Radius-Margin Bound on the Leave-One-Out Error of the LLW-M-SVM

Yann Guermeur[1] , Emmanuel Monfrini[2]

[1] LORIA-CNRS
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy cedex, France
E-mail: Yann.Guermeur@loria.fr
[2] TELECOM & Management SudParis
9 rue Charles Fourier
91011 Evry cedex, France
E-mail: Emmanuel.Monfrini@it-sudparis.eu

Abstract. To set the values of the hyperparameters of a support vector machine (SVM), one can use cross-validation. Its leave-one-out variant produces 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. The most popular one is the radius-margin bound. In this article, we introduce a generalized radius-margin bound dedicated to the multi-class SVM of Lee, Lin and Wahba.
Keywords: M-SVMs, model selection, leave-one-out error, radius-margin bound.

1 Introduction

Using a SVM [9] requires to set the values of two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. Several approaches are available to perform this model selection task (see for instance [6]). The solution of choice consists in applying a cross-validation procedure. The leave-one-out one presents the advantage to produce an estimator of the generalization error which is almost unbiased [9], and the drawback to be highly time consuming. Consequently, in recent years, a number of upper bounds on the leave-one-out error of the pattern recognition SVM have been proposed in literature (see [2] for a survey). Although the tightest one is the *span bound* [10], the results of Chapelle and co-workers [2] show that another bound, the *radius-margin* one [9], achieves equivalent performance for model selection while being far simpler to compute. This is the reason why it is currently the most popular bound. In this article, we introduce a generalized radius-margin bound on the leave-one-out error of the hard margin version of the multi-class SVM (M-SVM) of Lee, Lin and Wahba [5]. For lack of space, its proof is omitted. It can be found in the accompanying research report [4].

The organization of this paper is as follows. Section 2 offers a general introduction to the M-SVMs. Section 3 focuses on the M-SVM of Lee, Lin and Wahba (LLW-M-SVM). Section 4 is devoted to the formulation 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

Like the SVMs, the M-SVMs are *large margin classifiers* which are devised in the framework of Vapnik's statistical learning theory [9].

2.1 Formalization of the learning problem

We consider Q-category classification problems with $3 \leq Q < \infty$. An object is represented by its description $x \in \mathcal{X}$ and the set of categories $\mathcal{Y}$ can be identified with the set $[\![1, Q]\!]$. We assume that the link between objects and categories can be described by an unknown probability measure P on $\mathcal{X} \times \mathcal{Y}$. The learning problem consists in selecting in a class $\mathcal{G}$ 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. 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 $\mathcal{X}$ to $\mathcal{Y} \bigcup \{*\}$) associated with g. Ideally, the objective function to be minimized over $\mathcal{G}$ is the probability of error $P(f(X) \neq Y)$. In practice, since P is unknown, other criteria are used and the optimization process is based on empirical data. We assume that there exists a random pair (X, 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). Such learning problems raise two questions: how to choose $\mathcal{G}$ and how to determine the best candidate g^* in this class, using only D_m. We focus on the first one, named *model selection*, when the model considered is a M-SVM.

2.2 Architecture and training algorithms

M-SVMs, like all the SVMs, are *kernel machines* [7]. They operate on a class of functions spanned by a positive semidefinite function/kernel.

Definition 1 (Positive semidefinite function). A real-valued function κ on $\mathcal{X}^2$ is called a *positive semidefinite function* if it is symmetric and

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

Definition 2 (Reproducing kernel Hilbert space [1]). Let $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ be a Hilbert space of real-valued functions on $\mathcal{X}$. A real-valued function κ on $\mathcal{X}^2$ is a *reproducing kernel* of $\mathbf{H}$ if and only if

1. $\forall x \in \mathcal{X},\ \kappa_x = \kappa(x, \cdot) \in \mathbf{H}$;
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 real-valued functions which possesses a reproducing kernel is called a *reproducing kernel Hilbert space* (RKHS) or a *proper Hilbert space*.

The connection between positive semidefinite functions and RKHSs is provided by the Moore-Aronszajn theorem.

Theorem 1 (Moore-Aronszajn theorem [1]). *Let κ be a real-valued positive semidefinite function on $\mathcal{X}^2$. There exists only one Hilbert space $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ of real-valued functions on $\mathcal{X}$ with κ as reproducing kernel.*

Proposition 1. *Let κ be a real-valued positive semidefinite function on $\mathcal{X}^2$. There exists a map Φ from $\mathcal{X}$ into a Hilbert space $\left(E_{\Phi(\mathcal{X})}, \langle\cdot,\cdot\rangle\right)$ such that*

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

Let κ be a kernel on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ be the RKHS spanned by κ. Let $\bar{\mathcal{H}} = \mathbf{H}_\kappa^Q$ and $\mathcal{H} = \left(\mathbf{H}_\kappa + \{1\}\right)^Q$. $\mathcal{H}$ is the class of functions $h = (h_k)_{1 \leq k \leq Q}$ on $\mathcal{X}$ whose component functions are finite affine combinations of the form

$$h_k(\cdot) = \sum_{i=1}^{m_k} \beta_{ik} \kappa\left(x_{ik}, \cdot\right) + b_k,$$

as well as the limits of these functions as the sets $\{x_{ik} : 1 \leq i \leq m_k\}$ become dense in $\mathcal{X}$, in the norm induced by the inner product $\langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}$. Due to (1), $\mathcal{H}$ is also a multivariate affine model on $\Phi\left(\mathcal{X}\right)$. Thus, h can be rewritten as

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

where the vectors w_k belong to $E_{\Phi(\mathcal{X})}$. It is then 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 $\mathbf{b} = (b_k)_{1 \leq k \leq Q} \in \mathbb{R}^Q$. $\bar{\mathcal{H}}$ is a multivariate linear model on $\Phi\left(\mathcal{X}\right)$, endowed with a norm $\|\cdot\|_{\bar{\mathcal{H}}}$ given by

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

A generic definition of the M-SVMs can then be formulated as follows.

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

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

where the loss function $\ell_{\text{M-SVM}}$ is nonnegative and convex. If a M-SVM is trained subject to the constraint $\sum_{i=1}^{m} \ell_{\text{M-SVM}}\left(y_i, h\left(x_i\right)\right) = 0$, it is called a *hard margin* M-SVM. Otherwise, it is called a *soft margin* M-SVM.

2.3 Geometrical margins

The algorithms following Definition 3 select functions h^* that tend to maximize globally the $\binom{Q}{2}$ *geometrical margins* between the different categories.

Definition 4 (Geometrical margins). Let $d_n = \{(x_i, y_i) : 1 \leq i \leq n\}$ be a set of n examples (belonging to $\mathcal{X} \times \mathcal{Y}$). If $h \in \mathcal{H}$ classifies these examples without error, then its *margin between categories k and l* (computed with respect to d_n), $\gamma_{kl}(h)$, is defined as the smallest distance of a point of d_n either in k or l to the hyperplane separating those categories. Let us denote

$$d(h) = \min_{1 \leq k < l \leq Q} \left\{ \min_{i : y_i \in \{k,l\}} |h_k(x_i) - h_l(x_i)| \right\}$$

and for $1 \leq k < l \leq Q$, let $d_{kl}(h)$ be

$$d_{kl}(h) = \frac{1}{d(h)} \min_{i : y_i \in \{k,l\}} |h_k(x_i) - h_l(x_i)| - 1.$$

Then we have

$$\gamma_{kl}(h) = d(h) \frac{1 + d_{kl}(h)}{\|w_k - w_l\|}.$$

For the M-SVMs, the connection between the geometrical margins and the penalizer of $J_{\text{M-SVM}}$ is given by $\sum_{k<l} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2$.

3 The M-SVM of Lee, Lin and Wahba

Several models of M-SVMs can be found in literature (see [4] for a survey). The M-SVM of Lee, Lin and Wahba [5] is the only one which is Fisher consistent. It corresponds to the loss function ℓ_{LLW} given by

$$\ell_{\text{LLW}}(y, h(x)) = \sum_{k \neq y} \max \left\{ h_k(x) + \frac{1}{Q-1}, 0 \right\}.$$

The substitution in Definition 3 of $\ell_{\text{M-SVM}}$ with ℓ_{LLW} provides us with the expressions of the quadratic programming problems corresponding to the training algorithms of the two versions of the machine.

Problem 1 (Hard margin LLW-M-SVM, primal formulation).

$$\min_{\mathbf{w},\mathbf{b}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 \right\}$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \langle w_k, \Phi(x_i) \rangle + b_k \leq -\frac{1}{Q-1} \\ \sum_{k=1}^{Q} w_k = 0, \;\; \sum_{k=1}^{Q} b_k = 0 \end{cases}.$$

Problem 2 (Soft margin LLW-M-SVM, primal formulation).

$$\min_{\mathbf{w},\mathbf{b},\xi} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k \neq y_i} \xi_{ik} \right\}$$

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

In Problem 2, the ξ_{ik} are *slack variables* used to relax the constraints of good classification. C is used to control the trade-off between prediction accuracy on d_m and smoothness of h^*. Instead of directly solving Problems 1 and 2, one usually solves their dual. Let α_{ik} be the Lagrange multiplier corresponding to the constraint $\langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1}$ or $\langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}$. Let α be the vector of $\mathbb{R}_+^{Qm}$ such that its component of index $(i-1)Q+k$ is equal to α_{ik} if $k \neq y_i$, and to 0 otherwise. The dual of Problem 1 is:

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

$$\max_{\alpha} \left\{ -\frac{1}{2}\alpha^T H \alpha + \frac{1}{Q-1} 1_{Qm}^T \alpha \right\}$$

$$s.t. \begin{cases} \forall i \in [\![1,m]\!], \forall k \in [\![1,Q]\!] \setminus \{y_i\}, \ \ \alpha_{ik} \geq 0 \\ \forall k \in [\![1,Q-1]\!], \ \ \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l}\right) \alpha_{il} = 0 \end{cases}$$

with the general term of the Hessian matrix H being

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

The Wolfe dual of Problem 2 only differs from Problem 3 in the inequality constraints. The constraints $\alpha_{ik} \geq 0$ are replaced by $0 \leq \alpha_{ik} \leq C$.

4 Multi-Class Radius-Margin Bound

The proof of our radius-margin bound rests on two partial results.

Lemma 1 (Multi-class key lemma). *Let us consider a hard margin Q-category LLW-M-SVM on $\mathcal{X}$. Let $d_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ be its training set and α^* its vector of dual variables. 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*

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

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

The value of $\mathcal{D}_m$ is obtained by solving a quadratic programming problem.

Proposition 2. *For the hard margin Q-category LLW-M-SVM,*

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

Theorem 2 (Multi-class radius-margin bound). *Let us consider the LLW-M-SVM of Lemma 1. Let $\mathcal{L}_m$ be the number of errors resulting from applying a leave-one-out cross-validation procedure to this machine. Then, using the notations of Definition 4, the following upper bound holds true:*

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

5 Conclusions and Ongoing Research

We have introduced a generalization of Vapnik's radius-margin bound dedicated to the M-SVM of Lee, Lin and Wahba. In doing so, we have highlighted different features of the M-SVMs which make their study intrinsically more difficult than the one of bi-class SVMs. For instance, the formula expressing the geometrical margins as a function of the vector α^* is far more complicated than its bi-class counterpart. This work, which comes after the derivation of guaranteed risks for classifiers taking values in $\mathbb{R}^Q$ [3], thus provides us with new arguments suggesting that the study of multi-category classification should be tackled independently of the one of dichotomy computation.

An open question is the possibility to use our bound to set the value of C. It can be reformulated as follows: is there a variant of the soft margin LLW-M-SVM such that its training algorithm is equivalent to the training algorithm of a hard margin machine obtained by a simple change of kernel? In the bi-class case, the answer is positive, and the corresponding variant is the 2-norm SVM (see for instance Chapter 7 in [8]). Finding the answer in the multi-class case is the subject of an ongoing research, as well as the derivation of radius-margin bounds suitable for the other M-SVMs.

References

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

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

3.Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics*, 2009. (to appear).

4.Y. Guermeur and E. Monfrini. Radius-margin bound on the leave-one-out error of the LLW-M-SVM. Technical report, LORIA, http://www.loria.fr/~guermeur/GueMon09_long.pdf, 2009.

5.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.

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

7.B. Schölkopf and A.J. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

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

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

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