Bound on the Risk for M-SVMs

Yann Guermeur
LORIA-UHP
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy cedex, France
Yann.Guermeur@loria.fr

André Elisseeff
MPI for Biological Cybernetics
Spemannstraße 38
72076 Tübingen, Deutschland
andre.elisseeff@tuebingen.mpg.de

Dominique Zelus
Wiener Lab, CIBIO
Av Presidente Juan D. Perón 2991
2000 Rosario, Argentina
dzelus@wiener-lab.com.ar

Résumé

La conception d'architectures SVM dédiées aux tâches de discrimination à catégories multiples constitue actuellement l'un des problèmes centraux de la théorie statistique de l'apprentissage. Plusieurs contributions majeures au domaines ont été récemment recencées. Cependant, aucun argument définitif n'a à ce jour été avancé permettant d'établir laquelle des deux approches disponibles, décomposition du problème en un ensemble de dichotomies ou calcul direct de la polychotomie, devait être supérieure à l'autre. L'objet de cette communication est de proposer un cadre théorique simple permettant une étude comparée des capacités de généralisation des différents modèles existants et la conception de nouvelles machines. Elle devrait par là même contribuer à résoudre la question de l'approche la plus adéquate.

Keywords: multi-class SVMs, extended VC dimensions, guaranteed risk

1 Introduction

From a theoretical point of view, the most appealing property of SVMs [12] is the fact that they represent straightforward implementations of Vapnik's structural risk minimization (SRM) inductive principle [12]. Loosely speaking, starting from a uniform convergence result extending the famous Glivenko-Cantelli theorem, one ends up with a learning machine that is easy to implement and generalizes well. Unfortunately, this is no longer the case for multi-class discriminant analysis. This topic was initially tackled through decomposition schemes [11, 12]. The first multi-class SVM (M-SVM) published was proposed independently by different teams (see for instance [13, 12]). These studies, although theoretically well grounded, were not involving a specific expression of the "guaranteed risk". Recently, two new models became available. The first one, described in [4], uses an original expression of the empirical risk. The bound on the generalization error provided is directly borrowed from a tree-based decomposition approach called DAGSVM [10]. In [9], the machine is devised to asymptotically implement the Bayes rule. To sum up, numerous solutions, based on different principles, are currently available to compute polychotomies with SVMs. They are still to be compared in the very framework to which they belong, i.e. with respect to their guaranteed risk. This should be all the more helpful since comparisons based on empirical results have failed to highlight significant differences in performance [10, 4, 6].

Building upon an original uniform convergence result, we introduce a simple framework to compute guaranteed risks for multi-class discriminant models and apply it to the architecture of M-SVMs. The organization of the paper is as follows. Section 2 summarizes our bound on the risk of multi-class models. The capacity measure of interest, a covering number, is related to a new generalization of the VC dimension. This *M-fat-shattering dimension* is bounded in Section 3 in the case of a multivariate affine model, i.e. of a M-SVM.

2 Guaranteed risk for multi-class discriminant models

We consider the case of a Q-category pattern recognition problem, where $Q \geq 3$. Let $\mathcal{X}$ be the space of description and $\mathcal{C}$ the set of categories. There is assumed to be a joint probability distribution P, fixed but unknown, on $\mathcal{X} \times \mathcal{C}$. Our goal is to find, in a given set $\mathcal{H}$ of functions $h = [h_k]$ from $\mathcal{X}$ into $\mathbb{R}^Q$, a function with lowest risk. The risk of a function h is the risk of the corresponding discriminant function, obtained by assigning each pattern x of $\mathcal{X}$ to the category C_k in $\mathcal{C}$ satisfying: $h_k(x) = \max_l h_l(x)$. Hereafter, $C(x_i)$ will denote the category of pattern x_i. The uniform convergence result we established is based on the following definitions.

Definition 1 (Expected risk) *The* expected risk *of a function f from $\mathcal{X}$ into $\mathcal{C}$ is its* generalization error *given by:*

$$R(f) = \int_{\mathcal{X}\times\mathcal{C}} \mathbf{I}_{\{f(x)\neq C\}} dP(x, C)$$

where $\mathbf{I}$ *is the indicator function.*

Definition 2 (Empirical risk) *Let $s_m = \{(x_i, C(x_i))\} \in (\mathcal{X} \times \mathcal{C})^m$ be a set of examples i.i.d. according to P. The* empirical risk *of f on s_m is defined as:*

$$R_{s_m}(f) = 1/m \ \left|\{(x_i, C(x_i)) \in s_m \ / \ f(x_i) \neq C(x_i)\}\right|$$

Definition 3 (ϵ-cover) *Let (E, ρ) be a pseudo-metric space, and $B(v, r)$ the closed ball of centre v and radius r in E. An* ϵ-cover *of $H \subset E$ is a subset $\bar{H}$ of E such that $H \subset \bigcup_{v\in\bar{H}} B(v, \epsilon)$.*

Definition 4 (Covering number) *Let (E, ρ) be a pseudo-metric space. If $H \subset E$ has an ϵ-cover of finite cardinality, then its* covering number *$\mathcal{N}(\epsilon, H, \rho)$ is the smallest cardinality of its ϵ-covers. If there is no such finite cover, then the covering number is defined to be ∞.*

Definition 5 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, define the pseudo-metric $d_{l_\infty, l_\infty(s)}$ on $\mathcal{H}$ as: $\forall (h, \bar{h}) \in \mathcal{H}^2$, $d_{l_\infty, l_\infty(s)}(h, \bar{h}) = \max_{x\in s} \max_{k\in\{1,\dots,Q\}} \left|h_k(x) - \bar{h}_k(x)\right|$.*

Definition 6 *For all $h \in \mathcal{H}$ and all $x \in \mathcal{X}$, let $M_1(h, x)$ be the smallest index l such that $h_l(x) = \max_k h_k(x)$ and $M_2(h, x)$ the smallest index $l \neq M_1(h, x)$ such that $h_l(x) = \max_{k\neq M_1(h,x)} h_k(x)$. Define $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, as the function from $\mathcal{X}$ into $\mathbb{R}^Q$, satisfying*

$$\Delta h_k(x) = \begin{cases} 1/2 \left(h_k(x) - h_{M_2(h,x)}(x)\right) & if\ k = M_1(h, x) \\ 1/2 \left(h_k(x) - h_{M_1(h,x)}(x)\right) & otherwise \end{cases}$$

Definition 7 (Empirical margin risk) *The* empirical risk with margin $\gamma \in (0, 1]$ *of h on s_m is*

$$R^{\gamma}_{s_m}(h) = 1/m \ \left|\left\{(x_i, C(x_i)) \in s_m \ / \ \Delta h_{C(x_i)}(x_i) < \gamma\right\}\right|$$

For $\gamma \in (0, 1]$, let $\pi_\gamma : \mathbb{R} \to [-\gamma, \gamma]$ be the piecewise-linear squashing function defined as

$$\pi_\gamma(x) = \begin{cases} \gamma.sign(x) & if\ |x| \geq \gamma \\ x & otherwise \end{cases}$$

$\forall h \in \mathcal{H}$, $\Delta h^\gamma = [\Delta h^\gamma_k] = [\pi_\gamma \circ \Delta h_k]$, $(1 \leq k \leq Q)$. $\Delta\mathcal{H}^\gamma = \{\Delta h^\gamma \ / \ h \in \mathcal{H}\}$. Let $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, m) = \max_{s_m\in\mathcal{X}^m} \mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty, l_\infty(s_m)})$. With these definitions at hand, extending Lemma 4 from [1], as well as the basic lemma of Theorem 4.1 in [12], the following theorem was established in [7]:

Theorem 1 *Let s_m be a m-sample of examples drawn independently from P. With probability at least $1 - \delta$, for every value of γ in $(0, 1]$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant model $\mathcal{H}$ is bounded above by:*

$$R(h) \leq R^{\gamma}_{s_m}(h) + \sqrt{\frac{1}{2m}\left(\ln\left(2\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m} \tag{1}$$

In view of this uniform convergence result, in order to implement the SRM inductive principle for a given multi-class discriminant model taking its values in $\mathbb{R}^Q$, it suffices to bound its covering number $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m)$. To do so, we adopt the standard strategy consisting in relating it to an extended notion of VC dimension through a generalized Sauer's lemma. Our capacity measure, the M-fat-shattering dimension, can be seen either as an extension of the fat-shattering dimension to the multivariate case, or as a scale-sensitive version of the graph dimension.

Definition 8 (Fat-shattering dimension [8]) *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* γ-shattered *by $\mathcal{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$(h_y(x_i) - b_i)\, y_i \geq \gamma, \ (1 \leq i \leq m)$$

The fat-shattering dimension *$fat_{\mathcal{H}}$ of the set $\mathcal{H}$ is a function from the positive real numbers to the integers which maps a value γ to the size of the largest set γ-shattered by functions of $\mathcal{H}$, if this size is finite, or to infinity otherwise.*

Definition 9 (Graph dimension [5]) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in a countable set. For any $h \in \mathcal{H}$, the* graph *$\mathcal{G}$ of h is $\mathcal{G}(h) = \{(x, h(x)) \ / \ x \in \mathcal{X}\}$ and the* graph space *of $\mathcal{H}$ is $\mathcal{G}(\mathcal{H}) = \{\mathcal{G}(h) \ / \ h \in \mathcal{H}\}$. Then the* graph dimension *of $\mathcal{H}$ is defined to be the VC dimension of the space $\mathcal{G}(\mathcal{H})$.*

Definition 10 (M-fat-shattering dimension [7]) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be M-γ-*shattered *by $\mathcal{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ and a vector $v_c = [c_i] \in \{1, \ldots, Q\}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y = [h_{yk}]$, $(1 \leq k \leq Q) \in \mathcal{H}$ satisfying*

$$(h_{yc_i}(x_i) - b_i)\, y_i \geq \gamma, \ (1 \leq i \leq m)$$

*The M-*fat-shattering dimension *M-$fat_{\mathcal{H}}$ of the set $\mathcal{H}$ is a function from the positive real numbers to the integers which maps a value γ to the size of the largest set M-γ-shattered by functions of $\mathcal{H}$, if this size is finite, or to infinity otherwise.*

We established in [7] that the covering number appearing in (1) is related to the M-fat-shattering dimension of $\Delta\mathcal{H}^\gamma$ through the following result:

Theorem 2 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$ and $d = M$-$fat_{\Delta\mathcal{H}^\gamma}(\gamma/8)$. For every value of γ in $(0, 1]$ and every value of m satisfying $d < 2m$, the following bound is true:*

$$\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \leq 2\left(2mQ9^Q\right)^{d \log_2(18emQ/d)} \tag{2}$$

3 M-fat-shattering dimension of M-SVMs

3.1 Characterization of the architecture

The machines considered here are obtained by combining a multivariate affine model with a nonlinear mapping. Formally, the functions $h = [h_k]$ of the family $\mathcal{H}$ computed are defined by:

$$\forall k \in \{1, \ldots, Q\}, \ h_k(x) = w_k^T \Phi(x) + b_k \tag{3}$$

where Φ is a nonlinear map into a high dimensional reproducing kernel Hilbert space.

3.2 Pathway to bound the capacity measure

In a nutshell, the strategy implemented consists in deriving progressively a bound on M-fat$_{\Delta\mathcal{H}^\gamma}(\epsilon)$ in terms of simpler dimensions of basic discriminant models, for which sharp bounds have already been found. For simplicity, we first restrict to the case where the model is linear. Taking the biases b_k of (3) into account can be performed as in [14]. Furthermore, since M-fat$_{\Delta\mathcal{H}^\gamma}(\epsilon) \leq$ M-fat$_{\Delta\mathcal{H}}(\epsilon)$, $\Delta\mathcal{H}$ can be substituted to $\Delta\mathcal{H}^\gamma$ in the computations. In short, the problem of bounding the

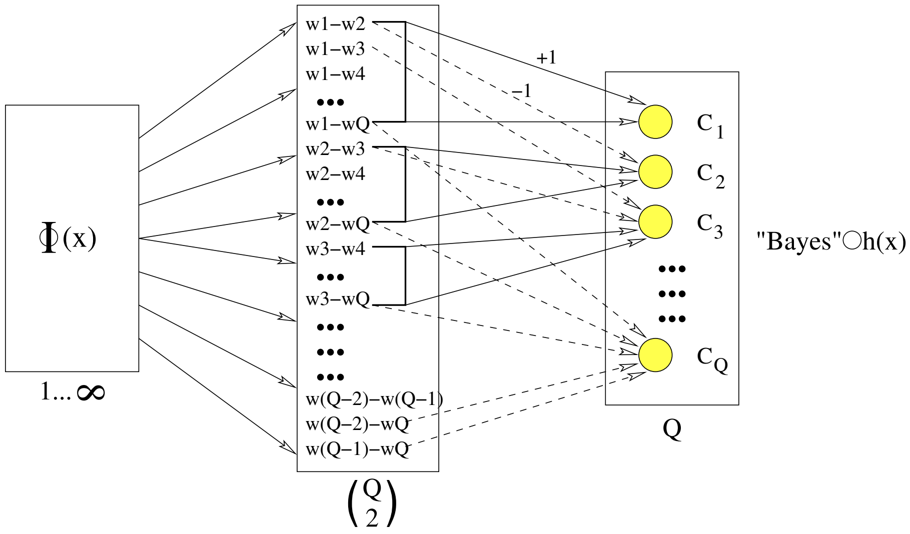

Figure 1: Architecture of a MLP with threshold units computing the same ϵ-margin discriminant functions as those associated with the functions h of $\mathcal{H}$. Each hidden unit computes a function $h_{k,l}(\Phi(x)) = t(1/2(w_k - w_l)^T \Phi(x))$, $(1 \leq k < l \leq Q)$, where $t(z) = 1$ if $z \geq \epsilon$, $t(z) = -1$ if $z \leq -\epsilon$ and $t(z) = 0$ otherwise. The weights of the output layer, either $+1$ (solid lines) or -1 (dashed lines), and the biases are chosen so that the output units compute a logical AND ($t(z) = 1$ if $z = Q - 1$, $t(z) = -1$ otherwise).

covering number of M-SVMs can be reduced to the one of bounding M-fat$_{\Delta\mathcal{H}}(\epsilon)$ in the case where $\mathcal{H}$ is a multivariate linear model. To do so, we consider the multi-layer perceptron (MLP) of Figure 1. The output of this MLP is the canonical coding of category C_k if and only if $\Delta h_k(x) \geq \epsilon$. If there is no index satisfying this property, then the output is simply equal to -1_Q (the example is not assigned to any category). As a consequence, M-fat$_{\Delta\mathcal{H}}(\epsilon)$ is exactly the graph dimension of the MLP. This dimension can be bounded from above using the approach described in [3]. To that end, the notion of growth function must preliminary be introduced.

Definition 11 (Growth function) *Let $\mathcal{F}$ be a set of binary-valued functions f on $\mathcal{X}$. For $s_m \in \mathcal{X}^m$, let $\Pi_{\mathcal{F}}(s_m)$ be the number of different functions corresponding to restrictions of functions in $\mathcal{F}$ to s_m. Then the* growth function *of $\mathcal{F}$ is the function with maps each positive integer to:*

$$\Pi_{\mathcal{F}}(m) = \max_{s_m \in \mathcal{X}^m} \Pi_{\mathcal{F}}(s_m)$$

Given a set $s_m \in \mathcal{X}^m$, a vector $v_c = [c_i] \in \{1, \ldots, Q\}^m$ and a function g computed by the MLP, let f be the binary function on s_m such that $f(x_i) = 1$ if $g(x_i)$ is the coding of category C_{c_i}, and 0 if it is the coding of another category (the case $g(x_i) = -1_Q$ is not permitted). Let $\mathcal{F}_{s_m, v_c}$ be the set of functions f. The growth function of the MLP is defined as follows:

$$\Pi_{MLP}(m) = \max_{v_c} \max_{s_m} \Pi_{\mathcal{F}_{s_m, v_c}}(s_m) \tag{4}$$

With this definition at hand, the proof of the following lemma is straightforward.

Lemma 1

$$\Pi_{MLP}(m) \leq \left(\Pi_{\Delta w}(m)\right)^{Q(Q-1)/2}$$

where $\Pi_{\Delta w}$ is the growth function of any of the hidden units, neglecting once more the state 0.

This lemma can be used to bound the graph dimension of the MLP in a way similar to what was done in [3]. It suffices then to bound $\Pi_{\Delta w}(m)$ in terms of the corresponding fat-shattering dimension, which can be done thanks to a generalized Sauer's lemma. This constitutes the last step of the decomposition, since sharp bounds on the fat-shattering dimension of linear classifiers are available. The following one, extending Theorem 4.6 in [2], is of particular interest here, since it involves a constraint on $\|w\|$ and makes no hypothesis on the dimensionality of the feature space.

Theorem 3 *Suppose that $\mathcal{X}$ is the ball of radius $\Lambda_{\mathcal{X}}$ in a Hilbert space $E_{\mathcal{X}}$ and consider the set $\mathcal{H}$ of linear functions h such that $h(x) = w^T x$ with $\|w\| \leq \Lambda_w$. Then, for all $\epsilon > 0$,*

$$fat_{\mathcal{H}}(\epsilon) \leq \left(\frac{\Lambda_{\mathcal{X}}\Lambda_w}{\epsilon}\right)^2$$

4 Conclusion

Simple extensions of standard results in statistical learning dealing with margin classifiers have provided us with a theoretical framework to study the generalization capabilities of multi-class discriminant systems, and more specifically multi-class SVMs. Other lines of attack can obviously be followed, as the one introduced in [14], leading to other expressions of the guaranteed risk and by way of consequence other implementations of the SRM inductive principle. However, our study has clearly highlighted a strong dependence of the guaranteed risk introduced in [7] on the control terms usually used in M-SVMs, namely $\sum_{k<l} \|w_k - w_l\|^2$ [13, 12, 6] and $\max_{k<l} \|w_k - w_l\|^2$ [6].

References

[1] P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[2] P.L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C.J.C Burges, and A. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*. The MIT Press, Cambridge, 1999.

[3] E.B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

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

[5] R.M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.*, 15(4):1306–1326, 1987.

[6] Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2):168–179, 2002.

[7] Y. Guermeur, A. Elisseeff, and D. Zelus. Bounding the capacity measure of multi-class discriminant models. Technical Report NC-TR-2002-123-R, NeuroCOLT2, 2002. (revised).

[8] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and Systems Sciences*, 48(3):464–497, 1994.

[9] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. Technical Report 1043, University of Wisconsin, Madison, Department of Statistics, 2001.

[10] J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *NIPS'12*, pages 547–553, 2000.

[11] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *ICKDDM'95*, pages 252–257, 1995.

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

[13] 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.

[14] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization Performance of Regularization Networks and Support Vector Machines *via* Entropy Numbers of Compact Operators. *IEEE Trans. on Information Theory*, 47(6):2516–2532, 2001.