Multi-Class Support Vector Machines

Yann Guermeur

LORIA - CNRS

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

Summer School NN2008

July 4, 2008

Overview

Guaranteed risk for large margin multi-category classifiers

- Theoretical framework
- Basic uniform convergence result
- γ-Ψ-dimensions
- Generalized Sauer-Shelah lemma
- Nature and rate of convergence

Multi-class SVMs

- Multi-category classification with binary SVMs
- Class of functions implemented by the M-SVMs
- General formulation of the training algorithm
- Three main models of M-SVMs
- Some variants of the main models
- Margins and support vectors

Overview

Guaranteed risks for multi-class SVMs

- Bounds on the covering numbers
- Use of the Rademacher complexity

Model selection for multi-class SVMs

- Algorithms fitting the entire regularization path
- Bounds on the leave-one-out cross-validation error

Conclusions and open problems

Hypotheses and goals

Characterization of the problem

- Study of the connection between objects $x \in \mathcal{X}$ and their categories $y \in \mathcal{Y} = [\![1, Q]\!]$
- Hypothesis: existence of a $\mathcal{X} \times \mathcal{Y}$-valued random pair (X, Y) distributed according to a probability measure P
- Problem: the joint probability measure P is unknown

What is available

- $D_m = ((X_i, Y_i))_{1 \leq i \leq m}$: i.i.d. m-sample from (X, Y)
- $\mathcal{G}$: class of functions g, from $\mathcal{X}$ into $\mathbb{R}^Q$ ($\mathcal{F}$: class of decision rules f, from $\mathcal{X}$ into $\mathcal{Y} \bigcup \{*\}$)
  $f(x) = \operatorname{argmax}_{1 \leq k \leq Q} g_k(x)$ or $f(x) = *$, in case of ex æquo

The goal

- ℓ, loss function: $\ell(y, g(x)) = \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}}$ ($\ell(y, f(x)) = \mathbb{1}_{\{f(x) \neq y\}}$)
- Selection of a function g^* minimizing over $\mathcal{G}$ the risk

$$R(g) = \mathbb{E}\left[\ell\left(Y, g\left(X\right)\right)\right] = P(f(X) \neq Y)$$

Multi-class margin and margin risk

Definition 1 (Function M) *Let M be the function from $\mathbb{R}^Q \times [\![1, Q]\!]$ into $\mathbb{R}$ given by:*

$$\forall (v, k) \in \mathbb{R}^Q \times [\![1, Q]\!], \; M(v, k) = \frac{1}{2}\left(v_k - \max_{l \neq k} v_l\right)$$

$M(v, \cdot) = \max_{1 \leq k \leq Q} M(v, k)$

Definition 2 (Multi-class margin of g on the example (x, y))

$$\forall (g, x, y) \in \mathcal{G} \times \mathcal{X} \times \mathcal{Y}, \;\; \mathcal{M}(g, x, y) = M\left(g(x), y\right)$$

Definition 3 (Operators Δ and Δ^*) $g = (g_k)_{1 \leq k \leq Q} \in \mathcal{G}$

- *The function $\Delta g = (\Delta g_k)_{1 \leq k \leq Q}$, from $\mathcal{X}$ into $\mathbb{R}^Q$, is given by:*

$$\forall x \in \mathcal{X}, \;\; \Delta g(x) = \left(M\left(g(x), k\right)\right)_{1 \leq k \leq Q}$$

- *The function $\Delta^* g = (\Delta^* g_k)_{1 \leq k \leq Q}$, from $\mathcal{X}$ into $\mathbb{R}^Q$, is given by:*

$$\forall x \in \mathcal{X}, \;\; \Delta^* g(x) = \left(\mathrm{sign}\left(\Delta g_k(x)\right) \cdot M\left(g(x), \cdot\right)\right)_{1 \leq k \leq Q}$$

Multi-class margin and margin risk

$\Delta^{\#}$ replaces Δ and Δ^{*} in the formulas that hold true for both operators (e.g., $R(g) = \mathbb{E}\left[\mathbb{1}_{\{\Delta^{\#} g_Y(X) \leq 0\}}\right]$)

Definition 4 (Margin risk) *Let* $\gamma \in \mathbb{R}_+^*$. *The* risk with margin γ *of* g *is defined as:*

$$R_\gamma(g) = \mathbb{E}\left[\mathbb{1}_{\{\Delta^{\#} g_Y(X) < \gamma\}}\right] = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{\Delta^{\#} g_y(x) < \gamma\}} dP(x, y)$$

Empirical risk with margin γ:

$$R_{\gamma,m}(g) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{\Delta^{\#} g_{Y_i}(X_i) < \gamma\right\}}$$

Class of functions of interest: $\Delta_\gamma^{\#} \mathcal{G}$

For $\epsilon \in \mathbb{R}_+^*$, let $\pi_\epsilon : \mathbb{R} \to [-\epsilon, \epsilon]$ be the linear squashing function defined as:

$$\pi_\epsilon(t) = \operatorname{sign}(t) \cdot \min\left\{|t|, \epsilon\right\}$$

$$\Delta_\gamma^{\#} g = \left(\Delta_\gamma^{\#} g_k\right)_{1 \leq k \leq Q}, \quad \Delta_\gamma^{\#} g_k = \pi_\gamma \circ \Delta^{\#} g_k, \quad \Delta_\gamma^{\#} \mathcal{G} = \left\{\Delta_\gamma^{\#} g : \; g \in \mathcal{G}\right\}$$

Capacity measure of $\Delta_{\gamma}^{\#}\mathcal{G}$: covering numbers

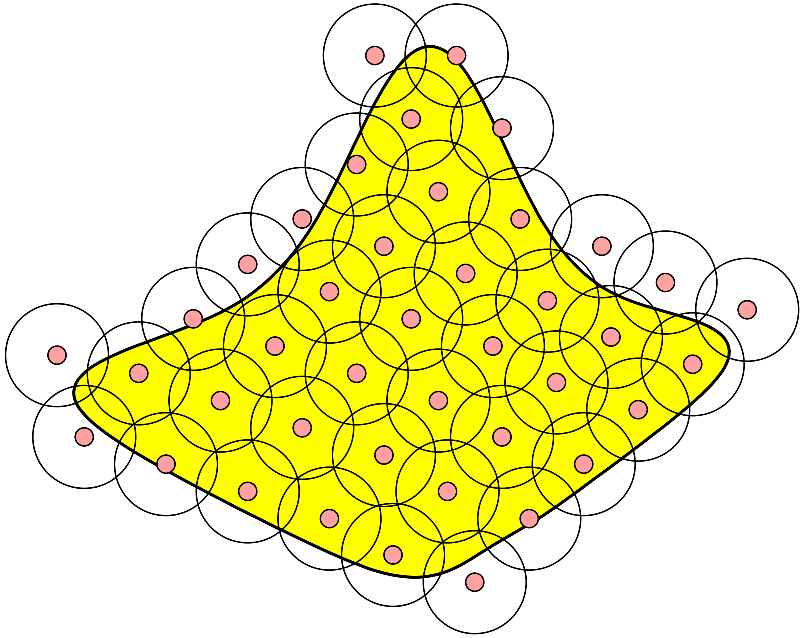

Figure 1: ϵ-*net* and ϵ-*cover* of a set E' in a pseudo-metric space (E, ρ)

Definition 5 (Covering numbers)

$\mathcal{N}(\epsilon, E', \rho)$: *minimal number of open balls of radius* ϵ *needed to cover* E' *(or* $+\infty$*)*

$\mathcal{N}^{(p)}(\epsilon, E', \rho)$: *the* ϵ*-nets considered are included in* E' *(proper to* E'*)*

Basic uniform convergence result
Classes of indicator functions

Theorem 1 (Guaranteed risk, Vapnik, 1998) *Let $\mathcal{F}$ be a class of indicator functions on a set $\mathcal{X}$. Let $N\left(\mathcal{F}, (X_i)_{1 \leq i \leq n}\right)$ be the number of different functions (dichotomies) that this class can implement on $(X_i)_{1 \leq i \leq n}$ and $\delta \in (0,1)$. With probability at least $1-\delta$, the risk of any function f in $\mathcal{F}$ is bounded from above as follows:*

$$R(f) \leq R_m(f) + \sqrt{\frac{1}{m}\left(\ln\left(\mathbb{E}N\left(\mathcal{F}, (X_i)_{1 \leq i \leq 2m}\right)\right) + \ln\left(\frac{4}{\delta}\right)\right)} + \frac{1}{m}.$$

$\ln\left(\mathbb{E}N\left(\mathcal{F}, (X_i)_{1 \leq i \leq 2m}\right)\right)$ is the *annealed entropy* of $\mathcal{F}$ on the sample $(X_i)_{1 \leq i \leq 2m}$.

Basic uniform convergence result
Classes of functions $\mathcal{G}$ (taking values in $\mathbb{R}^Q$)

Definition 6 (Pseudo-metric d_{x^n}) *Let $n \in \mathbb{N}^*$. For a sequence $x^n = (x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$, define the pseudo-metric d_{x^n} on $\mathcal{G}$ as:*

$$\forall (g, g') \in \mathcal{G}^2, \ d_{x^n}(g, g') = \max_{1 \leq i \leq n} \left\| g(x_i) - g'(x_i) \right\|_\infty .$$

For $\epsilon \in \mathbb{R}_+^*$, let $\mathcal{N}(\epsilon, \mathcal{G}, n) = \sup_{x^n \in \mathcal{X}^n} \mathcal{N}(\epsilon, \mathcal{G}, d_{x^n})$.

Theorem 2 (Guaranteed risk) *Let $\mathcal{G}$ be the class of functions that a large margin Q-category classifier on a domain $\mathcal{X}$ can implement. Let $\Gamma \in \mathbb{R}_+^*$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$, for every value of γ in $(0, \Gamma]$, the risk of any function g in $\mathcal{G}$ is bounded from above by:*

$$R(g) \leq R_{\gamma,m}(g) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}\left(\gamma/4, \Delta_\gamma^{\#}\mathcal{G}, 2m\right)\right) + \ln\left(\frac{2\Gamma}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

Growth function

Definition 7 (Growth function, Vapnik & Chervonenkis, 1971) *Let $\mathcal{F}$ be a class of indicator functions on a domain $\mathcal{X}$. For $n \in \mathbb{N}^*$, let $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ be a subset of $\mathcal{X}$ of cardinality n. Then, the* growth function *of $\mathcal{F}$, $\Pi_{\mathcal{F}}$, is defined by:*

$$\forall n \in \mathbb{N}^*, \ \Pi_{\mathcal{F}}(n) = \sup_{s_{\mathcal{X}^n} \subset \mathcal{X}} N(\mathcal{F}, s_{\mathcal{X}^n}).$$

Remark 1 *Some authors use the alternative definition:*

$$\forall n \in \mathbb{N}^*, \ \Pi_{\mathcal{F}}(n) = \ln\left(\sup_{s_{\mathcal{X}^n} \subset \mathcal{X}} N(\mathcal{F}, s_{\mathcal{X}^n})\right).$$

Remark 2 *In contrast with the annealed entropy, the growth function is distribution-free.*

VC dimension

Definition 8 (VC dimension, Vapnik & Chervonenkis, 1971) *Let $\mathcal{F}$ be a class of indicator functions on a domain $\mathcal{X}$. A subset $s_{\mathcal{X}^n} = \{x_i \ : 1 \leq i \leq n\}$ of $\mathcal{X}$ is said to be* shattered *by $\mathcal{F}$ if for each vector v_y in $\{1, 1\}^n$, there is a function f_y in $\mathcal{F}$ satisfying*

$$\left(f_y\left(x_i\right)\right)_{1 \leq i \leq n} = v_y.$$

The VC dimension *of $\mathcal{F}$, denoted by VC-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{X}$ shattered by $\mathcal{F}$, if this cardinality is finite. If no such maximum exists, $\mathcal{F}$ is said to have infinite VC dimension.*

Remark 3 *VC-dim$(\mathcal{F}) = d$ if and only if $\Pi_{\mathcal{F}}(d) = 2^d$ and $\Pi_{\mathcal{F}}(d+1) < 2^{d+1}$.*

Ψ-dimensions

Definition 9 (Ψ-dimensions, Ben-David *et al.*, 1995) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in the finite set $[\![1, Q]\!]$. Let Ψ be a family of mappings ψ from $[\![1, Q]\!]$ into $\{-1, 1, *\}$, where $*$ is thought of as a null element. A subset $s_{\mathcal{X}^n} = \{x_i \ : 1 \leq i \leq n\}$ of $\mathcal{X}$ is said to be* Ψ-shattered *by $\mathcal{F}$ if there is a mapping $\psi^n = \left(\psi^{(i)}\right)_{1 \leq i \leq n}$ in Ψ^n such that for each vector v_y in $\{-1, 1\}^n$, there is a function f_y in $\mathcal{F}$ satisfying*

$$\left(\psi^{(i)} \circ f_y(x_i)\right)_{1 \leq i \leq n} = v_y.$$

The Ψ-dimension *of $\mathcal{F}$, denoted by Ψ-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{X}$ Ψ-shattered by $\mathcal{F}$, if this cardinality is finite. If no such maximum exists, $\mathcal{F}$ is said to have infinite Ψ-dimension.*

Remark 4 *Let $\mathcal{F}$ and Ψ be defined as above. Extending the definition of the VC dimension so that it applies to classes of functions taking values in $\{-1, 1, *\}$, which has no incidence in practice, the following proposition holds true:*

$$\Psi\text{-}dim(\mathcal{F}) = VC\text{-}dim\left(\{(x, \psi) \mapsto \psi \circ f(x) : f \in \mathcal{F}, \psi \in \Psi\}\right).$$

Main examples of Ψ-dimensions

Definition 10 (Graph dimension, Dudley, 1987; Natarajan, 1989) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in $[\![1, Q]\!]$. The* graph dimension *of $\mathcal{F}$, G-dim($\mathcal{F}$), is the Ψ-dimension of $\mathcal{F}$ in the specific case where $\Psi = \{\psi_k : 1 \leq k \leq Q\}$, such that ψ_k takes the value 1 if its argument is equal to k and the value -1 otherwise. Reformulated in the context of multi-category classification, the functions ψ_k are the indicator functions of the categories.*

Definition 11 (Natarajan dimension, Natarajan, 1989) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in $[\![1, Q]\!]$. The* Natarajan dimension *of $\mathcal{F}$, N-dim($\mathcal{F}$), is the Ψ-dimension of $\mathcal{F}$ in the specific case where $\Psi = \{\psi_{k,l} : 1 \leq k \neq l \leq Q\}$, such that $\psi_{k,l}$ takes the value 1 if its argument is equal to k, the value -1 if its argument is equal to l, and $*$ otherwise.*

Remark 5 *The definition of the graph dimension is inspired from the one-against-all decomposition method whereas the definition of the Natarajan dimension is inspired from the one-against-one decomposition method.*

Fat-shattering or γ dimension

Definition 12 (Fat-shattering dimension, Kearns & Schapire, 1994) *Let $\mathcal{G}$ be a class of real-valued functions on a set $\mathcal{X}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ of $\mathcal{X}$ is said to be* γ-shattered *by $\mathcal{G}$ if there is a vector $v_b = (b_i)$ in $\mathbb{R}^n$ such that, for each vector $v_y = (y_i)$ in $\{-1, 1\}^n$, there is a function g_y in $\mathcal{G}$ satisfying*

$$\forall i \in [\![1, n]\!],\ y_i \left(g_y(x_i) - b_i\right) \geq \gamma.$$

The fat-shattering dimension with margin γ, *or P_γ* dimension, *of the class $\mathcal{G}$, P_γ-dim$(\mathcal{G})$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-shattered by $\mathcal{G}$, if this cardinality is finite. If no such maximum exists, $\mathcal{G}$ is said to have infinite P_γ dimension.*

γ-Ψ-dimensions

Let $\wedge$ denote the conjunction of two events.

Definition 13 (γ-Ψ-dimensions) *Let $\mathcal{G}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. Let Ψ be a family of mappings ψ from $[\![1, Q]\!]$ into $\{-1, 1, *\}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ of $\mathcal{X}$ is said to be γ-Ψ-*shattered *(Ψ-*shattered with margin* γ) by $\Delta^{\#}\mathcal{G}$ if there is a mapping $\psi^n = \left(\psi^{(i)}\right)_{1 \leq i \leq n}$ in Ψ^n and a vector $v_b = (b_i)$ in $\mathbb{R}^n$ such that, for each vector $v_y = (y_i)$ in $\{-1, 1\}^n$, there is a function g_y in $\mathcal{G}$ satisfying*

$$\forall i \in [\![1, n]\!], \begin{cases} \text{if } y_i = \quad 1, & \exists k : \psi^{(i)}(k) = \quad 1 \quad \wedge \ \Delta^{\#} g_{y,k}(x_i) - b_i \geq \gamma \\ \text{if } y_i = -1, & \exists l : \psi^{(i)}(l) = -1 \quad \wedge \ \Delta^{\#} g_{y,l}(x_i) + b_i \geq \gamma \end{cases} .$$

The γ-Ψ-dimension*, or* Ψ-dimension with margin γ*, of $\Delta^{\#}\mathcal{G}$, denoted by Ψ-dim$(\Delta^{\#}\mathcal{G}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-Ψ-shattered by $\Delta^{\#}\mathcal{G}$, if this cardinality is finite. If no such maximum exists, $\Delta^{\#}\mathcal{G}$ is said to have infinite γ-Ψ-dimension.*

This definition simplifies into the one of the fat-shattering dimension when $Q = 2$.

Natarajan dimension with margin γ

Definition 14 (Natarajan dimension with margin γ) *Let $\mathcal{G}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ of $\mathcal{X}$ is said to be* γ-N-shattered *(*N-shattered with margin γ*) by $\Delta^{\#}\mathcal{G}$ if there is a set*

$$I(s_{\mathcal{X}^n}) = \{(i_1(x_i), i_2(x_i)) : 1 \leq i \leq n\}$$

of n couples of distinct indexes in $[\![1, Q]\!]$ and a vector $v_b = (b_i)$ in $\mathbb{R}^n$ such that, for each vector $v_y = (y_i)$ in $\{-1, 1\}^n$, there is a function g_y in $\mathcal{G}$ satisfying

$$\forall i \in [\![1, n]\!], \begin{cases} \text{if } y_i = \quad 1, & \Delta^{\#} g_{y, i_1(x_i)}(x_i) - b_i \geq \gamma \\ \text{if } y_i = -1, & \Delta^{\#} g_{y, i_2(x_i)}(x_i) + b_i \geq \gamma \end{cases} .$$

The Natarajan dimension with margin γ *of the class $\Delta^{\#}\mathcal{G}$, N-dim$(\Delta^{\#}\mathcal{G}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-N-shattered by $\Delta^{\#}\mathcal{G}$, if this cardinality is finite. If no such maximum exists, $\Delta^{\#}\mathcal{G}$ is said to have infinite Natarajan dimension with margin γ.*

Sauer-Shelah lemma
(Classes of indicator functions)

Lemma 1 (Vapnik & Chervonenkis, 1971; Sauer, 1972; Shelah, 1972) *Let $\mathcal{F}$ be a class of indicator functions on a set $\mathcal{X}$ and let $\Pi_{\mathcal{F}}$ be its growth function. If its VC dimension d is finite, then for $n \geq d$,*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} C_n^i < \left(\frac{en}{d}\right)^d$$

where e is the base of the natural logarithm.

Generalized Sauer-Shelah lemma
Classes of functions from $\mathcal{X}$ into $[\![1, Q]\!]$

Lemma 2 (Haussler & Long, 1995) *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ into $[\![1, Q]\!]$ and let $\Pi_{\mathcal{F}}$ be its growth function. If its Natarajan dimension d is finite, then for $n \geq d$,*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} C_n^i \left(C_{Q+1}^2\right)^i < \left(\frac{(Q+1)^2 e n}{2d}\right)^d .$$

Generalized Sauer-Shelah lemma
Classes of real-valued functions

Lemma 3 (Alon *et al.*, 1997) *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[0,1]$. For every value of ϵ in $(0,1]$ and every integer value of n satisfying $n \geq P_{\epsilon/4}$-dim$(\mathcal{G})$, the following bound is true:*

$$\mathcal{N}(\epsilon, \mathcal{G}, n) < 2 \left(\frac{4n}{\epsilon^2} \right)^{d \log_2(2en/(d\epsilon))}$$

where $d = P_{\epsilon/4}$-dim$(\mathcal{G})$.

Generalized Sauer-Shelah lemma
Classes of functions from $\mathcal{X}$ into $\mathbb{R}^Q$

Lemma 4 *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$. For every value of ϵ in $(0, M_{\mathcal{G}}]$ and every integer value of n satisfying $n \geq$ N-dim$(\Delta\mathcal{G}, \epsilon/6)$, the following bound is true:*

$$\mathcal{N}^{(p)}(\epsilon, \Delta^*\mathcal{G}, n) < 2\left(n\, Q^2(Q-1)\left\lfloor \frac{3M_{\mathcal{G}}}{\epsilon} \right\rfloor^2\right)^{\left\lceil d \log_2\left(enC_Q^2\left(2\left\lfloor \frac{3M_{\mathcal{G}}}{\epsilon}\right\rfloor - 1\right)/d\right)\right\rceil}$$

where $d =$ N-dim$(\Delta\mathcal{G}, \epsilon/6)$.

The proof does not hold true anymore if the operator Δ^* is replaced with the operator Δ.

Nature and rate of convergence

Theorem 3 *Let $\mathcal{G}$ be the class of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$ that a large margin Q-category classifier can implement. Let $\delta \in (0,1)$. With probability at least $1-\delta$, uniformly for every value of γ in $(0, M_{\mathcal{G}}]$, the risk of any function g in $\mathcal{G}$ is bounded from above by:*

$$R(g) \leq R_{\gamma,m}(g) +$$

$$\sqrt{\frac{2}{m}\left(\ln\left(4\left(2m\,Q^2(Q-1)\left\lfloor\frac{12M_{\mathcal{G}}}{\gamma}\right\rfloor^2\right)^{\left\lceil d\log_2\left(emQ(Q-1)\left(2\left\lfloor\frac{12M_{\mathcal{G}}}{\gamma}\right\rfloor-1\right)/d\right)\right\rceil}\right) + \ln\left(\frac{2M_{\mathcal{G}}}{\gamma\delta}\right)\right) + \frac{1}{m}}$$

where $d = N\text{-}dim(\Delta\mathcal{G}, \gamma/24)$.

$$R(g) \leq R_{\gamma,m}(g) + c\ln(m)\sqrt{\frac{d}{m}}$$

Proposition 1 (Almost sure uniform convergences)

$$\lim_{m\to+\infty}\sup_P \mathbb{P}\left(\sup_{n\geq m}\sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,n}(g)\right) > \epsilon\right) = 0 \quad \lim_{m\to+\infty}\sup_P \mathbb{P}\left(\sup_{n\geq m}\sup_{g\in\mathcal{G}}\left|R_\gamma(g) - R_{\gamma,n}(g)\right| > \epsilon\right) = 0$$

Multi-category classification with binary SVMs

One-against-all method (Rifkin & Klautau, 2004)

- Q SVMs: the k-th one distinguishes category k from the $Q-1$ other ones
- Decision rule: "winner-takes-all"

One-against-one method/pairwise classification (Fürnkranz, 2002)

- $\binom{Q}{2}$ SVMs: one for each pair of classes
- Decision rule: "max-wins voting"

Use of error correcting output codes (ECOC) (Allwein *et al.*, 2000)

- $M = (m_{kl}) \in \mathcal{M}_{Q,N}(\{-1, 0, 1\})$: "coding matrix"
- N SVMs: one for each of the dichotomies defined by the columns of M
- Decision rule: computation of a loss function

Reproducing kernel Hilbert space

Let $\mathcal{X}$ be a space and $(H, \langle \cdot, \cdot \rangle_H)$ a Hilbert space of functions on $\mathcal{X}$ ($H \subset \mathbb{R}^{\mathcal{X}}$).

Definition 15 (Reproducing kernel, Aronszajn, 1950) *Let κ be a function from $\mathcal{X}^2$ into $\mathbb{R}$. $\forall x \in \mathcal{X}$, let κ_x be the function from $\mathcal{X}$ into $\mathbb{R}$ given by $\kappa_x : t \mapsto \kappa(x, t)$. κ is a* reproducing kernel *of H if and only if:*

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

Definition 16 (Reproducing kernel Hilbert space) *If H possesses a reproducing kernel, it is called a* reproducing kernel Hilbert space *(RKHS) or a* proper Hilbert space.

Positive semidefinite kernel and RKHS

Definition 17 (Positive semidefinite (positive type) kernel) *A function κ from $\mathcal{X}^2$ into $\mathbb{R}$ is called a* positive semidefinite kernel *(or a* positive type kernel*) if*

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

Theorem 4 (Moore-Aronszajn) *Let κ be a positive semidefinite kernel on $\mathcal{X}^2$. There exists only one Hilbert space $(H, \langle \cdot, \cdot \rangle_H)$ of functions on $\mathcal{X}$ with κ as reproducing kernel.*

Building a M-SVM starting from a kernel

Basic class of functions

Let κ be a positive semidefinite kernel on $\mathcal{X}$ and let $(H_\kappa, \langle \cdot, \cdot \rangle_{H_\kappa})$ be the corresponding RKHS.

Let $\bar{\mathcal{H}} = (H_\kappa, \langle \cdot, \cdot \rangle_{H_\kappa})^Q$ and $\mathcal{H} = ((H_\kappa, \langle \cdot, \cdot \rangle_{H_\kappa}) + \{1\})^Q$.

$\mathcal{H}$: class of functions $h = (h_k)_{1 \leq k \leq Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$ such that:

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

with $\{x_{ik} : 1 \leq i \leq m_k\} \subset \mathcal{X}$, $(\beta_{ik})_{1 \leq i \leq m_k} \in \mathbb{R}^{m_k}$ and $b_k \in \mathbb{R}$, 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 kernel

Class of functions implemented

convex subset of $\mathcal{H}$ (defined by constraints on an affine subspace)

Basic class of functions

An affine model in the feature space

Theorem 5 (Mercer's theorem) *For all Mercer kernel κ, there exists a map Φ such that:*

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

where $\langle \cdot, \cdot \rangle$ is the dot product of the ℓ_2 space.

Φ is called a *feature map*. Let $\Phi(\mathcal{X}) = \{\Phi(x) : x \in \mathcal{X}\}$.

A *feature space* is any of the Hilbert spaces $\left(E_{\Phi(\mathcal{X})}, \langle \cdot, \cdot \rangle\right)$ spanned by the $\Phi(\mathcal{X})$.

$\Longrightarrow \mathcal{H}$ can be seen as a class of multivariate affine functions on $\Phi(\mathcal{X})$

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

$\mathbf{w} = (w_k)_{1 \leq k \leq Q} \in E^Q_{\Phi(\mathcal{X})}$, $\mathbf{b} = (b_k)_{1 \leq k \leq Q} \in \mathbb{R}^Q$

Basic class of functions

Putting things the other way round: the "kernel trick"

Norms on $\bar{\mathcal{H}}$ and $E^Q_{\Phi(\mathcal{X})}$

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

$$\|\mathbf{w}\|_\infty = \max_{1 \leq k \leq Q} \|w_k\|$$

$Q \geq 3$: multi-class support vector machines

$((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times [\![1, Q]\!])^m$: *training set*

$\ell_{\text{M-SVM}}$: convex loss function (built around the *hinge loss*)

M-SVM: solution of a convex (quadratic) programming problem

Problem 1

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

$$s.t. \ \textstyle\sum_{k=1}^{Q} h_k = 0$$

Representer theorem

This theorem states that training (solving Problem 1) amounts to finding the values of the coefficients β_{ik} in

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

(the values of the “biases” b_k are deduced by application of the Kuhn-Tucker conditions).

A general framework that encompasses the bi-class case

$((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times \{-1, 1\})^m$: training set

$h = (h_1, h_2) = (h_1, -h_1)$, $\tilde{h}(x) = h_1(x) = \Delta^{\#} h_1(x) = \frac{1}{2} \left(\langle w_1 - w_2, \Phi(x) \rangle + b_1 - b_2 \right)$

$\ell_{\mathrm{SVM}}(y, \tilde{h}(x)) = \left(1 - y\tilde{h}(x)\right)_+$ (hinge loss)

SVM: solution of a convex (quadratic) programming problem

Problem 2

$$\min_{\tilde{h} \in \tilde{\mathcal{H}}} \left\{ \sum_{i=1}^{m} \ell_{SVM}\left(y_i, \tilde{h}(x_i)\right) + \lambda \left\| \bar{\tilde{h}} \right\|_{H_\kappa}^2 \right\}$$

Representer theorem

This theorem states that training (solving Problem 2) amounts to finding the values of the coefficients β_i in

$$\tilde{h}(\cdot) = \sum_{i=1}^{m} \beta_i \kappa(x_i, \cdot) + b$$

(the value of the “bias” b is deduced by application of the Kuhn-Tucker conditions).

Hard margin M-SVMs and geometrical margins

Geometrical margins

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

$$\forall (k,l), \;\; 1 \leq k < l \leq Q,$$

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

$$\forall (k,l), \;\; 1 \leq k < l \leq Q, \;\; \gamma_{kl} = d_{\text{M-SVM}} \frac{1 + d_{\text{M-SVM},kl}}{\| w_k - w_l \|}$$

Connection between the penalizer and the geometrical margins

$$\left(\sum_{k<l} \| w_k - w_l \|^2 = Q \sum_{k=1}^{Q} \| w_k \|^2 - \left\| \sum_{k=1}^{Q} w_k \right\|^2 \right) \wedge \sum_{k=1}^{Q} w_k = 0 \Longrightarrow$$

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

M-SVM of Weston and Watkins

Training algorithm - primal formulation

Problem 3 (M-SVM1, Vapnik & Blanz, 1998; Weston & Watkins, 1998; ...)

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

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

Remark 6 *The constraint* $\sum_{k=1}^{Q} h_k = 0$ *is implicit.*

M-SVM of Weston and Watkins

Training algorithm - dual formulation

α_{ik}: Lagrange multiplier corresponding to the constraint $\langle w_{y_i} - w_k, \Phi(x_i) \rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}$

$\alpha = (\alpha_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$, $(\alpha_{iy_i})_{1 \leq i \leq m} = 0$

Problem 4 (M-SVM1)

$$\min_{\alpha} \left\{ \frac{1}{2} \alpha^T H_{WW} \alpha - 1_{Qm}^T \alpha \right\}$$

$$s.t. \begin{cases} 0 \leq \alpha_{ik} \leq C, & (1 \leq i \leq m),\ (1 \leq k \neq y_i \leq Q) \\ \sum_{i: y_i = k} \sum_{l=1}^{Q} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0, & (1 \leq k \leq Q-1) \end{cases}$$

$$H_{\mathrm{WW}} = \left(\left(\delta_{y_i, y_j} - \delta_{y_i, l} - \delta_{y_j, k} + \delta_{k,l} \right) \kappa(x_i, x_j) \right)_{1 \leq i,j \leq m, 1 \leq k,l \leq Q}$$

$$w_k^* = \sum_{i: y_i = k} \sum_{l=1}^{Q} \alpha_{il}^* \Phi(x_i) - \sum_{i=1}^{m} \alpha_{ik}^* \Phi(x_i) = \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\delta_{y_i, k} - \delta_{k,l} \right) \alpha_{il}^* \Phi(x_i)$$

M-SVM of Crammer and Singer

Training algorithm - primal formulation

Problem 5 (M-SVM2, Crammer & Singer, 2001)

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

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

Remark 7 *The constraint* $\sum_{k=1}^{Q} \bar{h}_k = 0$ *is implicit.*

M-SVM of Crammer and Singer

Training algorithm - dual formulation

α_{ik}: Lagrange multiplier corresponding to the constraint $\langle w_{y_i} - w_k, \Phi(x_i)\rangle + \delta_{y_i,k} \geq 1 - \xi_i$

$\alpha = (\alpha_{ik})_{1\leq i\leq m, 1\leq k\leq Q}$, $\delta = (\delta_{y_i,k})_{1\leq i\leq m, 1\leq k\leq Q}$

Problem 6 (M-SVM2)

$$\min_{\alpha} \left\{ \frac{1}{2}\alpha^T H_{WW} \alpha + \delta^T \alpha \right\}$$

$$s.t. \begin{cases} \alpha_{ik} \geq 0, & (1 \leq i \leq m),\ (1 \leq k \leq Q) \\ \sum_{k=1}^{Q} \alpha_{ik} = C, & (1 \leq i \leq m) \end{cases}$$

M-SVM of Lee, Lin and Wahba

Training algorithm - primal formulation

Problem 7 (M-SVM3, Lee *et al.*, 2004)

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

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

Result of consistency (Zhang, 2004; Tewari & Bartlett, 2007)

This M-SVM is the only one for which training is Bayes/Fisher consistent.

M-SVM of Lee, Lin and Wahba

Training algorithm - dual formulation

α_{ik}: Lagrange multiplier corresponding to the constraint $\langle w_k, \Phi(x_i) \rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}$

$\alpha = (\alpha_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$, $(\alpha_{iy_i})_{1 \leq i \leq m} = 0$

Problem 8 (M-SVM3)

$$\min_{\alpha} \left\{ \frac{1}{2} \alpha^T H_{LLW} \alpha - \frac{1}{Q-1} 1_{Qm}^T \alpha \right\}$$

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

$$H_{\mathrm{LLW}} = \left(\left(\delta_{k,l} - \frac{1}{Q} \right) \kappa(x_i, x_j) \right)_{1 \leq i,j \leq m, 1 \leq k,l \leq Q}$$

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

Use of different norms on $\mathbf{w}$

Problem 9 (ℓ_∞-norm M-SVM)

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

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

ℓ_1-norm M-SVM (Wang *et al.*, 2006)

$\kappa(x, x') = x^T x'$ $(\Phi = Id)$

Problem 10 (ℓ_1-norm M-SVM)

$$\min_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{m} \ell_{M\text{-}SVM}(y_i, h(x_i)) \right\}$$

$$s.t. \begin{cases} \sum_{k=1}^{Q} \|w_k\|_1 \leq K \\ \sum_{k=1}^{Q} h_k = 0 \end{cases}$$

Use of a different norm on ξ: quadratic loss M-SVMs

Definition 18 (Quadratic loss M-SVM) *A* quadratic loss M-SVM *is a M-SVM for which the empirical term of the objective function, $\|\xi\|_1$, is replaced by a quadratic form, $\xi^T M_\xi \xi$, where M_ξ is a symmetric positive semidefinite matrix.*

Definition 19 (M-SVM^2) *Variant of the M-SVM of Lee, Lin and Wahba corresponding to*

$$M_\xi = \left(\left(\delta_{k,l} - \frac{1}{Q} \right) \delta_{i,j} \right)_{1 \leq i,j \leq m, 1 \leq k,l \leq Q}.$$

Training algorithm of the M-SVM^2

Primal formulation

Problem 11 (M-SVM^2)

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \xi^T M_\xi \xi \right\}$$

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

Dual formulation

Problem 12 (M-SVM^2)

$$\min_{\alpha} \left\{ \frac{1}{2} \alpha^T \left(H_{LLW} + \frac{1}{2C} M_\xi \right) \alpha - \frac{1}{Q-1} 1_{Qm}^T \alpha \right\}$$

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

Margins and support vectors of a M-SVM

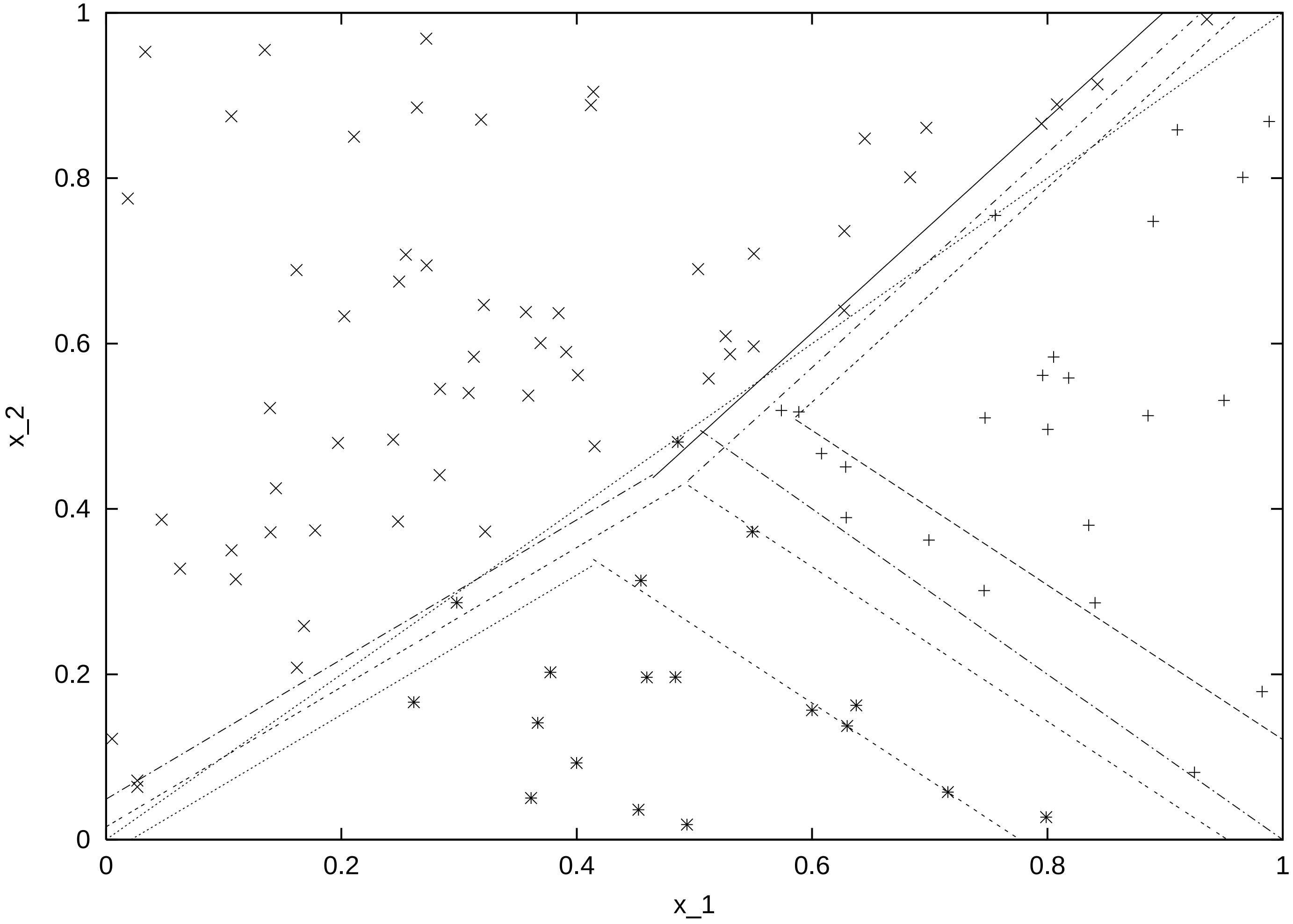

Figure 2: 3 categories linearly separable in $\mathbb{R}^2$

Margins and support vectors of a M-SVM

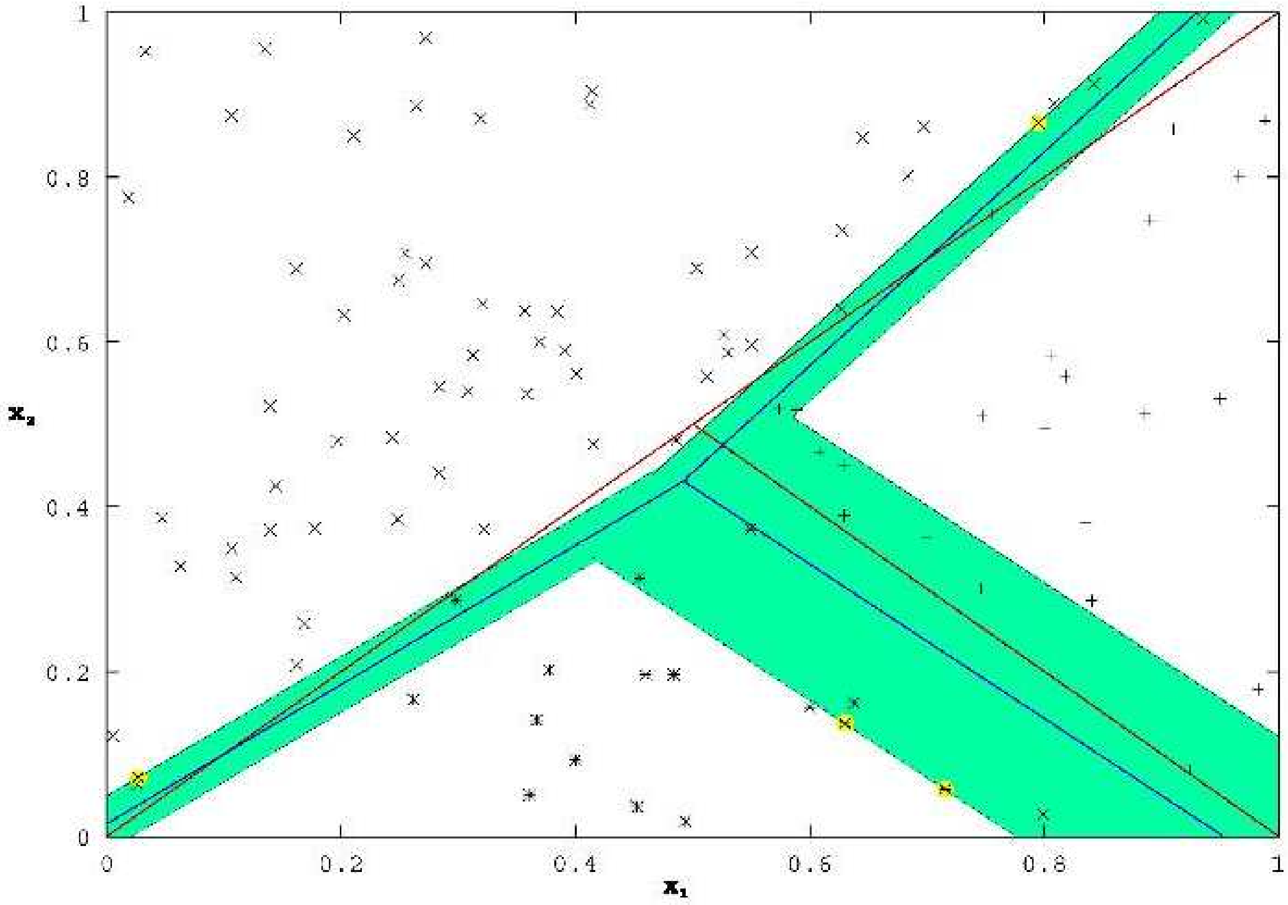

Figure 3: Separating hyperplanes and soft margins of a linear M-SVM1

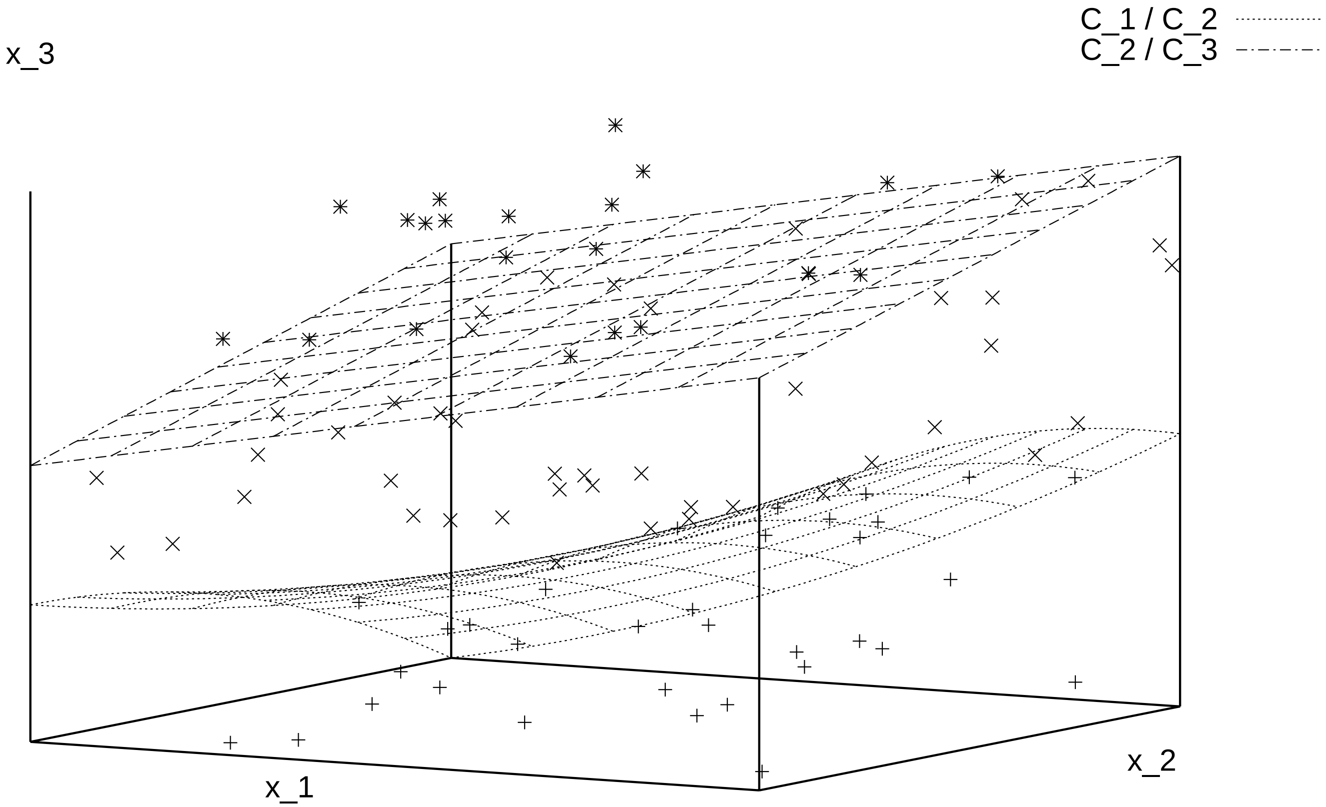

Figure 4: 3 categories non-linearly separable in $\mathbb{R}^3$

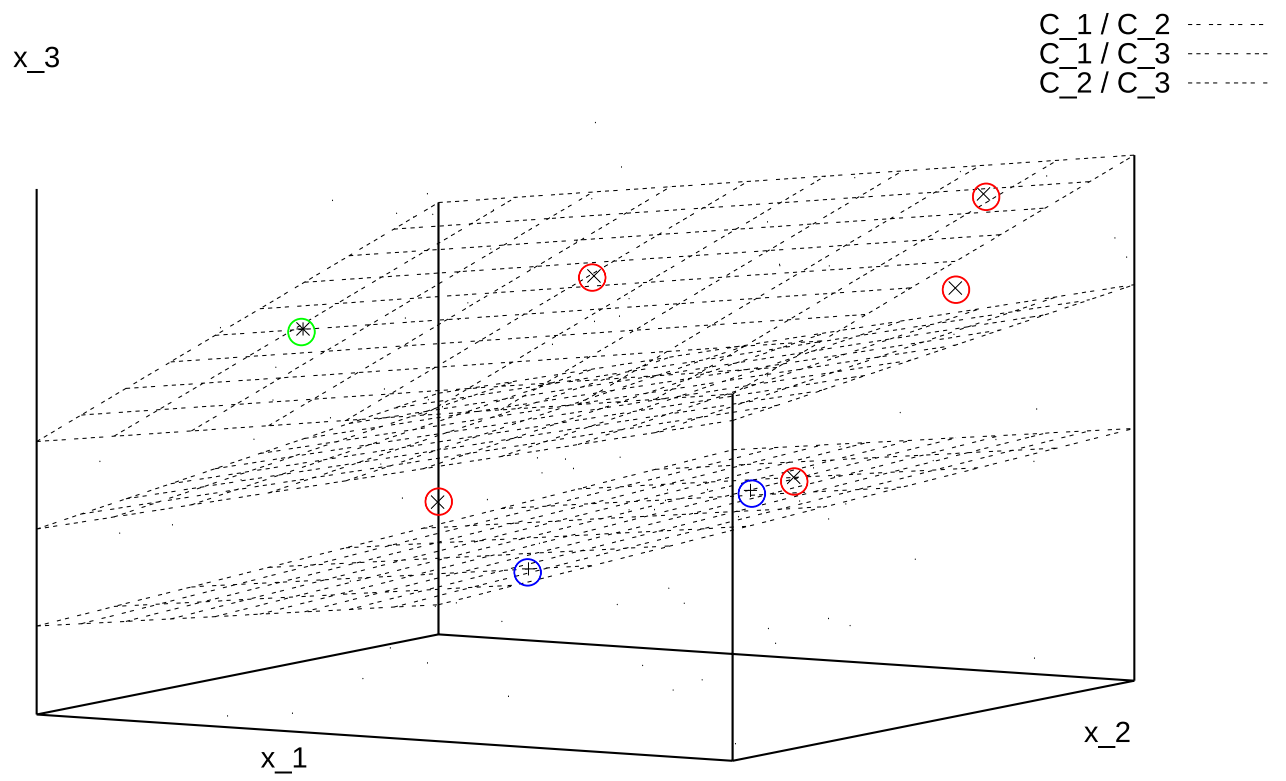

Figure 5: Separating hyperplanes and support vectors of a linear M-SVM1

Margin Natarajan dimension of the multi-class SVMs

Theorem 6 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under the hypothesis that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$, that the vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and that $\mathbf{b} = 0$. Then, for all $\epsilon \in \mathbb{R}_+^*$,*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \epsilon\right) \leq \binom{Q}{2}\left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2 .$$

The proof

- does not hold true anymore if the operator Δ is replaced by the operator Δ^*;
- calls for the use of the ℓ_∞-norm instead of the ℓ_2-norm (used by the penalizer);
- rests directly on the one-against-one decomposition scheme.

$$Q = 2 : \quad P_\epsilon\text{-dim}\left(H_\kappa\right) \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2$$

From covering numbers to entropy numbers

Definition 20 (Entropy numbers of a set) *Let (E, ρ) be a pseudo-metric space (or $(E, \|\cdot\|_E)$ a Banach space) and E' a bounded subset of E. Then, for $n \in \mathbb{N}^*$, the n-th entropy number of E', $\epsilon_n(E')$, is:*

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

Definition 21 (Entropy numbers of a bounded linear operator) *Let $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ be two Banach spaces. Let $\mathcal{L}(E, F)$ denote the Banach space of all (bounded linear) operators from $(E, \|\cdot\|_E)$ into $(F, \|\cdot\|_F)$ endowed with the norm:*
$\forall S \in \mathcal{L}(E, F), \|S\| = \sup_{e \in E : \|e\|_E = 1} \|S(e)\|_F$. The n-th entropy number of S is defined as

$$\epsilon_n(S) = \epsilon_n(S(U_E)).$$

From covering numbers to entropy numbers

Definition 22 (Evaluation operator) *For $n \in \mathbb{N}^*$, let $x^n \in \mathcal{X}^n$. The* evaluation operator S_{x^n} *on* $\bar{\mathcal{H}}$ *is defined as:*

$$\begin{array}{rccc} S_{x^n} : & \bar{\mathcal{H}} & \longrightarrow & \ell_\infty^{Qn} \\ & \bar{h} = (w_k)_{1 \leq k \leq Q} & \mapsto & S_{x^n}\left(\bar{h}\right) = \left(\langle w_k, \Phi(x_i)\rangle\right)_{1 \leq i \leq n,\ 1 \leq k \leq Q} \end{array}$$

Let $\mathcal{U}$ be the unit ball of $\bar{\mathcal{H}}$ in the ℓ_∞-norm ($\mathcal{U} = \left\{\bar{h} \in \bar{\mathcal{H}} : \|\mathbf{w}\|_\infty \leq 1\right\}$). The connection between $\mathcal{N}(\epsilon, \mathcal{U}, n)$ and the entropy numbers of S_{x^n} is provided by the following proposition:

Proposition 2 *Let $\epsilon \in \mathbb{R}_+^*$ and $n \in \mathbb{N}^*$.*

$$\sup_{x^n \in \mathcal{X}^n} \epsilon_p(S_{x^n}) \leq \epsilon \Longrightarrow \mathcal{N}(\epsilon, \mathcal{U}, n) \leq p.$$

Upper bound on the entropy numbers
Finite-dimensional feature space

Proposition 3 (Carl & Stephani, 1990) *Let E and F be Banach spaces and $S \in \mathfrak{L}(E, F)$. If S is of rank r, then for $n \in \mathbb{N}^*$,*

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

Theorem 7 *Let $\mathcal{H}$ be the class of functions that a Q-category M-SVM can implement under the hypothesis that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$, that the vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and $\mathbf{b} \in [-\beta, \beta]^Q$. If the dimensionality of the space $E_{\Phi(\mathcal{X})}$ is finite and equal to d, then for all $\gamma \in \mathbb{R}_+^*$,*

$$\mathcal{N}^{(p)}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m) \leq \left(2\left\lceil \frac{8\beta}{\gamma} \right\rceil + 1\right)^Q \cdot \left(\frac{64\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\gamma}\right)^{Qd}.$$

$$R(h) \leq R_{\gamma,m}(h) + O\left(\sqrt{\frac{1}{m}}\right)$$

Upper bound on the entropy numbers
Infinite-dimensional feature space

Theorem 8 (Maurey-Carl theorem, Carl & Stephani, 1990) *Let H be a Hilbert space and S an operator belonging to $\mathfrak{L}(\ell_1^n, H)$ or $\mathfrak{L}(H, \ell_\infty^n)$. Then, for each couple of integers (k, n) satisfying $1 \leq k \leq n$,*

$$e_k(S) \leq c \left(\frac{1}{k} \log_2 \left(1 + \frac{n}{k} \right) \right)^{1/2} \|S\|,$$

where the dyadic entropy number *$e_k(S)$ is equal to $\epsilon_{2^{k-1}}(S)$ and c is a universal constant.*

Theorem 9 *Let $\mathcal{H}$ be the class of functions that a Q-category M-SVM can implement under the hypothesis that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$, that the vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and $\mathbf{b} \in [-\beta, \beta]^Q$. Then, for all $\gamma \in \mathbb{R}_+^*$,*

$$\mathcal{N}^{(p)}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m) \leq \left(2 \left\lceil \frac{8\beta}{\gamma} \right\rceil + 1 \right)^Q \cdot 2^{\frac{16 c \Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\gamma} \sqrt{\frac{2Qm}{\ln(2)}} - 1}.$$

$$R(h) \leq R_{\gamma,m}(h) + O\left(\sqrt{\frac{1}{\sqrt{m}}} \right)$$

Basic probabilistic tools

Definition 23 (Rademacher average) *For $n \in \mathbb{N}^*$, let $\mathcal{A}$ be a bounded set of vectors $a = (a_i)_{1 \leq i \leq n}$ belonging to $\mathbb{R}^n$ and let $(\sigma_i)_{1 \leq i \leq n}$ be a Rademacher sequence. The* Rademacher average associated with $\mathcal{A}$, $\mathcal{R}_n(\mathcal{A})$, *is defined by:*

$$\mathcal{R}_n(\mathcal{A}) = \mathbb{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| .$$

Theorem 10 (Bounded differences inequality, McDiarmid, 1989) *Let $(T_i)_{1 \leq i \leq n}$ be a sequence of n independent random variables taking values in a set $\mathcal{T}$. Let g be a function from $\mathcal{T}^n$ into $\mathbb{R}$ such that there exists a sequence of nonnegative constants $(c_i)_{1 \leq i \leq n}$ satisfying:*

$$\forall i \in [\![1, n]\!], \sup_{(t_i)_{1 \leq i \leq n} \in \mathcal{T}^n, t_i' \in \mathcal{T}} |g(t_1, \ldots, t_n) - g(t_1, \ldots, t_{i-1}, t_i', t_{i+1}, \ldots, t_n)| \leq c_i .$$

Then, for all $\tau \in \mathbb{R}_+^*$, the random variable $g(T_1, \ldots, T_n)$ satisfies:

$$\mathbb{P}\left\{ g(T_1, \ldots, T_n) - \mathbb{E} g(T_1, \ldots, T_n) > \tau \right\} \leq e^{-\frac{2\tau^2}{c}}$$

$$\mathbb{P}\left\{ \mathbb{E} g(T_1, \ldots, T_n) - g(T_1, \ldots, T_n) > \tau \right\} \leq e^{-\frac{2\tau^2}{c}}$$

where $c = \sum_{i=1}^{n} c_i^2$.

Uniform convergence result

Convexified margin risk corresponding to the M-SVM of Crammer and Singer

$$\tilde{R}(h) = \mathbb{E}\left[\left(1 - \Delta h_Y\left(X\right)\right)_+\right]$$

Theorem 11 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$, that the vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and $\mathbf{b} = 0$. Let $K_{\bar{\mathcal{H}}} = \Lambda_w \Lambda_{\Phi(\mathcal{X})} + 1$ and $\delta \in (0,1)$. With probability at least $1-\delta$, the risk of any function $\bar{h}$ in $\bar{\mathcal{H}}$ is bounded from above by:*

$$R\left(\bar{h}\right) \leq \tilde{R}_m\left(\bar{h}\right) + \frac{4}{\sqrt{m}} + \frac{4Q(Q-1)\Lambda_w}{m}\sqrt{\sum_{i=1}^m \kappa\left(X_i, X_i\right)} + K_{\bar{\mathcal{H}}}\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}.$$

$$R\left(\bar{h}\right) \leq \tilde{R}_m\left(\bar{h}\right) + O\left(\sqrt{\frac{1}{m}}\right)$$

Radius-margin bound

Theorem 12 (Vapnik, 1998) *Let us consider a hard margin bi-class SVM. Let $\mathcal{L}_m$ be the number of errors that it makes in a leave-one-out cross-validation procedure and let $\gamma = \frac{1}{\|w\|}$ denote its geometrical margin. Then the following upper bound holds true:*

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

where $\mathcal{D}_m$ is the diameter of the smallest ball of the feature space containing the support vectors.

Radius-margin bound for the M-SVM of Weston and Watkins

$d_{\mathrm{WW}} = d_{\mathrm{CS}} = 1$

Theorem 13 *Let us consider a hard margin Q-category M-SVM of Weston and Watkins (or Crammer and Singer) 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:*

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

Constant K_{CV}

- The value of K_{CV} is obtained by solving as many QP problems as there are support vectors.
- For $Q = 2$, $K_{\mathrm{CV}} = 2$, and the bound reduces itself to the bi-class one.

Radius-margin bound for the M-SVM of Lee, Lin and Wahba

$d_{\mathrm{LLW}} = \frac{Q}{Q-1}$

Theorem 14 *Let us consider a hard margin Q-category 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:*

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

This bound does not reduce itself to the bi-class one for $Q = 2$.

Conclusions

Capacity measures of the classes of functions

- The γ-Ψ-dimensions play for the M-SVMs (and the MLPs!) the same role as the fat-shattering dimension for the bi-class SVMs.
- The current upper bounds on the covering numbers are suboptimal but in specific cases.
- If the use of the Rademacher complexity currently provides the sharpest bound, better bounds, adapted to the problem of interest, should result from implementing hybrid approaches.

Guaranteed risks

- These studies highlight the specific character of the multi-class case.
- Model selection should provide a touchstone to assess the different guaranteed risks derived.

Open problems and future work

Bounds on the risk of large margin multi-category classifiers

- Computation of a bound on the universal constant of the Maurey-Carl theorem
- Use of Dudley's method of chaining to improve the VC bound
- Derivation of dedicated PAC-Bayes bounds
- ...

Model selection for M-SVMs

- Assessment of the guaranteed risks and radius-margin bounds to select the value of the soft margin parameter C
- Integration in the applications implementing the M-SVMs of procedures choosing automatically the values of the hyperparameters