Qm}^T\alpha.$$

Due to the definitions of the matrices H and M, this is precisely Problem 3 with the kernel κ replaced by a kernel κ' such that:

$$\kappa'(x_i,x_j)=\kappa(x_i,x_j)+\frac{1}{2C}\delta_{i,j},\quad(1\leq i,j\leq m).$$

When $Q=2$, the M-SVM of Lee, Lin and Wahba, like the two other ones, is equivalent to the standard bi-class SVM (see for instance [7]). Furthermore, in that case, we get $\xi^TM\xi=\frac{1}{2}\|\xi\|_2^2$. The M-SVM^2 is thus equivalent to the 2-norm SVM.

4 Multi-Class Radius-Margin Bound on the Leave-One-Out Error of the M-SVM^2

To begin with, we must recall Vapnik's initial bi-class theorem (see Chapter 10 of [15]), which is based on an intermediate result of central importance known as the "key lemma".

4.1 Bi-class radius-margin bound

Lemma 1 (Bi-class key lemma) *Let us consider a hard margin bi-class SVM on a domain $\mathcal{X}$. Suppose that it is trained on a set $d_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ of m couples of $\mathcal{X} \times \{-1, 1\}$ (the points of which it separates without error). Consider now the same machine, trained on $d_m \setminus \{(x_p, y_p)\}$. If it makes an error on (x_p, y_p), then the inequality*

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

holds, where $\mathcal{D}_m$ is the diameter of the smallest sphere containing the images by the feature map of the support vectors of the initial machine.

Theorem 1 (Bi-class radius-margin bound) *Let γ be the geometrical margin of the hard margin SVM defined in Lemma 1, when trained on d_m. Let also $\mathcal{L}_m$ be the number of errors resulting from applying a leave-one-out cross-validation procedure to this machine. We have:*

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

The multi-class radius-margin bound that we propose in this report is a direct generalization of the one proposed by Vapnik. The first step of the proof consists in establishing a "multi-class key lemma". This is the subject of the following subsection.

4.2 Multi-class key lemma

Lemma 2 (Multi-class key lemma) *Let us consider a Q-category hard margin M-SVM of Lee, Lin and Wahba on a domain $\mathcal{X}$. Let $d_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ be its training set. Consider now the same machine trained on $d_m \setminus \{(x_p, y_p)\}$. If it makes an error on (x_p, y_p), then the inequality*

$$\max_{k \in [\![1,Q]\!]} \alpha_{pk}^0 \geq \frac{1}{Q(Q-1)\mathcal{D}_m^2}$$

holds, where $\mathcal{D}_m$ is the diameter of the smallest sphere of the feature space containing the set $\{\Phi(x_i) : 1 \leq i \leq m\}$.

Proof Let $(\mathbf{w}^p, \mathbf{b}^p)$ be the couple characterizing the optimal hyperplanes when the machine is trained on $d_m \setminus \{(x_p, y_p)\}$. Let

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

be the corresponding vector of dual variables. α^p belongs to $\mathbb{R}_+^{Qm}$, with $\left(\alpha_{pk}^p\right)_{1 \leq k \leq Q} = 0_Q$. 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 3 under the additional constraint $(\alpha_{pk})_{1 \leq k \leq Q} = 0_Q$. Let us define two more vectors in $\mathbb{R}_+^{Qm}$, $\lambda^p = (\lambda_{ik}^p)_{1 \leq i \leq m, 1 \leq k \leq Q}$ and $\mu^p = (\mu_{ik}^p)_{1 \leq i \leq m, 1 \leq k \leq Q}$. λ^p satisfies additional properties so that the vector $\alpha^0 - \lambda^p$ is a feasible solution of Problem 3 under the additional constraint that $\left(\alpha_{pk}^0 - \lambda_{pk}^p\right)_{1 \leq k \leq Q} = 0_Q$, i.e., $\alpha^0 - \lambda^p$ satisfies the same constraints as α^p. We have

$$\forall i \neq p, \forall k \neq y_i, \;\; \alpha_{ik}^0 - \lambda_{ik}^p \geq 0 \Longleftrightarrow \lambda_{ik}^p \leq \alpha_{ik}^0.$$

We deduce from the equality constraints of Problem 3 that:

$$\forall k, \;\; \sum_{i=1}^m \sum_{l=1}^Q \left(\alpha_{il}^0 - \lambda_{il}^p\right) \left(\frac{1}{Q} - \delta_{k,l}\right) = 0 \Longleftrightarrow \sum_{i=1}^m \sum_{l=1}^Q \lambda_{il}^p \left(\frac{1}{Q} - \delta_{k,l}\right) = 0.$$

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

$$\begin{cases} \forall k, \;\; \lambda_{pk}^p = \alpha_{pk}^0 \\ \forall i \neq p, \forall k, \;\; 0 \leq \lambda_{ik}^p \leq \alpha_{ik}^0 \\ \sum_{i=1}^m \sum_{l=1}^Q \lambda_{il}^p \left(\frac{1}{Q} - \delta_{k,l}\right) = 0, \;\; (1 \leq k \leq Q) \end{cases}. \tag{16}$$

The properties of vector μ^p are such that $\alpha^p + K_1 \mu^p$ satisfies the constraints of the same problem, where K_1 is a positive scalar the value of which will be specified in the sequel. We have thus:

$$\forall i, \;\; \alpha_{iy_i}^p + K_1 \mu_{iy_i}^p = 0 \Longleftrightarrow \mu_{iy_i}^p = 0.$$

Moreover, we have

$$\forall i, \forall k \neq y_i, \;\; \mu_{ik}^p \geq 0 \Longrightarrow \alpha_{ik}^p + K_1 \mu_{ik}^p \geq 0.$$

Finally,

$$\sum_{i=1}^m \sum_{l=1}^Q \left(\alpha_{il}^p + c\mu_{il}^p\right) \left(\frac{1}{Q} - \delta_{k,l}\right) = 0 \Longleftrightarrow \sum_{i=1}^m \sum_{l=1}^Q \mu_{il}^p \left(\frac{1}{Q} - \delta_{k,l}\right) = 0.$$

To sum up, vector μ^p satisfies the following constraints:

$$\begin{cases} \forall i, \;\; \mu_{iy_i}^p = 0 \\ \forall i, \forall k \neq y_i, \;\; \mu_{ik}^p \geq 0 \\ \sum_{i=1}^m \sum_{l=1}^Q \mu_{il}^p \left(\frac{1}{Q} - \delta_{k,l}\right) = 0, \;\; (1 \leq k \leq Q) \end{cases}. \tag{17}$$

In the sequel, for the sake of simplicity, we write J in place of $J_{\mathrm{LLW,d}}$. By construction of vectors λ^p and μ^p, we have $J(\alpha^0 - \lambda^p) \leq J(\alpha^p)$ and $J\left(\alpha^p + K_1 \mu^p\right) \leq J(\alpha^0)$, and by way of consequence,

$$J(\alpha^0) - J(\alpha^0 - \lambda^p) \geq J(\alpha^0) - J(\alpha^p) \geq J\left(\alpha^p + K_1 \mu^p\right) - J(\alpha^p). \tag{18}$$

The expression of the first term is

$$J(\alpha^0) - J(\alpha^0 - \lambda^p) = \frac{1}{2}\lambda^{p\,T} H \lambda^p + \left(-H\alpha^0 + \frac{1}{Q-1}1_{Qm}\right)^T \lambda^p. \tag{19}$$

Given (6) and the definition of matrix H,

$$\begin{aligned}\left(-H\alpha^0 + \frac{1}{Q-1}1_{Qm}\right)^T \lambda^p &= \sum_{i=1}^m \sum_{k \neq y_i} \left(\langle w_k^0, \Phi(x_i)\rangle + \frac{1}{Q-1}\right) \lambda_{ik}^p \\ &= \sum_{i=1}^m \sum_{k \neq y_i} \left(h_k^0\left(x_i\right) + \frac{1}{Q-1}\right) \lambda_{ik}^p - \sum_{i=1}^m \sum_{k \neq y_i} b_k^0 \lambda_{ik}^p. \end{aligned} \tag{20}$$

Due to the constraints of correct classification and the nonnegativity of the components of vector λ^p, the first double sum of the right-hand side of (20) is nonpositive. Furthermore, making use of the equality constraints of (16) and $\sum_{k=1}^Q b_k^0 = 0$ gives:

$$\sum_{i=1}^m \sum_{k=1}^Q b_k^0 \lambda_{ik}^p = \sum_{k=1}^Q b_k^0 \sum_{i=1}^m \lambda_{ik}^p = \left(\sum_{k=1}^Q b_k^0\right)\left(\sum_{i=1}^m \sum_{l=1}^Q \frac{1}{Q}\lambda_{il}^p\right) = 0.$$

Thus,

$$\left(-H\alpha^0 + \frac{1}{Q-1}1_{Qm}\right)^T \lambda^p \leq 0.$$

A substitution into (19) provides us with the following upper bound on $J(\alpha^0) - J(\alpha^0 - \lambda^p)$:

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

and equivalently, by definition of H,

$$J(\alpha^0) - J(\alpha^0 - \lambda^p) \leq \frac{1}{2}\sum_{k=1}^Q \left\| \sum_{i=1}^m \sum_{l=1}^Q \lambda_{il}^p \left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_i)\right\|^2. \tag{21}$$

We now turn to the right-hand side of (18). The line of reasoning already used for the left-hand side gives:

$$J\left(\alpha^p + K_1\mu^p\right) - J(\alpha^p) =$$

$$K_1\left(-H\alpha^p + \frac{1}{Q-1}1_{Qm}\right)^T \mu^p - \frac{K_1^2}{2}\sum_{k=1}^Q \left\| \sum_{i=1}^m \sum_{l=1}^Q \mu_{il}^p \left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_i)\right\|^2 \tag{22}$$

with

$$\left(-H\alpha^p + \frac{1}{Q-1}1_{Qm}\right)^T \mu^p = \sum_{i=1}^m \sum_{k \neq y_i} \left(h_k^p\left(x_i\right) + \frac{1}{Q-1}\right) \mu_{ik}^p. \tag{23}$$

By hypothesis, the M-SVM trained on $d_m \setminus \{(x_p, y_p)\}$ does not classify x_p correctly. This means that there exists $n \in [\![1, Q]\!] \setminus \{y_p\}$ such that $h_n^p(x_p) \geq 0$. Let $\mathcal{I}$ be a mapping from $[\![1, Q]\!] \setminus \{n\}$ to $[\![1, m]\!] \setminus \{p\}$ such that

$$\forall k \in [\![1, Q]\!] \setminus \{n\}, \;\; \alpha^p_{\mathcal{I}(k)n} > 0.$$

We know that such a mapping exists, otherwise, given the equality constraints of Problem 3, vector α^p would be equal to the null vector. For $K_2 \in \mathbb{R}_+^*$, let μ^p be the vector of $\mathbb{R}^{Qm}$ that only differs from the null vector in the following way:

$$\begin{cases} \mu^p_{pn} = K_2 \\ \forall k \in [\![1, Q]\!] \setminus \{n\}, \;\; \mu^p_{\mathcal{I}(k)k} = K_2 \end{cases} .$$

Obviously, this solution is feasible (satisfies the constraints 17). Indeed, $\frac{1}{Q}\sum_{i=1}^m \sum_{k=1}^Q \mu^p_{ik} = K_2$ and $\sum_{i=1}^m \mu^p_{ik} = K_2$, $(1 \leq k \leq Q)$. With this definition of vector μ^p, the right-hand side of (23) simplifies into:

$$K_2 \left(h_n^p(x_p) + \sum_{k \neq n} h_k^p\left(x_{\mathcal{I}(k)}\right) + \frac{Q}{Q-1} \right).$$

Vector μ^p has been specified so as to make it possible to exhibit a nontrivial lower bound on this last expression. By definition of n, $h_n^p(x_p) \geq 0$. Furthermore, the Kuhn-Tucker optimality conditions:

$$\alpha^p_{ik}\left(\langle w_k^p, \Phi(x_i) \rangle + b_k^p + \frac{1}{Q-1} \right) = 0, \;\; (1 \leq i \neq p \leq m), (1 \leq k \neq y_i \leq Q)$$

imply that $\left(h_k^p\left(x_{\mathcal{I}(k)}\right)\right)_{1 \leq k \neq n \leq Q} = -\frac{1}{Q-1}1_{Q-1}$. As a consequence, a lower bound on the right-hand side of (23) is provided by:

$$\sum_{i=1}^m \sum_{k \neq y_i} \left(h_k^p(x_i) + \frac{1}{Q-1} \right) \mu^p_{ik} \geq \frac{K_2}{Q-1}.$$

It springs from this bound and (22) that

$$J\left(\alpha^p + K_1\mu^p\right) - J(\alpha^p) \geq \frac{K_1K_2}{Q-1} - \frac{K_1^2}{2}\sum_{k=1}^Q \left\| \sum_{i=1}^m \sum_{l=1}^Q \mu^p_{il}\left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_i) \right\|^2 . \qquad (24)$$

Combining (18), (21) and (24) finally gives:

$$\frac{1}{2}\sum_{k=1}^Q \left\| \sum_{i=1}^m \sum_{l=1}^Q \lambda^p_{il}\left(\frac{1}{Q} - \delta_{k,l}\right)\Phi(x_i) \right\|^2 \geq$$

$$\frac{K_1 K_2}{Q-1} - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \mu_{il}^p \left(\frac{1}{Q} - \delta_{k,l} \right) \Phi(x_i) \right\|^2 . \tag{25}$$

Let $\nu^p = (\nu_{ik}^p)_{1 \leq i \leq m, 1 \leq k \leq Q}$ be the vector of $\mathbb{R}_+^{Qm}$ such that $\mu^p = K_2 \nu^p$. The value of the scalar $K_3 = K_1 K_2$ maximizing the right-hand side of (25) is:

$$K_3^* = \frac{\frac{1}{Q-1}}{\sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \nu_{il}^p \left(\frac{1}{Q} - \delta_{k,l} \right) \Phi(x_i) \right\|^2}.$$

By substitution in (25), this means that:

$$(Q-1)^2 \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \lambda_{il}^p \left(\frac{1}{Q} - \delta_{k,l} \right) \Phi(x_i) \right\|^2 \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \nu_{il}^p \left(\frac{1}{Q} - \delta_{k,l} \right) \Phi(x_i) \right\|^2 \geq 1.$$

For η in $\mathbb{R}^{Qm}$, let $K(\eta) = \frac{1}{Q} \sum_{i=1}^{m} \sum_{k=1}^{Q} \eta_{ik}^p$. We have:

$$\left\| \frac{1}{Q} \sum_{i=1}^{m} \sum_{l=1}^{Q} \lambda_{il}^p \Phi(x_i) - \sum_{i=1}^{m} \lambda_{ik}^p \Phi(x_i) \right\|^2 = K\left(\lambda^p\right)^2 \left\| \mathrm{conv}_1(\Phi(x_i)) - \mathrm{conv}_2(\Phi(x_i)) \right\|^2$$

where $\mathrm{conv}_1(\Phi(x_i))$ and $\mathrm{conv}_2(\Phi(x_i))$ are two convex combinations of the $\Phi(x_i)$. As a consequence, $\left\| \mathrm{conv}_1(\Phi(x_i)) - \mathrm{conv}_2(\Phi(x_i)) \right\|^2$ can be bounded from above by $\mathcal{D}_m^2$. Since the same reasoning applies to ν^p, we get:

$$(Q-1)^2 Q^2 K\left(\lambda^p\right)^2 K\left(\nu^p\right)^2 \mathcal{D}_m^4 \geq 1. \tag{26}$$

By construction, $K\left(\nu^p\right) = 1$. We now construct a vector λ^p minimizing the objective function K. First, note that due to the equality constraints satisfied by this vector,

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^{m} \lambda_{ik}^p = \frac{1}{Q} \sum_{i=1}^{m} \sum_{l=1}^{Q} \lambda_{il}^p.$$

As a consequence,

$$\forall (k,l) \in [\![1, Q]\!]^2, \;\; \sum_{i=1}^{m} \lambda_{ik}^p = \sum_{i=1}^{m} \lambda_{il}^p.$$

This implies that:

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^{m} \lambda_{ik}^p \geq \max_{l \in [\![1, Q]\!]} \alpha_{pl}^0.$$

Obviously, both the box constraints in (16) and the nature of K call for the choice of small values for the components λ_{ik}^{p}. Thus, there is a feasible solution λ^{p*} such that:

$$\forall k \in [\![1, Q]\!], \quad \sum_{i=1}^{m} \lambda_{ik}^{p\;*} = \max_{l \in [\![1,Q]\!]} \alpha_{pl}^{0}.$$

This solution is such that $K\left(\lambda^{p*}\right) = \max_{k \in [\![1,Q]\!]} \alpha_{pk}^{0}$. The substitution of the values of $K\left(\nu^{p}\right)$ and $K\left(\lambda^{p*}\right)$ in (26) provides us with:

$$\left(\max_{k \in [\![1,Q]\!]} \alpha_{pk}^{0}\right)^{2} \geq \frac{1}{(Q-1)^{2} Q^{2} \mathcal{D}_{m}^{4}}.$$

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

4.3 Multi-class radius-margin bound

Theorem 2 (Multi-class radius-margin bound) *Let us consider a Q-category hard margin M-SVM of Lee, Lin and Wahba on a domain $\mathcal{X}$. Let $d_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ be its training set, $\mathcal{L}_m$ the number of errors resulting from applying a leave-one-out cross-validation procedure to this machine, and $\mathcal{D}_m$ the diameter of the smallest sphere of the feature space containing the set $\{\Phi(x_i) : 1 \leq i \leq m\}$. Then the following upper bound holds true:*

$$\mathcal{L}_m \leq Q^{2} \mathcal{D}_{m}^{2} \sum_{k<l} \left(\frac{1 + d_{LLW,kl}}{\gamma_{kl}}\right)^{2}.$$

Proof Lemma 2 exhibits a non trivial lower bound on $\max_{k \in [\![1,Q]\!]} \alpha_{pk}^{0}$ when the machine trained on the set $d_m \setminus \{(x_p, y_p)\}$ makes an error on (x_p, y_p), i.e., when (x_p, y_p) contributes to $\mathcal{L}_m$. As a consequence,

$$1_{Qm}^{T} \alpha^{0} \geq \sum_{i=1}^{m} \max_{k \in [\![1,Q]\!]} \alpha_{ik}^{0} \geq \frac{\mathcal{L}_m}{Q(Q-1)\mathcal{D}_{m}^{2}}. \tag{27}$$

According to Proposition 3, $1_{Qm}^{T} \alpha^{0} = \frac{Q}{Q-1} \sum_{k<l} \left(\frac{1+d_{\mathrm{LLW},kl}}{\gamma_{kl}}\right)^{2}$. A substitution in (27) thus provides us with the result announced. ■

5 Conclusions and Future Work

In this report, we have introduced a variant of the M-SVM of Lee, Lin and Wahba that strictly generalizes to the multi-class case the 2-norm SVM. For this quadratic loss M-SVM, named M-SVM^2, we have then established a generalization of Vapnik's radius-margin bound. We conjecture that this bound could be improved by a Q^2 factor. As it is, it can already be compared with those proposed in [18] for model selection. This, with a general study of the quadratic loss M-SVMs, is the subject of an ongoing research.

Acknowledgements

The work of E. Monfrini was supported by the Decrypthon program of the "Association Française contre les Myopathies" (AFM), the CNRS and IBM.

Contents

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers, Boston, 2004.

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

[3] O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

[4] C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[5] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[6] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, Chichester, second edition, 1987.

[7] Y. Guermeur. *SVM multiclasses, théorie et applications.* Habilitation à diriger des recherches, UHP, 2007. (in French).

[8] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[10] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

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

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

[13] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, 2004.

[14] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[15] V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.

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

[17] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 6, pages 69–88. The MIT Press, Cambridge, MA, 1999.

[18] 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.

[19] 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.

[20] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.