Bound on the Risk for M-SVMs

Yann Guermeur
LORIA
Université Nancy I

André Elisseeff
MPI for Biological Cybernetics
Tuebingen

Dominique Zelus
Wiener Lab, CIBIO
Rosario

http://www.loria.fr/~guermeur/

Overview

Guaranteed risk for multi-class discriminant models

- Statistical multi-class pattern recognition
- Margin-based bound on the risk : bi-class case
- Margin-based bound on the risk : multi-class case

M-fat-shattering dimension of M-SVMs

- Architecture and training algorithms of M-SVMs
- Capacity measure of M-SVMs and graph dimension of threshold MLPs
- Dependence of the capacity measure on the control term of the objective function

Multi-class pattern recognition

Hypotheses : empirical data characterizing a joint probability distribution

- Q-category discrimination problem
- $Z = (X, Y)$: random variable on a probability space
- $X(\Omega) = \mathcal{X}$: input space (set of descriptions), $Y(\Omega) = \mathcal{Y}$: finite set of categories
- P : joint probability distribution function on $\mathcal{X} \times \mathcal{Y}$, fixed but unknown
- $s = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset (\mathcal{X} \times \mathcal{Y})^m$, learning set : observations i.i.d. according to P
- $\mathcal{H}$: **family of vector-valued functions** $h = [h_k]$, $(1 \leq k \leq Q)$, from $\mathcal{X}$ into $\mathbb{R}^Q$

Goal : for a given pattern, find its category

Find in $\mathcal{H}$ a function associated with the lowest expected risk (generalization error)

$$R(h) = R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{f(x) \neq y\}} dP(x, y)$$

f **: discriminant function** corresponding to h, obtained by choosing the category associated with the **index of the highest output**

Empirical margin risk and uniform convergence result - the bi-class case

$\mathcal{Y} = \{-1, 1\}$

Definition 1 (Empirical margin risk (Bartlett 98)) *Let h be a real-valued function on $\mathcal{X}$. For a training data sequence $s_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$ of length m and a real number $\gamma > 0$*

$$R_{s_m}^{\gamma}(h) = \frac{1}{m} \left| \{(x_i, y_i) \in s_m \ / \ y_i h(x_i) < \gamma\} \right|$$

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

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

Empirical margin risk and uniform convergence result - the bi-class case

Capacity measure : covering numbers

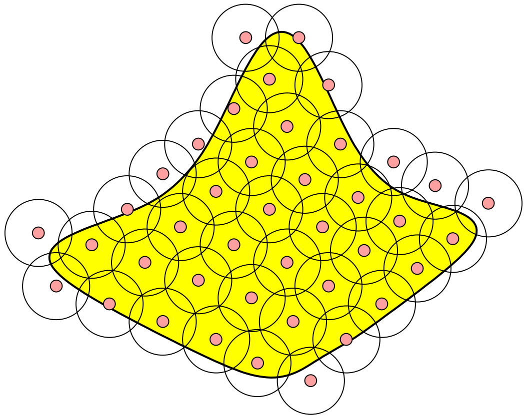

FIG. 1 – ϵ-net of a set $\mathcal{G}$ in a pseudo-metric or Banach space

Definition 2 (Covering numbers)
$\mathcal{N}(\epsilon, \mathcal{G}, \|.\|)$ = *minimum number of balls of radius* ϵ *required to cover the set* $\mathcal{G}$

Empirical margin risk and uniform convergence result - the bi-class case

Theorem 1 (Bartlett 98) *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 bi-class discriminant model $\mathcal{H}$ is bounded above by :*

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

where $\mathcal{H}^\gamma = \{\pi_\gamma \circ h \ / \ h \in \mathcal{H}\}$

$$\forall s_m \in \mathcal{X}^m,\ \forall (h^{(1)}, h^{(2)}) \in \mathcal{H}^2,\ d_{l_\infty(s_m)}(h^{(1)}, h^{(2)}) = \max_{x_i \in s_m} \left| h^{(1)}(x_i) - h^{(2)}(x_i) \right|$$

$$\mathcal{N}_\infty(\gamma/2, \mathcal{H}^\gamma, 2m) = \max_{s_{2m} \in \mathcal{X}^{2m}} \mathcal{N}(\gamma/2, \mathcal{H}^\gamma, d_{l_\infty(s_{2m})})$$

Empirical margin risk and uniform convergence result - the multi-class case

Definition 3 (Canonical function)
$h = [h_k] : \mathcal{X} \longrightarrow \mathbb{R}^Q$

$M_1(h, x)$ *: smallest index* l *such that* $h_l(x) = \max_k h_k(x)$

$M_2(h, x)$ *: smallest index* $l \neq M_1(h, x)$ *such that* $h_l(x) = \max_{k \neq M_1(h,x)} h_k(x)$

$\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, *function from* $\mathcal{X}$ *into* $\mathbb{R}^Q$ *satisfying*

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

Definition 4 (Empirical margin risk (Elisseeff & al. 99)) *The empirical risk with margin* $\gamma \in (0, 1]$ *of* h *on a set* $s_m = \{(x_1, C(x_1)), \ldots, (x_m, C(x_m))\}$ *of size* m *is*

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

Empirical margin risk and uniform convergence result - the multi-class case

Theorem 2 (Elisseeff & al. 99) *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 :*

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

where $\Delta h^{\gamma} = [\pi_{\gamma} \circ \Delta h_k]$, $(1 \leq k \leq Q)$, $\Delta\mathcal{H}^{\gamma} = \{\Delta h^{\gamma} \ / \ h \in \mathcal{H}\}$

$$\forall s_m \in \mathcal{X}^m, \ \forall (h^{(1)}, h^{(2)}) \in \mathcal{H}^2, \ d_{l_\infty, l_\infty(s_m)}(h^{(1)}, h^{(2)}) = \max_{x_i \in s_m} \max_{k \in \{1,\dots,Q\}} \left| h^{(1)}_k(x_i) - h^{(2)}_k(x_i) \right|$$

$$\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^{\gamma}, 2m) = \max_{s_{2m} \in \mathcal{X}^{2m}} \mathcal{N}(\gamma/2, \Delta\mathcal{H}^{\gamma}, d_{l_\infty, l_\infty(s_{2m})})$$

Bound on the covering numbers - bi-class case

Theorem 3 (Alon & al. 97) *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $[0,1]$. For every value of γ in $(0,1]$ and every value of m in $\mathbb{N}^*$, the following bound is true :*

$$\mathcal{N}_\infty(\gamma, \mathcal{H}, m) \leq 2 \left(\frac{4m}{\gamma^2} \right)^{d \log_2(2em/(d\gamma))}$$

where $d = fat_{\mathcal{H}}(\gamma/4)$.

Extended notions of VC dimension

Definition 5 (Fat-shattering dimension (Kearns & Schapire 90)) *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 vector v_b is then said to witness *the γ-shattering of s_m by $\mathcal{H}$. 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 6 (Graph dimension (Dudley 87, Natarajan 89)) *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 $\mathcal{G}(\mathcal{H})$.*

M-fat-shattering dimension

Definition 7 (M-fat-shattering dimension (Guermeur & al. 02)) *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 $v_c = [c_i] \in \{1, \dots, Q\}^m$ 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 couple (v_b, v_c) is then said to witness *the M-γ-shattering of s_m by $\mathcal{H}$. 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.*

M-fat-shattering dimension : extension of the fat-shattering dimension to the multivariate case and scale-sensitive version of the graph dimension

Bound on the covering numbers - multi-class case

Theorem 4 (Guermeur & al. 02) *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For every value of γ in $(0,1]$ and every value of m in $\mathbb{N}^*$, the following bound is true :*

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

where $d = M\text{-}fat_{\Delta\mathcal{H}^\gamma}(\gamma/8)$.

Multi-class Support Vector Machines

Architecture

The functions $h = [h_k]$ of the family $\mathcal{H}$ considered are defined by :

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

Φ is a nonlinear map into the *feature space*

Training algorithm

Let K be the *kernel* associated with Φ :

$$\forall (x^{(1)}, x^{(2)}) \in \mathcal{X}^2, \; K(x^{(1)}, x^{(2)}) = \langle \Phi(x^{(1)}), \Phi(x^{(2)}) \rangle$$

and let $s_m = \{(x_1, C(x_1)), \dots, (x_m, C(x_m))\}$ be the training set

In its dual formulation, training consists in finding the values of the coefficients β_{ik} in :

$$\forall k \in \{1, \dots, Q\}, \; h_k(x) = \sum_{i=1}^{m} \beta_{ik} K(x_i, x) + b_k$$

Training algorithms of M-SVMs (primal formulation)

Problem 1 (M-SVM1 (Vapnik & Blanz 98, Weston & Watkins 98))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

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

Problem 2 (M-SVM2 (Guermeur 02))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} t^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.t. \begin{cases} \|w_k - w_l\|^2 \leq t^2, \ (1 \leq k < l \leq Q) \\ \textit{Contraints of Problem 1} \end{cases}$$

M-fat-shattering dimension of M-SVMs and graph dimension of a MLP

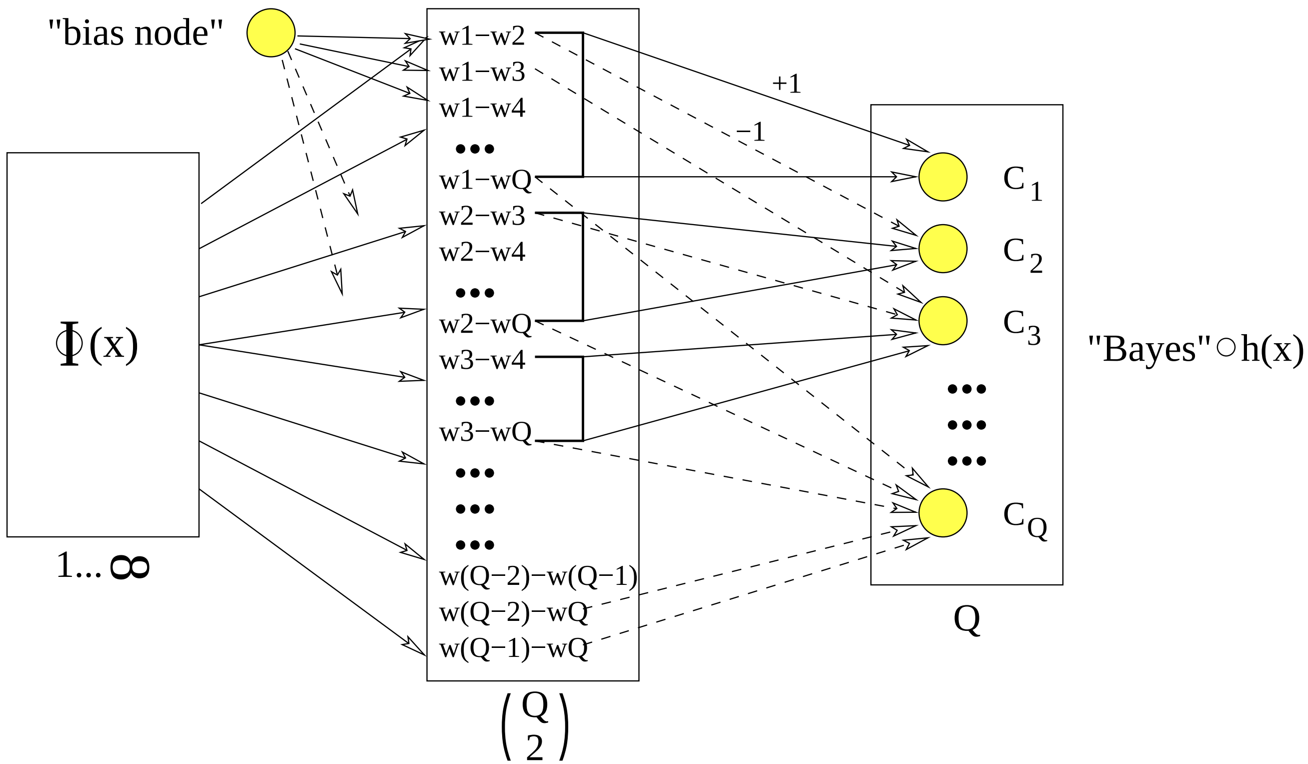

FIG. 2 – MLP computing the same discriminant functions as the M-SVMs

$$h_{k,l}(\Phi(x)) = t_h(1/2(w_k - w_l)^T \Phi(x) + b_{k,l}), \; (1 \le k < l \le Q)$$

$$t_h(z) = 1 \text{ if } z \ge \epsilon, \; t_h(z) = -1 \text{ if } z \le -\epsilon \text{ and } t_h(z) = 0 \text{ otherwise}$$

M-fat-shattering dimension of M-SVMs and graph dimension of a MLP

Definition 8 (uniform M-fat-shattering dimension) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, the* uniform M-fat-shattering dimension *UM-fat$_{\mathcal{H}}$ of $\mathcal{H}$ is simply M-fat$_{\mathcal{H}}$ in the case where the components of vector v_b are constrained to take only Q different values, one for each category. In other words, if two components of the vector v_c are equal, then the correponding components of the vector v_b are also equal.*

Pathway linking the capacity measures of the two models

(1) $\text{M-fat}_{\text{M-SVM}}(\epsilon) \leq K_{\gamma,\epsilon}\ \text{UM-fat}_{\text{M-SVM}}(\epsilon/2)$
(2) The MLP must be adapted to output a category different from $C(x_i)$ when $y_i = -1$
(3) $\text{UM-fat}_{\text{M-SVM}}(\epsilon)$ is inferior or equal to the graph dimension of the MLP

Graph dimension of the MLP

(1) The *growth function* Π_{MLP} of the MLP is inferior or equal to the product of the growth functions of each hidden unit **(Baum & Haussler 89)**

(2) The growth function of each hidden unit can be bounded in terms of the corresponding fat-shattering dimension d_ϵ (Vapnik-Chervonenkis-Sauer-Shela lemma)

(3)

$$\Pi_{MLP}(m) < \left(\frac{em}{d_\epsilon}\right)^{1/2Q(Q-1)d_\epsilon}$$

(4)

$$d_{graph}(MLP) < Q(Q-1)\log_2\left[eQ(Q-1)\right] d_\epsilon$$

$$\Longrightarrow$$

The fat-shattering dimension of linear classifiers appears to be the central parameter to study

Fat-shattering dimension of hyperplanes and objective functions of M-SVMs

Theorem 5 (Bartlett & Shawe-Taylor 99) *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 $\epsilon > 0$,*

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

Remarks

- $E_{\mathcal{X}}$ can be an infinite dimensional space
- The model is affine (not linear) $\Longrightarrow$ additional multiplicative coefficient

$$\Longrightarrow$$

Possible control terms

- $\sum_{k<l}^{Q} \|w_k - w_l\|^2$,
- $\max_{k<l} \|w_k - w_l\|^2$,
- ...

Objective functions of standard M-SVMs

Multi-class SVM	Objective function	Add. const.
Vapnik & Blanz 98	$J_1(w,b,\xi) = \sum_{k=1}^{Q} \|w_k\|^2 + C_1 1^T \xi$	-
Weston & Watkins 98	$J_1(w,b,\xi) = \sum_{k=1}^{Q} \|w_k\|^2 + C_1 1^T \xi$	-
Bredensteiner & al. 99	$J_2(w,b,\xi) = \sum_{k<l}^{Q} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2 + C_2 1^T \xi$	-
Guermeur & al. 00	$J_3(w,b,\xi) = \sum_{k<l}^{Q} \|w_k - w_l\|^2 + C_3 1^T \xi$	$\sum_{k=1}^{Q} w_k = 0_d$

Objective function	Add. const.	C	Solution
$J_1(w,b,\xi)$	-	C_1	$\left(w^{(1)}, b^{(1)}, \xi^{(1)}, \alpha^{(1)}, \beta^{(1)}\right)$
$J_2(w,b,\xi)$	-	$(Q+1)C_1$	$\left(w^{(1)}, b^{(1)}, \xi^{(1)}, (Q+1)\alpha^{(1)}, (Q+1)\beta^{(1)}\right)$
$J_3(w,b,\xi)$	$\sum_{k=1}^{Q} w_k = 0_d$	QC_1	$\left(w^{(1)}, b^{(1)}, \xi^{(1)}, Q\alpha^{(1)}, Q\beta^{(1)}, 0_d\right)$

The same set of primal variables generates solutions for the three problems

$\Longrightarrow$ **All these multi-class SVMs are equivalent**

Conclusions and future work

Conclusions

- New pathway to bound the generalization performance of multi-class discriminant models
- New justification of the control terms used for the M-SVMs
- Possibility to develop new machines

Future work

- Comparison with the direct approach involving the *entropy numbers of a linear operator* **(Williamson & al. 01)**
- Comparison with works involving *data dependent capacity measures* **(Boucheron & al. 99, Bartlett & al. 02, Bousquet 02)**
- Design of optimization methods devoted to the new machines