Model Selection for Multi-class SVMs

Yann Guermeur	Myriam Maumy	Frédéric Sur
LORIA - CNRS	IRMA - ULP	LORIA - CNRS

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

http://www-irma.u-strasbg.fr/~mmaumy/

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

Overview

Multi-class kernel machines

- Statistical framework
- M-SVMs
- Model selection - hyperparameters

Guaranteed risks for M-SVMs

- Basic uniform convergence result
- Covering numbers of the M-SVM and entropy numbers of an evaluation operator
- Bounding the entropy numbers of the evaluation operator

Deriving the hyperparameters from the bound

- Algorithm
- Experimental results

Motivation of the study

1. Support vector machines (SVMs) have been applied with success to many problems in pattern recognition
2. Theory suggests that multi-class SVMs (M-SVMs) should prove superior to decomposition schemes to compute polychotomies
3. Model selection (setting the hyperparameters ...) is a central problem in statistical learning
4. No solution has been proposed so far to tune the hyperparameters of M-SVMs

$\Longrightarrow$ **Significant benefits should result from the specification of a model selection procedure dedicated to M-SVMs**

Statistical learning - Pattern recognition

Hypotheses : empirical data characterizing a joint probability measure

- Q-category discrimination problem

- $\mathcal{X}$: *space of description*, $\mathcal{C}$: finite set of *categories*, $\mathcal{C} = \{1, \ldots, Q\}$
- $\mathcal{B}$, σ-algebra on $\mathcal{X} \times \mathcal{C}$
- μ : joint probability measure on $(\mathcal{X} \times \mathcal{C}, \mathcal{B})$, *fixed but unknown*
- $s_m = ((x_1, C(x_1)), \ldots, (x_m, C(x_m))) \in (\mathcal{X} \times \mathcal{C})^m$, *learning set* : observations i.i.d. according to μ

- $\mathcal{H}$: class 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)*

$$R(h) = R(f) = \mu\left\{(x, k) : \ f(x) \neq k\right\} = \int_{\mathcal{X} \times \mathcal{C}} \mathbb{1}_{\{f(x) \neq k\}}(x, k) d\mu(x, k)$$

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

Multi-class Support Vector Machines (M-SVMs)

Architecture / class of functions computed (common to all machines)

1. Choice of a *positive definite kernel* $\kappa \Longrightarrow$ implicit choice of a *reproducing kernel Hilbert space* (RKHS) $(H, \langle ., . \rangle_H)$
2. Functions $h = (h_k)_{1 \leq k \leq Q}$ computed by the architecture are defined by :

$$\forall x \in \mathcal{X}, \ \forall k \in \{1, \ldots, Q\}, \ h_k(x) = \tilde{h}_k(x) + b_k = \langle w_k, \Phi(x) \rangle_H + b_k$$

   where Φ is any function satisfying

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

- The vectors $\Phi(x)$ belong to the *feature space* $E_{\Phi(\mathcal{X})} = (H, \langle ., . \rangle_H)$
- Φ is "the" *feature map* from $\mathcal{X}$ into the feature space $E_{\Phi(\mathcal{X})}$
- $\tilde{\mathcal{H}} = \left\{ \tilde{h} = \left(\tilde{h}_k \right) \right\} \subset E^Q_{\Phi(\mathcal{X})}$

Training algorithm of M-SVMs : general principle

Let $s_m = ((x_1, C(x_1)), \ldots, (x_m, C(x_m)))$ be the training set

Primal formulation (with *soft margins*)

Problem 1

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \|\tilde{h}\|^2_{\tilde{\mathcal{H}}} + C \sum_{i=1}^{m} \ell\left(h(x_i), C(x_i)\right) \right\}$$

Dual formulation

A representer theorem establishes that training (solving Problem 1) amounts to finding the values of the coefficients β_{ik} in :

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

Training algorithms of M-SVMs (primal formulation)

Problem 2 (M-SVM1 (Vapnik & Blanz, 1998, Weston & Watkins, 1998, . . .))

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

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

Problem 3 (M-SVM2 (Guermeur, 2002))

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

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

Training algorithms of M-SVMs (primal formulation)

Problem 4 (M-SVM3 (Crammer & Singer, 2001))

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

$$s.t.\ \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)} - b_k + \delta_{C(x_i),k} \geq 1 - \xi_i,\ (1 \leq i \leq m), (1 \leq k \leq Q)$$

Problem 5 (M-SVM4 (Lee *et al.*, 2001))

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

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

Behaviour of a M-SVM (M-SVM1)

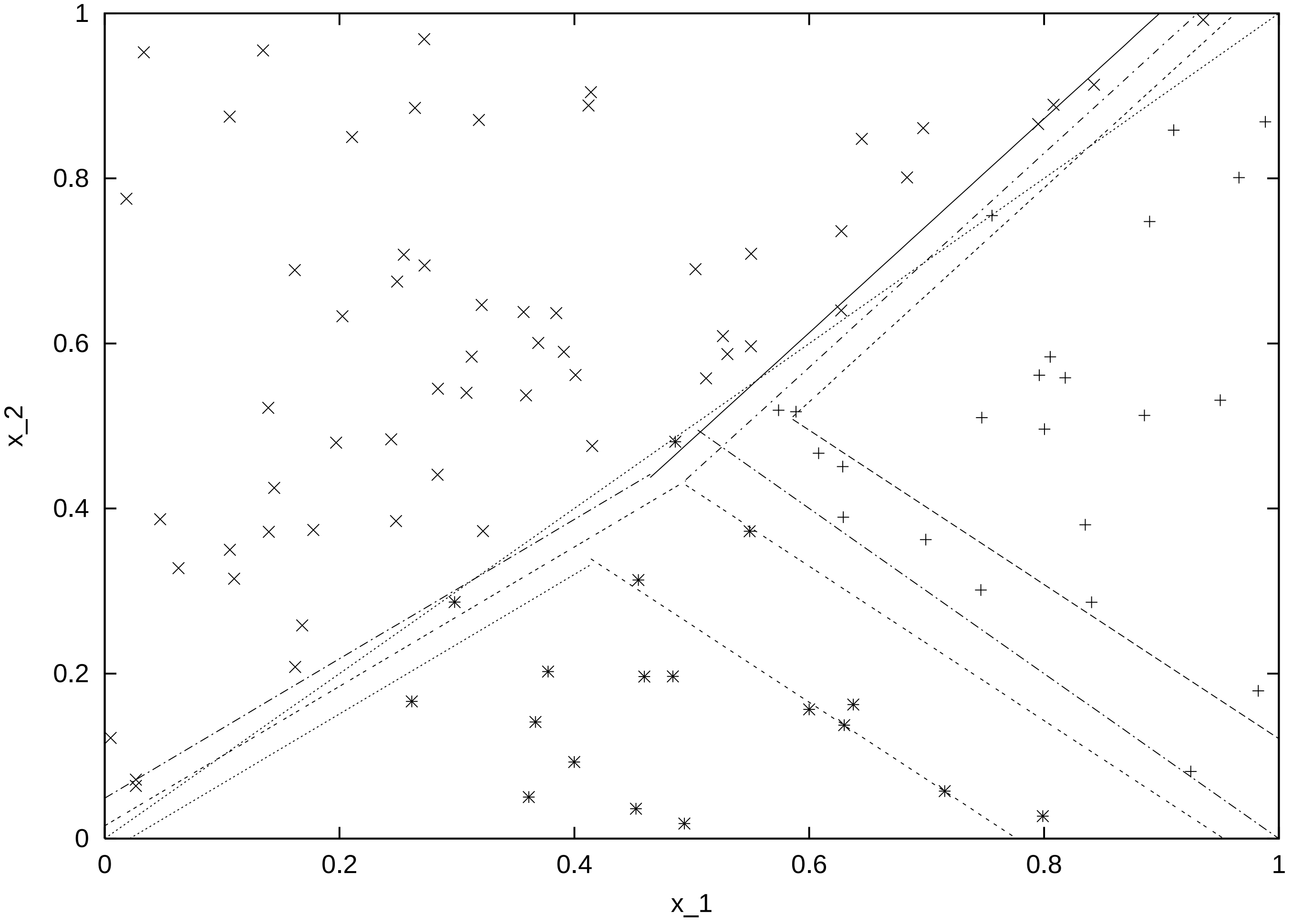

FIG. 1 – 3 categories linearly separable in 2D

Behaviour of a M-SVM (M-SVM1)

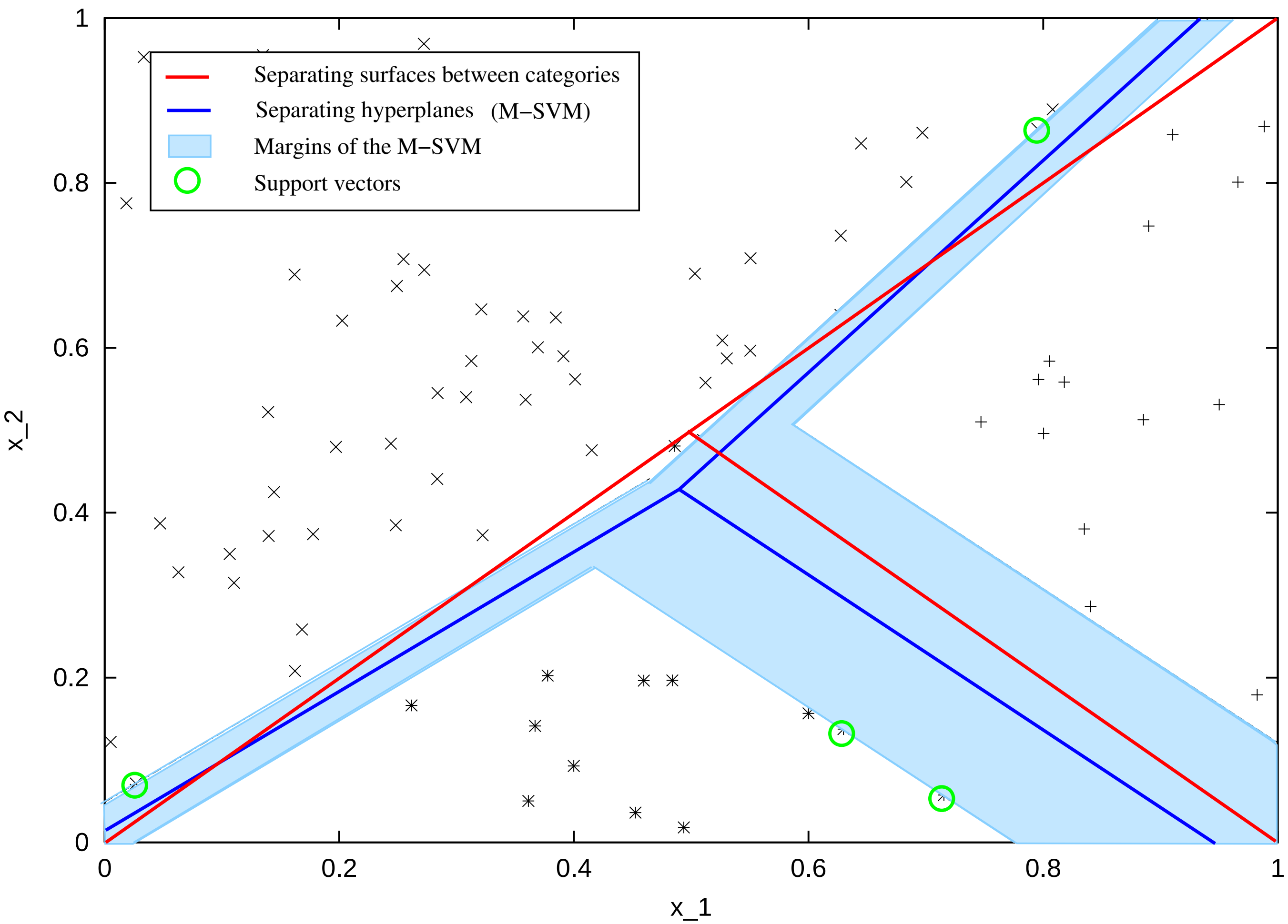

FIG. 2 – Hyperplanes of the linear M-SVM and support vectors

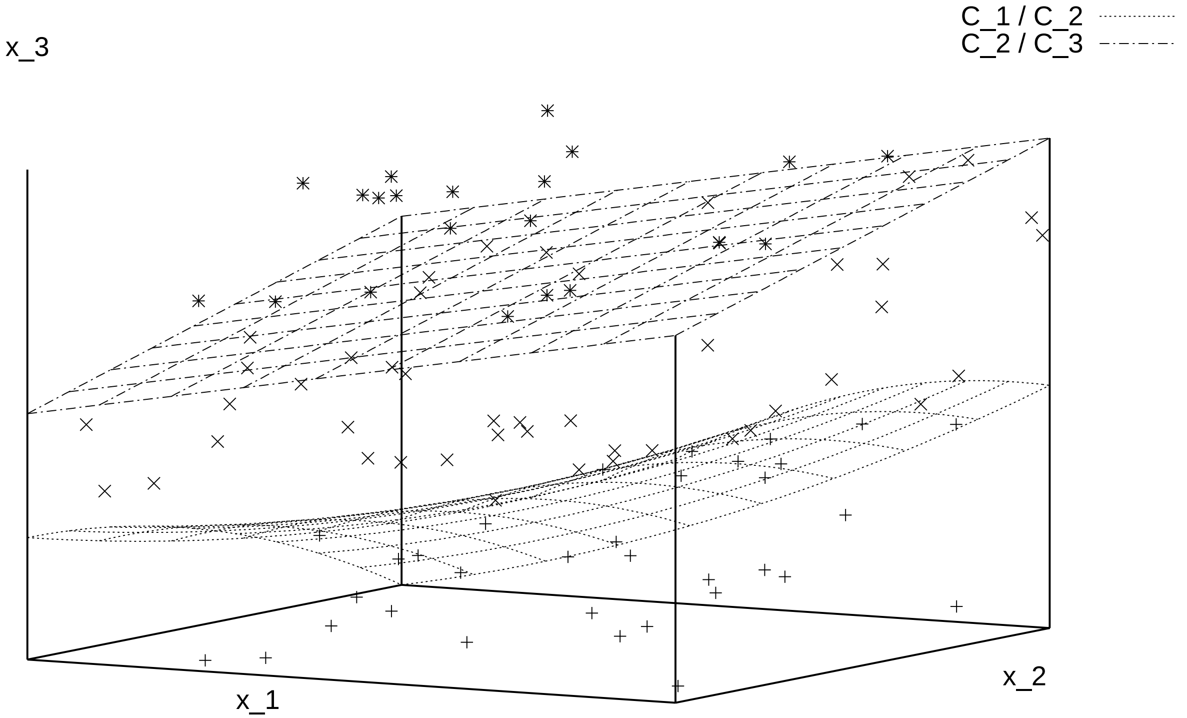

FIG. 3 – 3 categories nonlinearly separable in 3D

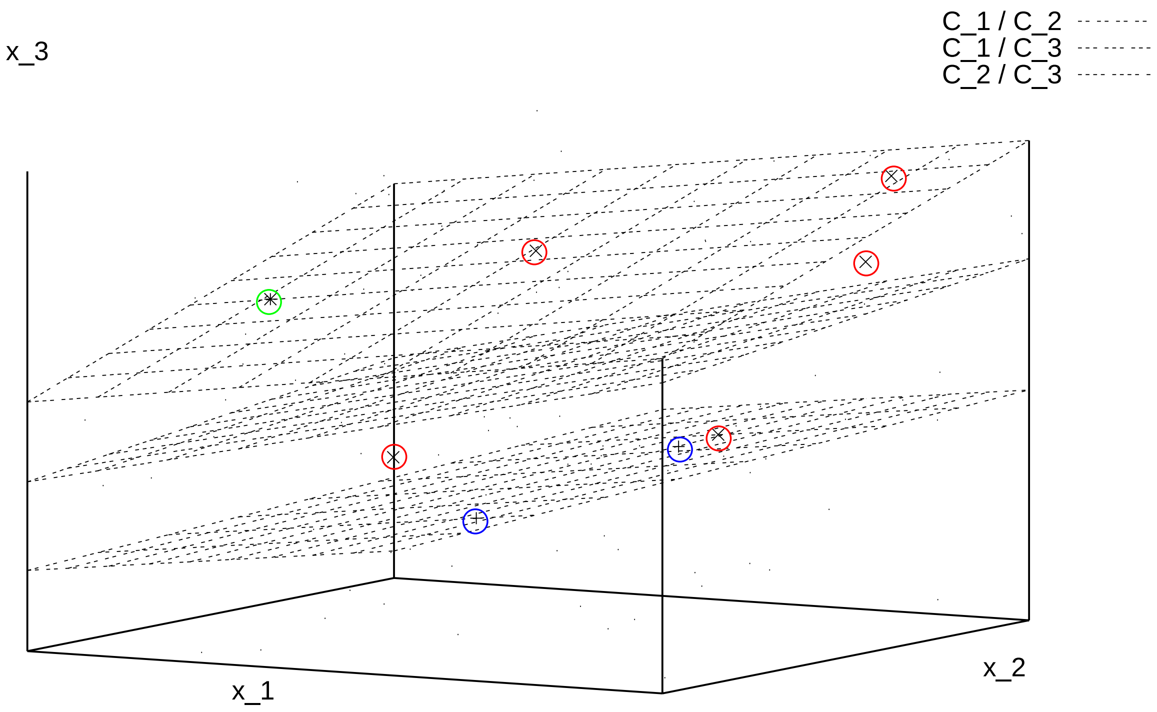

FIG. 4 – Hyperplanes of the linear M-SVM and support vectors

Model selection for M-SVMs

Soft margin parameter C

The larger C, the lower the empirical risk, but the higher the capacity (confidence interval) $\Longrightarrow$ a compromise is to be found

Parameters of the kernel

- r.b.f. (Gaussian) kernel

$$\kappa(x, x') = \exp\left(-\mu \|x - x'\|^2\right)$$

- Polynomial kernel

$$\kappa(x, x') = (\langle x, x' \rangle + \mu)^d$$

- Sigmoid kernel

$$\kappa(x, x') = \tanh\left(\mu \langle x, x' \rangle + \nu\right)$$

Guaranteed risk

General expression of the (one-sided) bounds

Let $\mathcal{H}$ be the class of functions implemented by... and s_m a training set of cardinality m. With probability at least $1 - \delta$, for all function h in $\mathcal{H}$,

$$R(h) \leq R_{s_m}(h) + \Psi\left(\operatorname{cap}\left(\mathcal{H}\right), \delta, m\right)$$

where $\operatorname{cap}(\mathcal{H})$ is the *capacity measure* of the class $\mathcal{H}$.

Guaranteed risk - multi-class case - vector-valued functions

Capacity measure : covering numbers

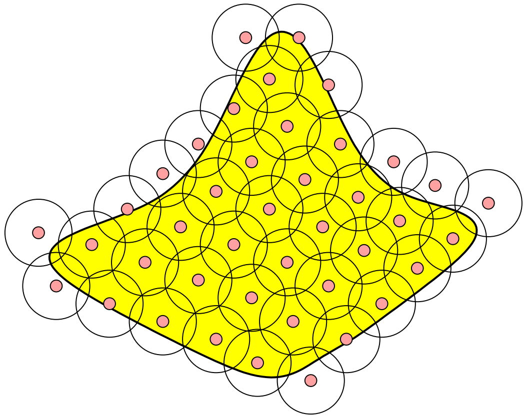

FIG. 5 – ϵ-*net* and ϵ-*cover* of a set F in a pseudo-metric space (E, ρ_E) or a Banach space $(E, \|.\|_E)$

Definition 1 (Covering numbers)
$\mathcal{N}(\epsilon, F, \|.\|_E) =$ *minimum number of balls of radius* ϵ *in* E *required to cover the set* F

Guaranteed risk - multi-class case - vector-valued functions

Definition 2 (Δ, Δ_γ operators (Elisseeff *et al.*, 1999)) *Let $h = (h_k)_{1 \leq k \leq Q}$ be a function from $\mathcal{X}$ into $\mathbb{R}^Q$.*

– *Function $\Delta h = (\Delta h_k)_{1 \leq k \leq Q}$, from $\mathcal{X}$ into $\mathbb{R}^Q$, is given by :*

$$\forall k \in \{1, \dots, Q\}, \quad \Delta h_k(x) = \frac{1}{2} \min_{l \neq k} \{h_k(x) - h_l(x)\} = \frac{1}{2} \left\{ h_k(x) - \max_{l \neq k} h_l(x) \right\}.$$

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

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

– *For $\gamma > 0$, function $\Delta_\gamma h = (\Delta_\gamma h_k)_{1 \leq k \leq Q}$, from $\mathcal{X}$ into $[-\gamma, \gamma]^Q$, is given by :*

$$\forall k \in \{1, \dots, Q\}, \quad \Delta_\gamma h_k(x) = \pi_\gamma (\Delta h_k(x)).$$

Guaranteed risk - multi-class case - vector-valued functions

Empirical margin risk

Definition 3 (Empirical margin risk (Elisseeff *et al.*, 1999)) *The empirical risk with margin $\gamma > 0$ of h on a set $s_m = \{(x_1, C(x_1)), \ldots, (x_m, C(x_m))\}$ of size m is :*

$$R_{\gamma, s_m}(h) = \frac{1}{m} \cdot \# \left\{ (x_i, C(x_i)) \in s_m : \; \Delta h_{C(x_i)}(x_i) < \gamma \right\}$$

Notations

$$\Delta_\gamma \mathcal{H} = \{\Delta_\gamma h : \; h \in \mathcal{H}\}$$

$$\forall s_m \in \mathcal{X}^m, \; \forall (g, g') \in \mathcal{G}^2, \; d_{s_m}(g, g') = \max_{x_i \in s_m} \|g(x_i) - g'(x_i)\|_\infty$$

$$\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{G}, m) = \sup_{s_m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \mathcal{G}, d_{s_m})$$

Basic capacity measure

The capacity of $\mathcal{H}$ to discriminate points in $\mathcal{X}$ with margin γ is characterized by

$$\mathcal{N}_{\infty,\infty}(\epsilon, \Delta_\gamma \mathcal{H}, m)$$

Guaranteed risk - multi-class case - vector-valued functions

Uniform convergence result (extension of the central result in (Bartlett, 1998))

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

$$R(h) \leq R_{\gamma,s_m}(h) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

Simple pathway to bound the covering numbers of M-SVMs (after (Williamson *et al.*, 2001))

Simplifying the expression of the covering numbers

Lemma 2 *Let* $\mathcal{H}$ *be a class of functions from* $\mathcal{X}$ *into* $\mathbb{R}^Q$. *For every value of* ϵ *and* γ *satisfying* $0 < \epsilon \leq \gamma \leq 1$, *the following bounds hold true :*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \Delta_\gamma \mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \Delta \mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$$

Lemma 3 *Let* $\mathcal{H}$ *be a class of functions computed by a M-SVM. Suppose that* $b = (b_k) \in [-\beta, \beta]^Q$. *Then, for all positive value of* ϵ *:*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m) \leq \left(2\left\lceil \frac{\beta}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,\infty}(\epsilon/2, \tilde{\mathcal{H}}, m)$$

A function $\tilde{h}$ in $\tilde{\mathcal{H}}$ is characterized by the vector $\mathbf{w} = (w_k)_{1 \leq k \leq Q}$ in $E^Q_{\Phi(\mathcal{X})}$.

Lemma 4 *Let* $\tilde{\mathcal{H}}$ *be defined as above, under the additional constraint that* $\|\mathbf{w}\| \leq \Lambda_w$. *Let* $\mathcal{U}$ *be its restriction to the functions satisfying* $\|\mathbf{w}\| \leq 1$.

$$\forall \epsilon > 0, \;\; \mathcal{N}_{\infty,\infty}(\Lambda_w \epsilon, \tilde{\mathcal{H}}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m).$$

From covering numbers to entropy numbers

Definition 4 (Entropy numbers) *Let (E, ρ) be a pseudo-metric space and F a subset of E. Then, for n in $\mathbb{N}^*$, the n^{th} entropy number of F, $\epsilon_n(F)$, is :*

$$\epsilon_n(F) = \inf\left\{\epsilon > 0 : \ \mathcal{N}(\epsilon, F, \rho) \leq n\right\}$$

Definition 5 (Entropy numbers of a bounded linear operator) *Let E and F be Banach spaces endowed with the norms $\|.\|_E$ and $\|.\|_F$. Let $\mathcal{L}(E, F)$ be the Banach space of all bounded linear operators between $(E, \|.\|_E)$ and $(F, \|.\|_F)$. Let $S \in \mathcal{L}(E, F)$.*

$$\|S\| = \sup_{x \in U_E} \|Sx\|_F \quad \epsilon_n(S) = \epsilon_n(S(U_E))$$

From covering numbers to entropy numbers

Definition 6 (Evaluation operator) *Let s_m be any element of $\mathcal{X}^m$. We define S_{s_m} as the linear operator given by :*

$$\begin{array}{rccl} S_{s_m}: & E^Q_{\Phi(\mathcal{X})} & \longrightarrow & \ell_\infty^{Qm} \\ & \mathbf{w} & \mapsto & S_{s_m}(\mathbf{w}) = \left(\langle w_k, \Phi(x_i) \rangle \right)_{1 \leq k \leq Q,\ 1 \leq i \leq m} \end{array}$$

The connection between $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m)$ and the entropy numbers of S_{s_m} is given by the following proposition.

Lemma 5 *If for all $s_m \in \mathcal{X}^m$, $\epsilon_n(S_{s_m}) \leq \epsilon$, then $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m) \leq n$.*

From covering numbers to entropy numbers

To bound $\epsilon_n(S_{s_m})$, a result due to Maurey and Carl can be used.

Lemma 6 (Carl & Stephani, 1990) *Let H be a Hilbert space, m a positive integer and $S \in \mathfrak{L}(H, \ell_\infty^m)$. Then, for $1 \leq n \leq m$,*

$$\epsilon_{2^{n-1}}(S) \leq c\|S\| \left(\frac{1}{n} \log \left(1 + \frac{m}{n} \right) \right)^{1/2},$$

where c is a universal constant and by log *we denote the logarithm to base* 2.

Lemma 6 still holds without the hypothesis $n \leq m$.

Theorem 7 *Let $\mathcal{H}$ be the class of functions computed by a Q-category M-SVM under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the constraints that $\|\mathbf{w}\| \leq \Lambda_w$ and $b \in [-\beta, \beta]^Q$. For every value of γ in $(0, 1]$,*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m) \leq \left(2 \left\lceil \frac{4\beta}{\gamma} \right\rceil + 1 \right)^Q \cdot 2^{\frac{8c\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\gamma} \sqrt{\frac{2Qm}{\ln(2)}} - 1}.$$

From covering numbers to entropy numbers

In the case when the feature space $E_{\Phi(\mathcal{X})}$ has finite dimension, a simpler result can be used to bound the entropy numbers.

Proposition 8 (Carl & Stephani, 1990) *If an operator $S \in \mathfrak{L}(E, F)$ acting between arbitrary real Banach spaces E and F is of rank r, then for $n \in \mathbb{N}^*$,*

$$\epsilon_n(S) \leq 4\|S\| \cdot n^{-1/r}.$$

Theorem 9 *Let $\mathcal{H}$ be the class of functions computed by a Q-category M-SVM under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the constraints that $\|\mathbf{w}\| \leq \Lambda_w$ and $b \in [-\beta, \beta]^Q$. Suppose further that the dimensionality of $E_{\Phi(\mathcal{X})}$ is finite and equal to d. Then, for every value of γ in $(0, 1]$,*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m) \leq \left(2\left\lceil\frac{4\beta}{\gamma}\right\rceil + 1\right)^Q \cdot \left(\frac{32\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\gamma}\right)^{Qd}. \tag{1}$$

Training algorithms of M-SVMs (dual formulation)

$$
\begin{aligned}
J_d(\alpha) &= \frac{1}{2}\left\{ \begin{array}{l} \sum_{i \simeq j} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik}\alpha_{jl}\kappa(x_i, x_j) - 2\sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}\alpha_{jC(x_i)}\kappa(x_i, x_j) \\ + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}\alpha_{jk}\kappa(x_i, x_j) \end{array} \right\} \\
&\quad - \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \\
&= \frac{1}{2}\alpha^T H \alpha - 1_{Qm}^T \alpha
\end{aligned}
$$

where $i \simeq j$ means that x_i and x_j belong to the same category $C(x_i)$ and the $\alpha_{iC(x_i)}$ are dummy variables equal to 0.

Problem 6 (M-SVM1)

$$\min_{\alpha} J_d(\alpha)$$

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

Tuning the hyperparameters thanks to the bound on the risk

Principle : explore the space of hyperparameters with the bound as objective function

Illustration : tuning the *soft margin parameter* C

/* Initialization */
  $C_0 := C^{(0)}$, $\alpha^{(0)} := 0_{Qm}$, $\epsilon > 0$;
/* Main loop */
For $i := 1$ to nb_iter do
  train_SVM(C_{i-1}, s_m, $\alpha^{(i-1)}$) $\longrightarrow \alpha^{(i)}$;
  $C_i := C_{i-1} + \epsilon$;
done
/* Termination */
  $i_0 := \text{Argmin}_{1 \leq i \leq \texttt{nb\_iter}} \{$ compute_bound(C_i, s_m, $\alpha^{(i)}$) $\}$;
  $C := C_{i_0}$;

The exploration of the *regularization path* can be performed as in **(Hastie *et al.*, 2004)**

Experimental results

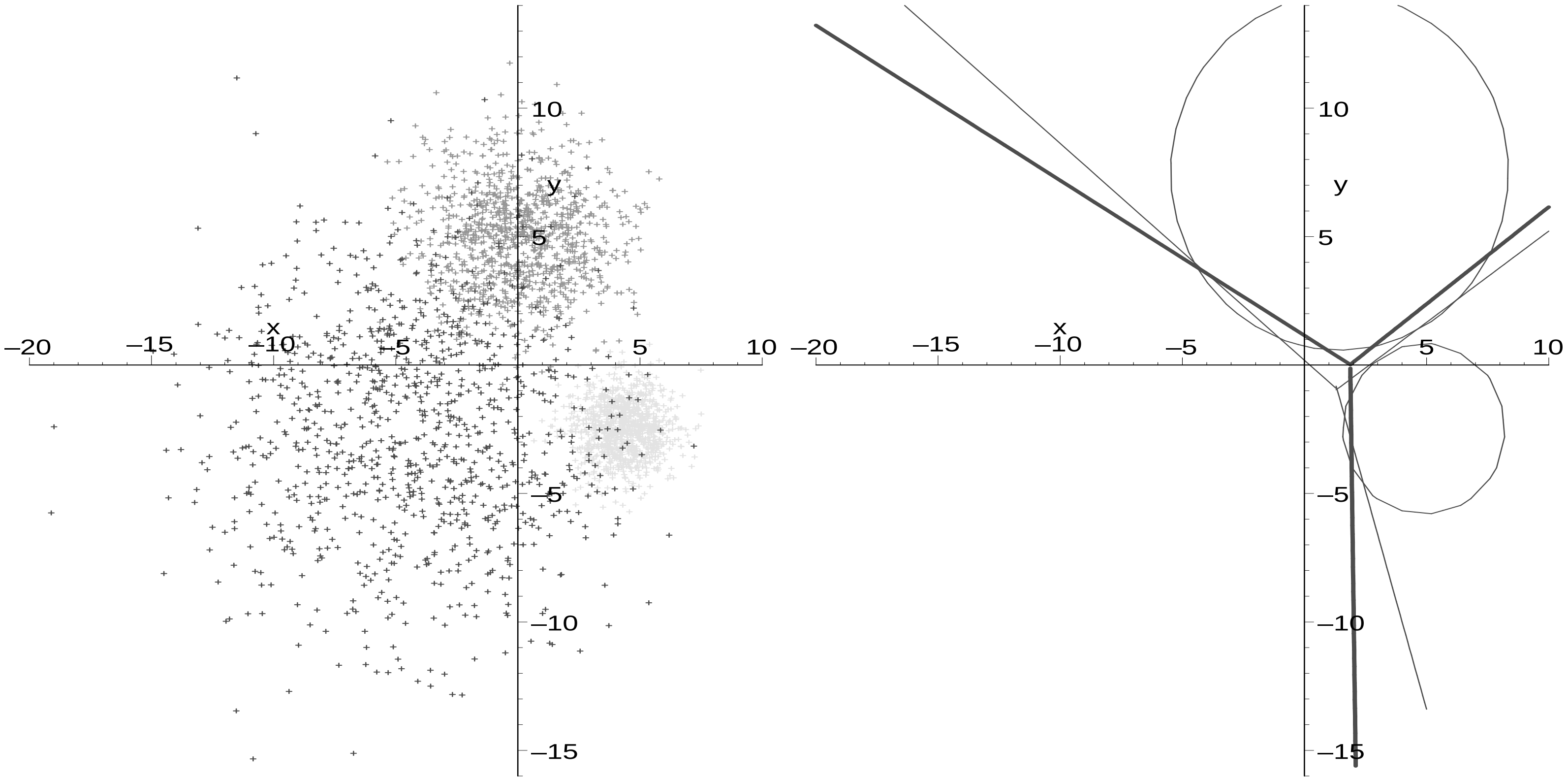

FIG. 6 – Separating 3 Gaussian-distributed categories in $\mathbb{R}^2$. **Left :** training set. **Right :** Bayes' classifier (circles), optimal linear classifier and boundaries computed by the linear M-SVM for the (estimated) optimal value of C (thick lines).

Experimental results

Classifier	Expected risks
Bayes' classifier	5.27 %
Selected M-SVM (polynomial kernel)	5.45 %
Opt. linear separator	5.85 %
Selected M-SVM (linear kernel)	6.30 %

Although the selected M-SVM exhibits the lowest expected risk among all the M-SVMs considered ($C \in \left[10^{-1}, 10^{2}\right]$), the variations observed are too small to be significant.

Additional experiments are underway with real-world problems and richer classes of functions (nonlinear kernels with finite dimensional feature spaces).

Conclusions and future work

Conclusions

- Dedicated solutions are needed to tune the hyperparameters of multi-class kernel machines
- Part of the standard (bi-class/regression) results can be extended simply... but not all !
- New tools/approaches are needed to derive tighter bounds on the expected risk

Future work

- Extension of the bounds based on the leave-one-out procedure
- Study of alternative measures of complexity (Donsker classes ...)
- Derivation of bounds in the framework of the empirical process theory