

**Risques garantis et sélection de modèle pour les SVM multi-classes,
application à la prédiction de la structure secondaire des protéines**

Yann Guermeur

LORIA - CNRS

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

AAFD 2008

Plan de la présentation

Risque garanti pour les modèles multi-classes à grande marge

- Cadre théorique
- Résultat de convergence uniforme de base
- γ - Ψ -dimensions
- Lemme de Sauer-Shelah généralisé

SVM multi-classes

- Famille de fonctions réalisables
- Trois modèles principaux

Risques garantis et sélection de modèle pour les SVM multi-classes

- Borne sur la dimension de Natarajan à marge
- Utilisation d'une moyenne de Rademacher
- Bornes sur l'erreur de validation croisée "leave-one-out" des SVM multi-classes

Application à la prédiction de la structure secondaire des protéines

- Un problème central en biologie structurale
- M-SVM dédiée à la prédiction de la structure secondaire

Hypothèses et objectif

Caractérisation du problème

- Etude du lien associant des objets $x \in \mathcal{X}$ à leurs catégories $y \in \mathcal{Y} = \{1, \dots, Q\}$
- Hypothèse : existence d'un couple aléatoire (X, Y) à valeurs dans $\mathcal{X} \times \mathcal{Y}$, distribué suivant une mesure de probabilité P
- Problème : la loi jointe de (X, Y) est inconnue

Ce dont on dispose

- $D_m = ((X_i, Y_i))_{1 \leq i \leq m}$: m -échantillon constitué de copies de (X, Y) indépendantes
- \mathcal{G} : famille de fonctions g , de \mathcal{X} dans \mathbb{R}^Q (\mathcal{F} : famille de règles de décision f , de \mathcal{X} dans $\mathcal{Y} \cup \{*\}$)
 $f(x) = \operatorname{argmax}_{1 \leq k \leq Q} g_k(x)$ ou $f(x) = *$, en cas d'ex æquo

Ce que l'on cherche à faire

- ℓ , fonction de perte : $\ell(y, g(x)) = \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}}$ ($\ell(y, f(x)) = \mathbb{1}_{\{f(x) \neq y\}}$)
- Sélection d'une fonction g^* minimisant sur \mathcal{G} le risque

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

Marge multi-classe et risque à marge

Définition 1 (Fonction M) Soit M la fonction de $\mathbb{R}^Q \times \{1, \dots, Q\}$ dans \mathbb{R} définie par :

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

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

Définition 2 (Marge multi-classe de g sur l'exemple (x, y))

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

Définition 3 (Opérateurs Δ et Δ^*) $g = (g_k)_{1 \leq k \leq Q}$: fonction de \mathcal{X} dans \mathbb{R}^Q

– La fonction $\Delta g = (\Delta g_k)_{1 \leq k \leq Q}$, de \mathcal{X} dans \mathbb{R}^Q , est donnée par :

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

– La fonction $\Delta^* g = (\Delta^* g_k)_{1 \leq k \leq Q}$, de \mathcal{X} dans \mathbb{R}^Q , est donnée par :

$$\forall x \in \mathcal{X}, \Delta^* g(x) = (\max \{ \Delta g_k(x), -M(g(x), \cdot) \})_{1 \leq k \leq Q}$$

Marge multi-classe et risque à marge

$\Delta^\#$ remplace Δ et Δ^* dans les expressions valables pour les deux opérateurs
(ex. : $R(g) = \mathbb{E} [\mathbb{1}_{\{\Delta^\# g_Y(X) \leq 0\}}]$)

Définition 4 (Risque à marge) Soit $\gamma \in \mathbb{R}_+^*$. Le risque à marge γ de g se définit comme :

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

Risque empirique à marge γ :

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

Famille de fonctions à étudier : $\Delta_\gamma^\# \mathcal{G}$

Pour $\gamma \in \mathbb{R}_+^*$, soit $\pi_\gamma : \mathbb{R} \rightarrow [-\gamma, \gamma]$ la fonction linéaire saturée définie par

$$\pi_\gamma(t) = \text{signe}(t) \cdot \min(|t|, \gamma)$$

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

Mesure de capacité de $\Delta_\gamma^\# \mathcal{G}$: nombres de couverture

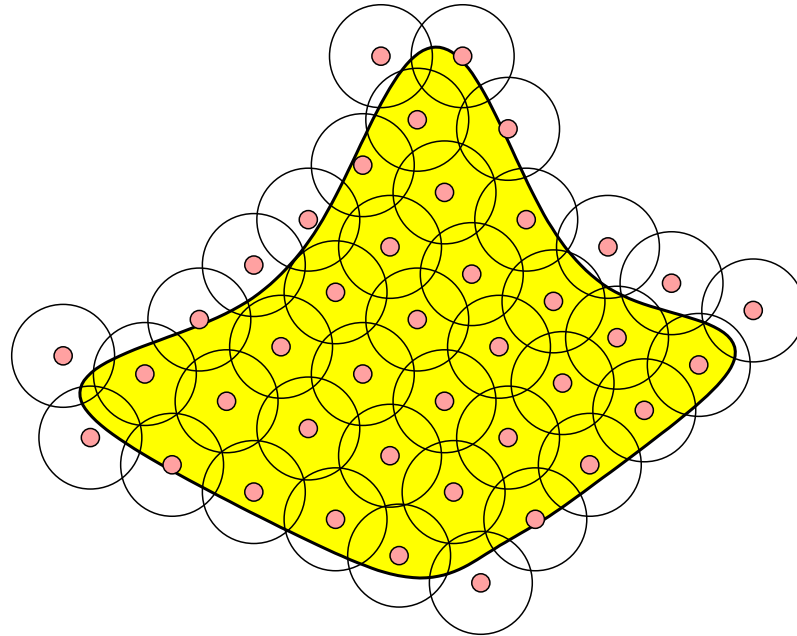


FIG. 1 – ϵ -réseau et ϵ -couverture d'un sous-ensemble E' d'un espace pseudo-métrique (E, ρ)

Définition 5 (Nombres de couverture)

$\mathcal{N}(\epsilon, E', \rho)$: nombre minimum de boules ouvertes de rayon ϵ requis pour couvrir E' (ou $+\infty$)

$\mathcal{N}^{(p)}(\epsilon, E', \rho)$: les ϵ -réseaux considérés sont ceux inclus dans E' (propres à E')

Résultat de convergence uniforme de base Classes de fonctions à valeurs binaires

Théorème 1 (Risque garanti, Vapnik, 1998) Soit \mathcal{F} une famille de fonctions sur un domaine \mathcal{X} à valeurs binaires. Soient $N\left(\mathcal{F}, (X_i)_{1 \leq i \leq n}\right)$ le nombre de fonctions (dichotomies) que cette famille peut calculer sur $(X_i)_{1 \leq i \leq n}$ et $\delta \in]0, 1[$. Avec une probabilité au moins égale à $1 - \delta$, le risque de toute fonction f de \mathcal{F} est borné supérieurement de la manière suivante :

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

$\ln \left(\mathbb{E} N \left(\mathcal{F}, (X_i)_{1 \leq i \leq 2m} \right) \right)$ est l'entropie recuite de \mathcal{F} sur l'échantillon $(X_i)_{1 \leq i \leq 2m}$.

Résultat de convergence uniforme de base

Classes de fonctions à valeurs dans \mathbb{R}^Q

Définition 6 (Pseudo-métrique d_{x^n}) *Etant donnée une suite $x^n = (x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$, on définit la pseudo-métrique d_{x^n} sur \mathcal{G} de la manière suivante :*

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

Soit $\mathcal{N}(\epsilon, \mathcal{G}, n) = \sup_{x^n \in \mathcal{X}^n} \mathcal{N}(\epsilon, \mathcal{G}, d_{x^n})$

Théorème 2 (Risque garanti) *Soit \mathcal{G} la famille de fonctions sur un domaine \mathcal{X} à valeurs dans \mathbb{R}^Q que peut réaliser un classifieur à Q catégories à grande marge. Soient $\Gamma \in \mathbb{R}_+^*$ et $\delta \in]0, 1[$. Avec une probabilité au moins égale à $1 - \delta$, uniformément pour toute valeur de γ dans $]0, \Gamma]$, le risque de toute fonction g de \mathcal{G} est borné supérieurement de la manière suivante :*

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

Ψ -dimensions

Définition 7 (Ψ -dimensions, Ben-David *et al.*, 1995) *Soit \mathcal{F} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans l'ensemble fini $\{1, \dots, Q\}$. Soit Ψ une famille d'applications ψ de $\{1, \dots, Q\}$ dans $\{-1, 1, *\}$, où le symbole $*$ représente une valeur prise par défaut. Un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de \mathcal{X} est dit être Ψ -pulvérisé par \mathcal{F} s'il existe une application $\psi^n = (\psi^{(i)})_{1 \leq i \leq n}$ dans Ψ^n telle que, pour tout vecteur v_y de $\{-1, 1\}^n$, il existe une fonction f_y dans \mathcal{F} satisfaisant*

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

La Ψ -dimension de \mathcal{F} , notée $\Psi\text{-dim}(\mathcal{F})$, est le cardinal du plus grand sous-ensemble de \mathcal{X} Ψ -pulvérisé par \mathcal{F} , si ce cardinal est fini, et l'infini dans le cas contraire.

Remarque 1 *Soient \mathcal{F} et Ψ les familles de fonctions définies ci-dessus. En étendant la définition de la dimension VC, VC-dim, de manière qu'elle s'applique aux familles de fonctions prenant leurs valeurs dans $\{-1, 1, *\}$, ce qui ne change rien en pratique, on dispose de la caractérisation suivante des Ψ -dimensions :*

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

Principaux exemples de Ψ -dimensions

Définition 8 (Dimension graphique, Dudley, 1987 ; Natarajan, 1989) Soit \mathcal{F} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans $\{1, \dots, Q\}$. La dimension graphique de \mathcal{F} , $G\text{-dim}(\mathcal{F})$, est la Ψ -dimension de \mathcal{F} dans le cas particulier où $\Psi = \{\psi_k : 1 \leq k \leq Q\}$, l'application ψ_k prenant la valeur 1 si son argument est égal à k et la valeur -1 dans le cas contraire. En reformulant cette définition dans le contexte de la discrimination multi-classe, les fonctions ψ_k sont les fonctions indicatrices des catégories.

Définition 9 (Dimension de Natarajan, Natarajan, 1989) Soit \mathcal{F} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans $\{1, \dots, Q\}$. La dimension de Natarajan de \mathcal{F} , $N\text{-dim}(\mathcal{F})$, est la Ψ -dimension de \mathcal{F} dans le cas particulier où $\Psi = \{\psi_{k,l} : 1 \leq k \neq l \leq Q\}$, l'application $\psi_{k,l}$ prenant la valeur 1 si son argument est égal à k , la valeur -1 si son argument est égal à l , et la valeur $*$ partout ailleurs.

La définition de la dimension graphique s'inspire de la méthode de décomposition un contre tous, la définition de la dimension de Natarajan s'inspirant de la méthode de décomposition un contre un.

Dimension "fat-shattering" ou dimension γ

Définition 10 (Dimension "fat-shattering", Kearns & Schapire, 1994) Soit \mathcal{G} une famille de fonctions sur \mathcal{X} à valeurs réelles. Pour $\gamma \in \mathbb{R}_+^*$, un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de \mathcal{X} est dit être γ -pulvérisé par \mathcal{G} s'il existe un vecteur $v_b = (b_i) \in \mathbb{R}^n$ tel que, pour tout vecteur $v_y = (y_i)$ de $\{-1, 1\}^n$, il existe une fonction g_y dans \mathcal{G} satisfaisant

$$\forall i \in \{1, \dots, n\}, y_i (g_y(x_i) - b_i) \geq \gamma$$

La dimension fat-shattering à marge γ , ou P_γ dimension, de la famille \mathcal{G} , $P_\gamma\text{-dim}(\mathcal{G})$, est le cardinal du plus grand sous-ensemble de \mathcal{X} γ -pulvérisé par \mathcal{G} , si ce cardinal est fini, et l'infini dans le cas contraire.

γ - Ψ -dimensions

Définition 11 (γ - Ψ -dimensions) Soit \mathcal{G} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans \mathbb{R}^Q . Soit Ψ une famille d'applications ψ de $\{1, \dots, Q\}$ dans $\{-1, 1, *\}$, où le symbole $*$ représente une valeur prise par défaut. Pour $\gamma \in \mathbb{R}_+^*$, un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de \mathcal{X} est dit être γ - Ψ -pulvérisé (Ψ -pulvérisé avec une marge γ) par $\Delta^\# \mathcal{G}$ s'il existe une application $\psi^n = (\psi^{(i)})_{1 \leq i \leq n}$ dans Ψ^n et un vecteur $v_b = (b_i)$ de \mathbb{R}^n tels que, pour tout vecteur $v_y = (y_i)$ de $\{-1, 1\}^n$, il existe une fonction g_y dans \mathcal{G} satisfaisant

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

La γ - Ψ -dimension, ou Ψ -dimension à marge γ , de $\Delta^\# \mathcal{G}$, notée $\Psi\text{-dim}(\Delta^\# \mathcal{G}, \gamma)$, est le cardinal du plus grand sous-ensemble de \mathcal{X} γ - Ψ -pulvérisé par $\Delta^\# \mathcal{G}$, si ce cardinal est fini, et l'infini dans le cas contraire.

Cette définition se réduit à celle de la dimension fat-shattering lorsque $Q = 2$.

Dimension de Natarajan à marge γ

Définition 12 (Dimension de Natarajan à marge γ) Soit \mathcal{G} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans \mathbb{R}^Q . Pour $\gamma \in \mathbb{R}_+^*$, un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de \mathcal{X} est dit être γ -N-pulvérisé (N -pulvérisé avec une marge γ) par $\Delta^\# \mathcal{G}$ s'il existe un ensemble

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

de n couples d'indices distincts dans $\{1, \dots, Q\}$ et un vecteur $v_b = (b_i)$ de \mathbb{R}^n tels que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^n$, il existe une fonction g_y de \mathcal{G} satisfaisant

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

La dimension de Natarajan à marge γ de la famille $\Delta^\# \mathcal{G}$, $N\text{-dim}(\Delta^\# \mathcal{G}, \gamma)$, est le cardinal du plus grand sous-ensemble de \mathcal{X} γ -N-pulvérisé par $\Delta^\# \mathcal{G}$, si ce cardinal est fini, et l'infini dans le cas contraire.

Lemme de Sauer-Shelah généralisé

Familles de fonctions de \mathcal{X} dans \mathbb{R}^Q

Lemme 1 *Soit \mathcal{G} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$. Pour toute valeur d' ϵ dans $]0, M_{\mathcal{G}}]$ et toute valeur entière de n satisfaisant $n \geq N\text{-dim}(\Delta\mathcal{G}, \epsilon/6)$, on dispose de la borne suivante :*

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

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

La preuve n'est plus valable si l'opérateur Δ^* est remplacé par l'opérateur Δ .

Nature et vitesse de la convergence

Théorème 3 Soit \mathcal{G} la famille de fonctions sur un domaine \mathcal{X} à valeurs dans $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$ que peut réaliser un classifieur à Q catégories à grande marge. Soit $\delta \in]0, 1[$. Avec une probabilité au moins égale à $1 - \delta$, uniformément pour toute valeur de γ dans $]0, M_{\mathcal{G}}]$, le risque de toute fonction g de \mathcal{G} est borné supérieurement de la manière suivante :

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

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

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

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

Proposition 1 (Convergences presque sûrement uniformes)

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

M-SVM - famille de fonctions réalisables

Famille de fonctions de base

Soient κ un noyau symétrique semi-défini positif sur \mathcal{X} et $(H_\kappa, \langle \cdot, \cdot \rangle_{H_\kappa})$ le RKHS correspondant

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

\mathcal{H} : famille des fonctions $h = (h_k)_{1 \leq k \leq Q}$ de \mathcal{X} dans \mathbb{R}^Q telles que :

$$\forall k \in \{1, \dots, Q\}, h_k(\cdot) = \sum_{i=1}^{m_k} \beta_{ik} \kappa(x_{ik}, \cdot) + b_k$$

avec $\{x_{ik} : 1 \leq i \leq m_k\} \subset \mathcal{X}$, $(\beta_{ik})_{1 \leq i \leq m_k} \in \mathbb{R}^{m_k}$ et $b_k \in \mathbb{R}$, ainsi que les limites de ces fonctions lorsque les ensembles $\{x_{ik} : 1 \leq i \leq m_k\}$ deviennent denses dans \mathcal{X} au sens de la norme induite par le noyau

Famille de fonctions réalisables

Sous-ensemble convexe de \mathcal{H} (défini par des contraintes sur un sous-espace affine)

Famille de fonctions de base

"Astuce du noyau"

Pour tout noyau de Mercer κ , il existe une fonction Φ telle que :

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

où $\langle \cdot, \cdot \rangle$ est le produit scalaire de l'espace ℓ_2 . Soit $\Phi(\mathcal{X}) = \{\Phi(x) : x \in \mathcal{X}\}$

Espace de représentation : l'un des espaces de Hilbert $(E_{\Phi(\mathcal{X})}, \langle \cdot, \cdot \rangle)$ engendrés par les $\Phi(\mathcal{X})$

$\implies \mathcal{H}$ peut être considérée comme une famille de fonctions affines multivariées sur $\Phi(\mathcal{X})$

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

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

Normes sur $\bar{\mathcal{H}}$ et $E_{\Phi(\mathcal{X})}^Q$

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

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

$Q \geq 3$: machines à vecteurs support multi-classes (M-SVM)

$((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times \{1, \dots, Q\})^m$: ensemble d'apprentissage

ℓ_{M-SVM} : fonction de perte convexe (construite autour de la fonction de perte charnière)

M-SVM : solution d'un problème de programmation convexe (quadratique)

Problème 1

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

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

Théorème de représentation

Ce théorème établit que l'apprentissage (la résolution du problème 1) revient à trouver les valeurs des coefficients β_{ik} dans :

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

(les valeurs des "biais" b_k s'en déduisent par application des conditions de Kuhn-Tucker)

M-SVM de Weston et Watkins

Algorithme d'apprentissage - formulation primale

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

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

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

M-SVM à marge dure et marges géométriques

Marges géométriques

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

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

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

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

Lien entre le pénalisateur et les marges géométriques

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

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

M-SVM de Crammer et Singer

Algorithme d'apprentissage - formulation primale

Problème 3 (M-SVM2, Crammer & Singer, 2001)

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

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

Algorithme d'apprentissage - formulation duale

Soient $\alpha_i. = (\alpha_{ik})_{1 \leq k \leq Q}$, $\delta_{y_i,.} = (\delta_{y_i, k})_{1 \leq k \leq Q}$, $\tau_i. = (\tau_{ik})_{1 \leq k \leq Q} = C\delta_{y_i,.} - \alpha_i.$ et $\tau = (\tau_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$

Problème 4 (M-SVM2)

$$\min_{\tau} \left\{ \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \tau_i^T \tau_j \kappa(x_i, x_j) - \sum_{i=1}^m \tau_i^T \delta_{y_i,.} \right\}$$

$$s.c. \begin{cases} \tau_{ik} \leq C\delta_{y_i, k}, & (1 \leq i \leq m), (1 \leq k \leq Q) \\ 1_Q^T \tau_i. = 0, & (1 \leq i \leq m) \end{cases}$$

M-SVM de Lee, Lin et Wahba

Algorithme d'apprentissage - formulation primale

Problème 5 (M-SVM3, Lee *et al.*, 2004)

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

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

Résultat de consistance (Zhang, 2004; Tewari & Bartlett, 2007)

Cette M-SVM est la seule dont l'apprentissage soit Bayes-consistant.

Nouvelles M-SVM

Problème 6 (M-SVM de Weston et Watkins utilisant la norme $\|\cdot\|_\infty$)

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

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

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

Problème 7 (M-SVM de Lee, Lin et Wahba "à coût quadratique")

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

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

Comportement d'une M-SVM (M-SVM1)

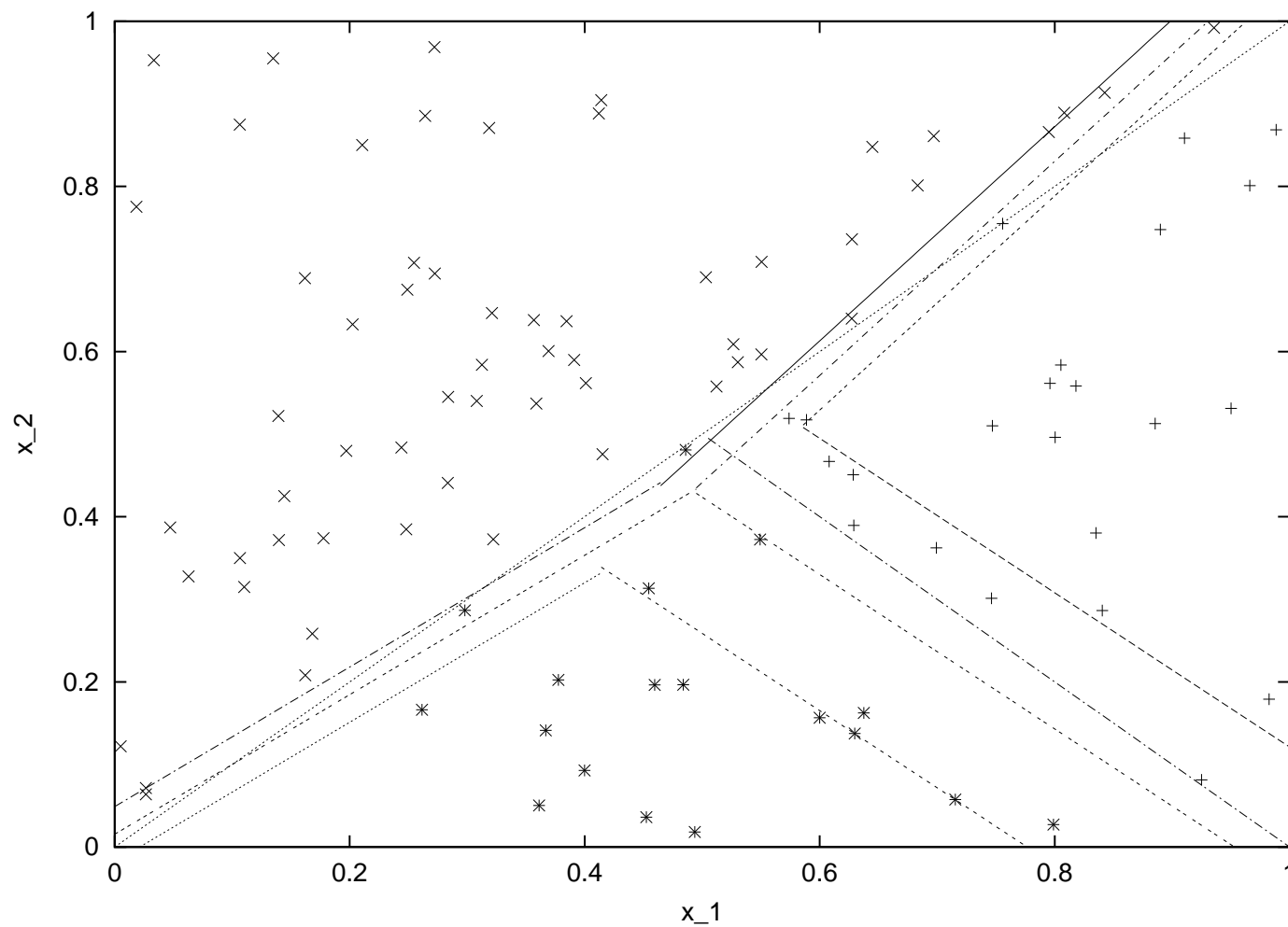


FIG. 2 – 3 catégories linéairement séparables dans \mathbb{R}^2

Comportement d'une M-SVM (M-SVM1)

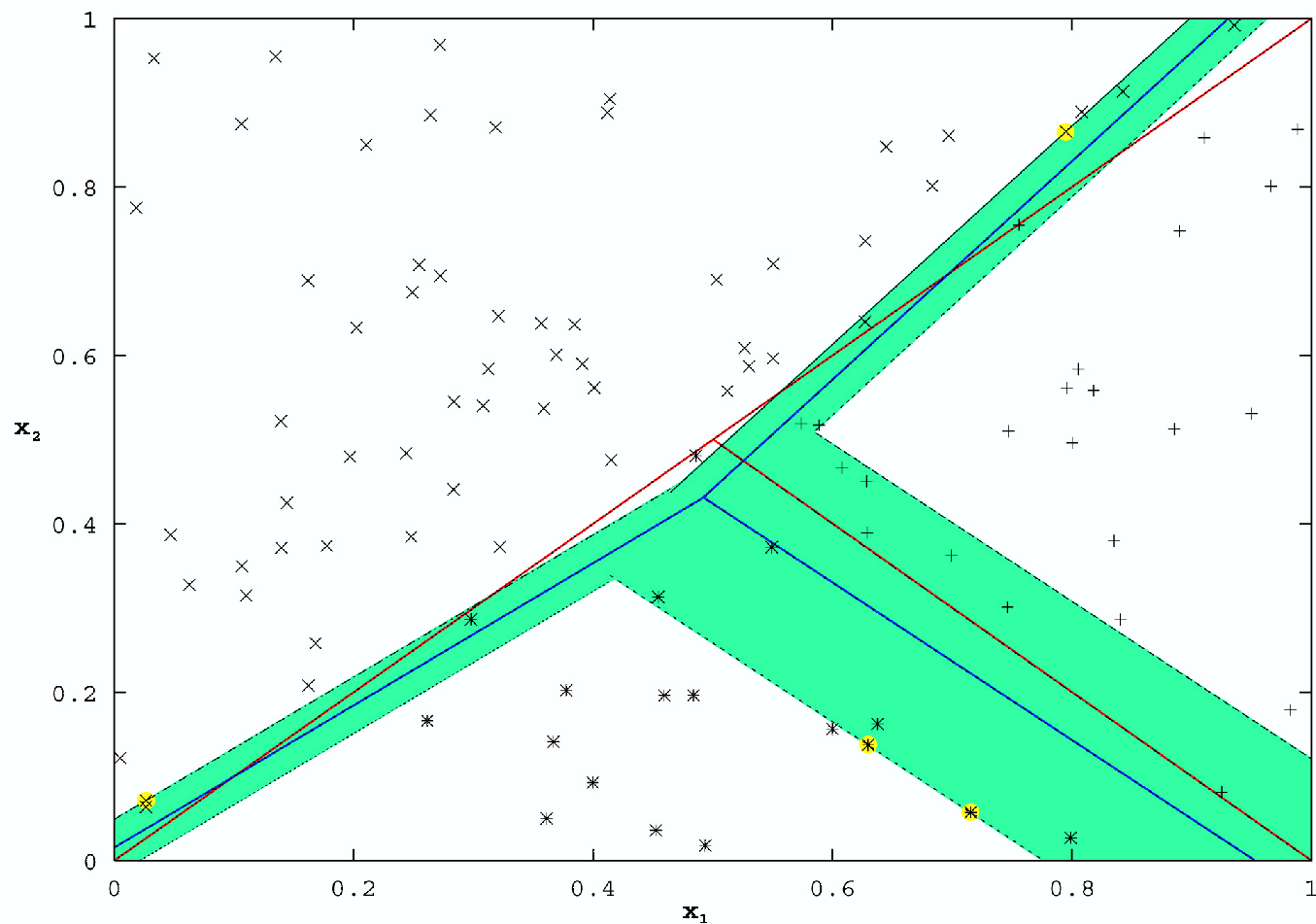


FIG. 3 – Hyperplans séparateurs et marges douces de la M-SVM linéaire

Dimension de Natarajan à marge des M-SVM

Théorème 4 Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$, que le vecteur \mathbf{w} vérifie $\|\mathbf{w}\|_{\infty} \leq \Lambda_w$ et que $\mathbf{b} = 0$. Alors, pour tout $\epsilon \in \mathbb{R}_+^*$,

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

La preuve

- n'est plus valable si l'opérateur Δ est remplacé par l'opérateur Δ^*
- s'appuie directement sur le principe de décomposition un contre un et le principe des tiroirs
- appelle l'utilisation de la norme $\|\cdot\|_{\infty}$ et non celle de la norme $\|\cdot\|$ (utilisée par le pénalisateur)

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

Utilisation d'une moyenne de Rademacher

Définition 13 (Moyenne de Rademacher) *Pour $n \in \mathbb{N}^*$, soient \mathcal{A} un ensemble borné de vecteurs $a = (a_i)_{1 \leq i \leq n}$ appartenant à \mathbb{R}^n et $(\sigma_i)_{1 \leq i \leq n}$ une suite de Rademacher. La moyenne de Rademacher associée à \mathcal{A} , $\mathcal{R}_n(\mathcal{A})$, est définie par :*

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

Théorème 5 (Inégalité des différences bornées, McDiarmid, 1989) *Pour $n \in \mathbb{N}^*$, soit $(T_i)_{1 \leq i \leq n}$ une suite de n variables aléatoires indépendantes à valeurs dans un ensemble \mathcal{T} . Soit g une fonction de \mathcal{T}^n dans \mathbb{R} telle qu'il existe une suite de constantes positives $(c_i)_{1 \leq i \leq n}$ vérifiant :*

$$\forall i \in \{1, \dots, n\}, \quad \sup_{(t_i)_{1 \leq i \leq n} \in \mathcal{T}^n, t'_i \in \mathcal{T}} |g(t_1, \dots, t_n) - g(t_1, \dots, t_{i-1}, t'_i, t_{i+1}, \dots, t_n)| \leq c_i.$$

Alors, pour tout $\tau \in \mathbb{R}_+^$, la variable aléatoire $g(T_1, \dots, T_n)$ satisfait :*

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

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

où $c = \sum_{i=1}^n c_i^2$.

Utilisation d'une moyenne de Rademacher

Risque à marge convexifié associé à la M-SVM de Crammer et Singer

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

Théorème 6 Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$, que le vecteur \mathbf{w} vérifie $\|\mathbf{w}\|_{\infty} \leq \Lambda_w$ et que $\mathbf{b} = 0$. Soit $K_{\bar{\mathcal{H}}} = \Lambda_w \Lambda_{\Phi(\mathcal{X})} + 1$. Avec une probabilité au moins égale à $1 - \delta$, le risque de toute fonction \bar{h} de $\bar{\mathcal{H}}$ est borné supérieurement de la manière suivante :

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

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

Borne "rayon-marge"

Théorème 7 (Vapnik, 1998) *Considérons une SVM bi-classe à marge dure. Soit \mathcal{L}_m son nombre d'erreurs en validation croisée "leave-one-out". En notant $\gamma = \frac{1}{\|w\|}$ sa marge géométrique sur l'ensemble d'apprentissage, on obtient pour majorant de \mathcal{L}_m :*

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

où \mathcal{D}_m est le diamètre de la plus petite boule de l'espace de représentation contenant les vecteurs support.

Borne "rayon-marge" pour la M-SVM de Weston et Watkins

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

Théorème 8 *Considérons une M-SVM de Weston et Watkins (ou Crammer et Singer) à Q catégories à marge dure. Soit \mathcal{L}_m le nombre de ses erreurs en validation croisée "leave-one-out". On a :*

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

où \mathcal{D}_m est le diamètre de la plus petite boule de l'espace de représentation contenant les vecteurs support.

Constante K_{VC}

- La valeur de K_{VC} s'obtient par résolution de programmes quadratiques en nombre égal au nombre de vecteurs support
- Pour $Q = 2$, $K_{\text{VC}} = 2$, et la borne se réduit à la borne "rayon-marge" bi-classe

Borne "rayon-marge" pour la M-SVM de Lee, Lin et Wahba

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

Théorème 9 *Considérons une M-SVM de Lee et co-auteurs à Q catégories à marge dure. Soit \mathcal{L}_m le nombre de ses erreurs en validation croisée "leave-one-out". On a :*

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

où \mathcal{D}_m est le diamètre de la plus petite boule de l'espace de représentation contenant les vecteurs support.

Cette borne se réduit encore à la borne bi-classe pour $Q = 2$.

Prédiction de la structure secondaire des protéines

Contexte biologique Exploitation fonctionnelle des informations provenant des grands programmes de séquençage des génomes : passe par la **connaissance de la structure 3D des protéines**.

1. Arrivée massive de séquences protéiques (croissance exponentielle des bases)

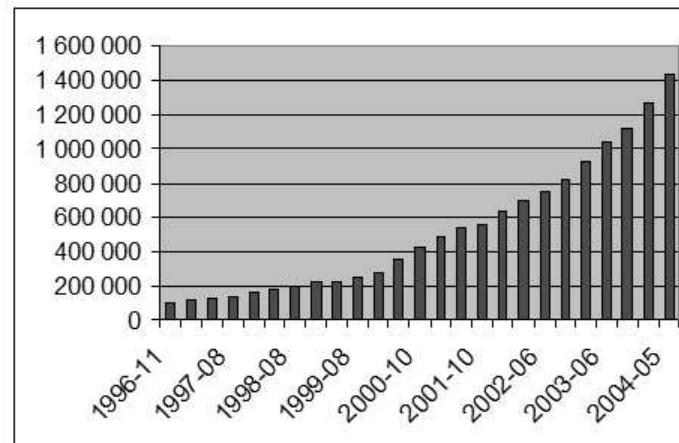


FIG. 4 – Croissance de la banque internationale TREMBL de 1996 à 2005

2. Détermination expérimentale de la structure 3D : tâche très lourde... lorsqu'elle est réalisable
⇒ **nécessité de passer d'une approche biochimique à une approche prédictive**

**Problème central en biologie permettant d'aborder
l'essentiel des grandes questions ouvertes en traitement de données séquentielles**

Différents niveaux d'organisation structurale des protéines

- Séquence ou structure primaire ($1.6 \cdot 10^6$ séquences connues)

MEEKLKKAKIIFVVGPGSGKGTQCEKIVQKYGYTHLSTC...

- Structure secondaire

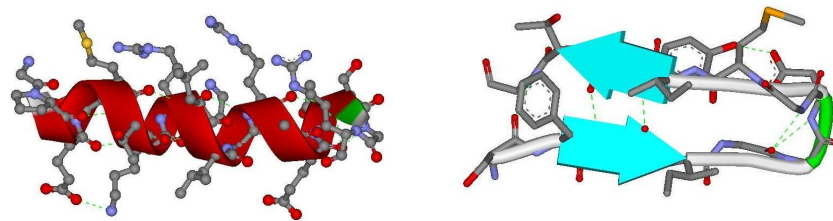
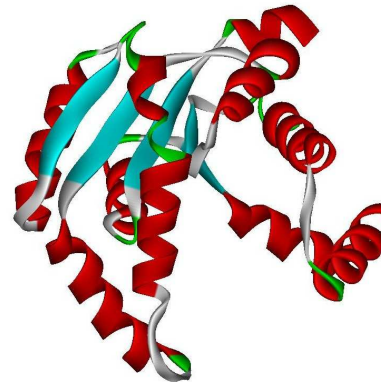


FIG. 5 – Elements structuraux périodiques : hélice α (à gauche) et brins β (à droite)

- Structure tertiaire ($2.7 \cdot 10^4$ structures 3D connues)



Noyau RBF dédié à la prédiction de la structure secondaire

Exploitation de la séquence seule

$\mathbf{x} = (x_i)_{-n \leq i \leq n}$: vecteur codant un polypeptide (contenu d'une fenêtre d'analyse de taille $2n + 1$)

$$\kappa_{\theta, G}(\mathbf{x}, \mathbf{x}') = \exp \left(- \sum_{i=-n}^n \theta_i^2 \|x_i - x'_i\|^2 \right)$$

Exploitation des alignements multiples

$\tilde{\mathbf{x}} = (\tilde{x}_i)_{-n \leq i \leq n}$ tel que $\tilde{x}_i = \sum_{j=1}^{22} \theta_{ij} a_j$ (combinaison convexe)

$$\langle \tilde{x}_i, \tilde{x}'_i \rangle = \left\langle \sum_{j=1}^{22} \theta_{ij} a_j, \sum_{k=1}^{22} \theta'_{ik} a_k \right\rangle = \sum_{j=1}^{22} \sum_{k=1}^{22} \theta_{ij} \theta'_{ik} \langle a_j, a_k \rangle$$

Prise en compte de la nature des substitutions (matrice G)

G	2																				
P	1	3																			
D	0	0	2																		
E	0	-1	1	2																	
A	0	-1	0	1	2																
N	0	0	1	0	0	3															
Q	0	0	0	1	0	1	2														
S	0	0	0	0	1	0	0	2													
T	0	0	0	0	0	0	0	0	2												
K	0	0	0	0	0	1	0	0	0	2											
R	0	0	0	0	0	0	0	0	0	1	2										
H	0	0	0	0	0	0	0	0	0	0	0	2									
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2								
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2							
M	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2						
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2					
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2				
F	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	1	0	-1	0	2			
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	0	1	2		
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	0	-1	0	0	0	2	
	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W	

FIG. 6 – S : matrice de substitution dédiée à la prédiction de la structure secondaire (Levin *et al.*, 1986)

Approximation de S par une matrice de Gram

- $A = (a_{ij}) \in \mathcal{M}_{22,22}(\mathbb{R})$: représentation vectorielle des acides aminés (en ligne)
- $G = AA^T$: matrice des produits scalaires = approximation de S par une matrice symétrique semi-définie positive

Diagonalisation de S :

$$S = PDP^{-1} = PDP^T$$

(P est orthogonale puisque S est symétrique)

$$AA^T = PD_+P^T$$

où D_+ est déduite de D en remplaçant par 0 les valeurs propres négatives

On en déduit

$$A = P\sqrt{D_+}$$

Calcul du vecteur de pondération θ

Alignement de noyaux (Cristianini *et al.*, 2002)

$$A(\kappa, \kappa') = \frac{\langle \kappa, \kappa' \rangle}{\|\kappa\| \|\kappa'\|} = \frac{\int_{\mathcal{X}^2} \kappa(x, x') \kappa'(x, x') dP_{\mathcal{X}}(x) dP_{\mathcal{X}}(x')}{\sqrt{\int_{\mathcal{X}^2} \kappa(x, x')^2 dP_{\mathcal{X}}(x) dP_{\mathcal{X}}(x')} \sqrt{\int_{\mathcal{X}^2} \kappa'(x, x')^2 dP_{\mathcal{X}}(x) dP_{\mathcal{X}}(x')}}}$$

Alignement noyau-cible empirique

$$\hat{A}_{D_m}(K_{\theta, G}, K_t) = \frac{\langle K_{\theta, G}, K_t \rangle_F}{\|K_{\theta, G}\|_F \|K_t\|_F}$$

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} \hat{A}_{D_m}(K_{\theta, G}, K_t)$$

Noyau cible pour la discrimination multi-classe (Vert, 2002)

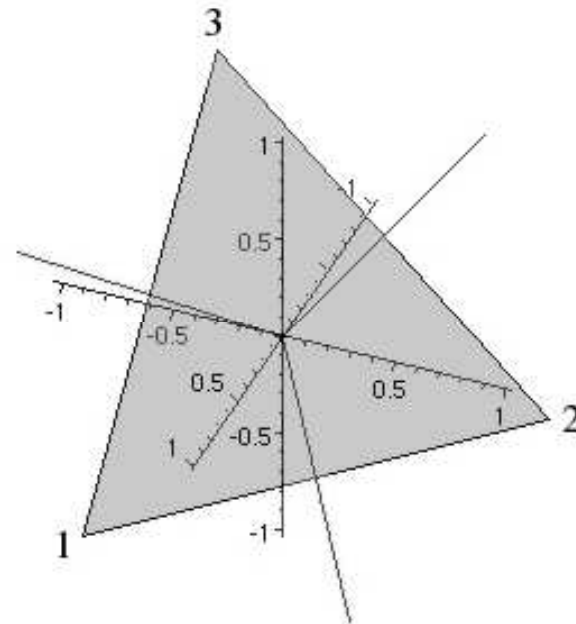
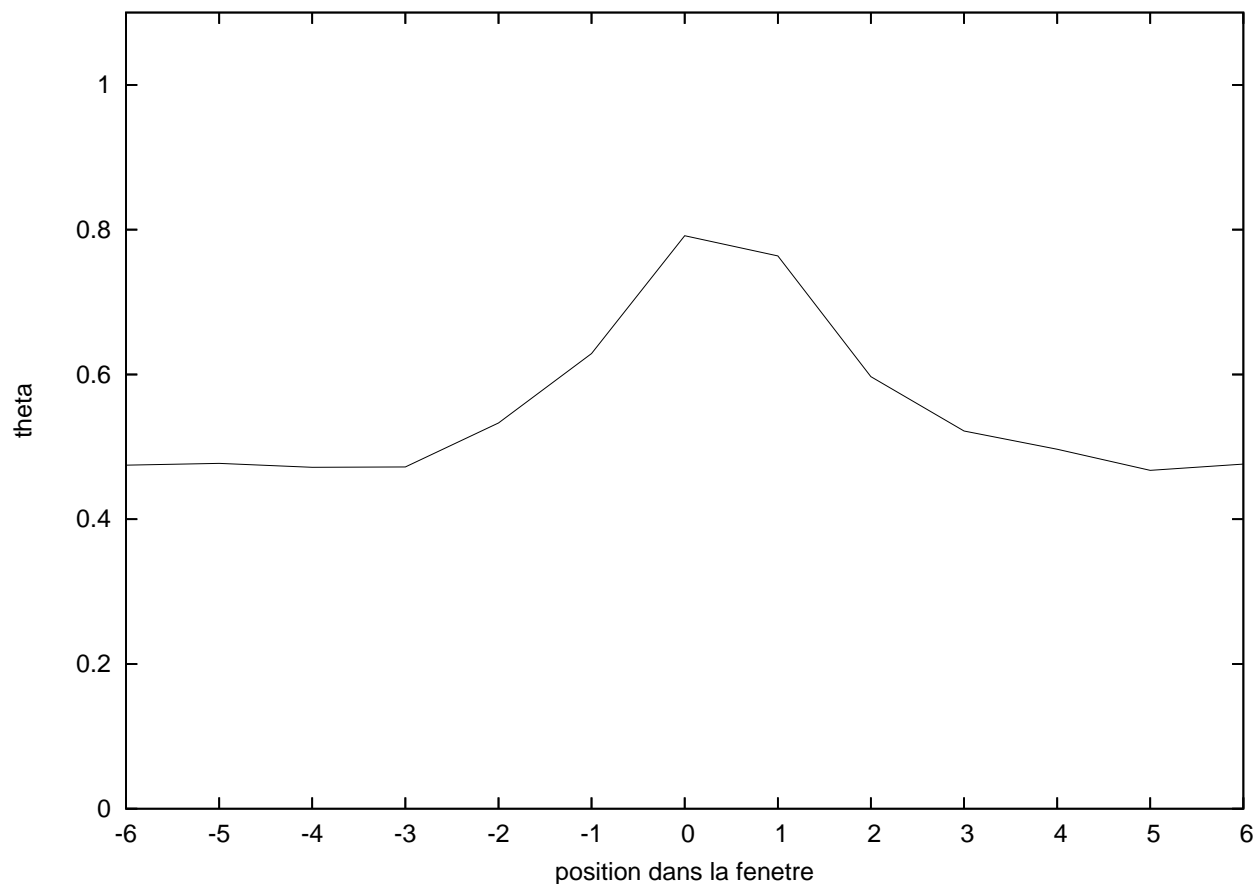


FIG. 7 – $Q = 3$: représentants optimaux des catégories

$$\begin{cases} \text{si } y = y', & \kappa_t(x, x') = 1 \\ \text{si } y \neq y', & \kappa_t(x, x') = -\frac{1}{Q-1} \end{cases}$$

Estimation du vecteur θ^* FIG. 8 – Vecteur θ obtenu par alignement noyau-cible

Algorithme d'apprentissage : descente en gradient stochastique

Résultats expérimentaux

Validation croisée à 5 pas sur une base de **1096 protéines** constituées de **268575 résidus**
(taux d'identité < 30%)

	séquence		alignement - profil		alignement - sortie	
	PMC	M-SVM	PMC	M-SVM	PMC	M-SVM
Q_3	61.6	62.7	72.0	72.3	68.9	69.6
C_α	0.46	0.47	0.63	0.64	0.55	0.59
C_β	0.33	0.38	0.53	0.54	0.42	0.48
C_c	0.38	0.41	0.53	0.54	0.47	0.46
Sov	53.9	54.5	65.1	65.3	64.0	64.8
Sov_α	57.8	57.9	66.5	66.7	64.4	65.0
Sov_β	44.7	46.8	61.5	62.3	58.4	61.6
Sov_c	57.3	57.7	66.7	66.8	64.2	66.1

Conclusions et perspectives

Risques garantis et sélection de modèle pour les M-SVM

- Les γ - Ψ -dimensions jouent pour les M-SVM (et les PMC!) le même rôle que la dimension fat-shattering pour les SVM bi-classes.
- D'importants progrès restent à effectuer dans la formulation de la borne VC et la majoration des mesures de capacité.
- La sélection de modèle fournit un pierre de touche pour l'évaluation des risques garantis et des bornes sur le risque empirique que nous établissons.

Prédiction de la structure secondaire des protéines

- Les M-SVM doivent pouvoir être utilisées au côté des PMC comme éléments de base des méthode de prédiction de la stucture secondaire des protéines.
- Le développement de modèles hybrides, intégrant systèmes discriminants et génératifs, constitue l'une des principales options pour obtenir des progrès en biologie structurale prédictive.

Pour en savoir plus...

Lemme de Sauer-Shelah

Lemme 2 (Vapnik & Chervonenkis, 1971 ; Sauer, 1972 ; Shelah, 1972) Soient \mathcal{F} une famille de fonctions sur un domaine \mathcal{X} à valeurs binaires, $\Pi_{\mathcal{F}}$ sa fonction de croissance, qui à $n \in \mathbb{N}^*$ associe $\Pi_{\mathcal{F}}(n) = \sup_{S \subset \mathcal{X}} N(\mathcal{F}, S)$ et d sa dimension VC ($\Pi_{\mathcal{F}}(d) = 2^d$ et $\Pi_{\mathcal{F}}(d+1) < 2^{d+1}$). Alors pour $n \geq d$,

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

où e est la base des logarithmes népériens.

Lemme de Sauer-Shelah généralisé

Familles de fonctions de \mathcal{X} dans $\{1, \dots, Q\}$

Lemme 3 (Haussler & Long, 1995) Soient \mathcal{F} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans $\{1, \dots, Q\}$, $\Pi_{\mathcal{F}}$ sa fonction de croissance et d sa dimension de Natarajan. Alors pour $n \geq d$,

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

Lemme de Sauer-Shelah généralisé

Familles de fonctions de \mathcal{X} dans \mathbb{R}

Lemme 4 (Alon et al., 1997) *Soit \mathcal{G} une famille de fonctions sur un domaine \mathcal{X} à valeurs dans $[0, 1]$. Pour toute valeur d' ϵ dans $]0, 1]$ et toute valeur entière de n satisfaisant $n \geq P_{\epsilon/4}\text{-dim}(\mathcal{G})$, on dispose de la borne suivante :*

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

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

Dimension graphique et consistance du principe inductif ERM

Lemme de Sauer-Shelah

S'applique directement sous sa forme initiale, car la dimension graphique de \mathcal{F} est simplement la dimension VC de l'espace des graphes de \mathcal{F} , \mathcal{GF}

Résultat de convergence uniforme

Théorème 10 (Anthony, 1997) Soit $\Pi_{\mathcal{GF}}$ la fonction de croissance de l'espace des graphes de \mathcal{F} . Alors :

$$\mathbb{P}_{D_m} \left(\sup_{f \in \mathcal{F}} \left(\frac{R(f) - R_m(f)}{\sqrt{R(f)}} \right) > \epsilon \right) \leq 4\Pi_{\mathcal{GF}}(2m) \exp \left(-\frac{m\epsilon^2}{4} \right)$$

Espaces de Hilbert à noyau reproduisant

Soient \mathcal{X} un espace quelconque et $(H, \langle \cdot, \cdot \rangle_H)$ un espace de Hilbert de fonctions sur \mathcal{X} ($H \subset \mathbb{R}^{\mathcal{X}}$).

Noyau reproduisant (Aronszajn, 1950)

Soit κ une fonction de $\mathcal{X} \times \mathcal{X}$ dans \mathbb{R}

$\forall x \in \mathcal{X}$, soit κ_x la fonction de \mathcal{X} dans \mathbb{R} définie par $\kappa_x : t \mapsto \kappa(x, t)$

κ est appelé un *noyau reproduisant* pour H si et seulement si :

1. $\forall x \in \mathcal{X}, \kappa_x \in H$
2. $\forall x \in \mathcal{X}, \forall h \in H, \langle h, \kappa_x \rangle_H = h(x)$ (*propriété de reproduction*)

Espace de Hilbert à noyau reproduisant

Si un noyau reproduisant existe, H est appelé un *espace de Hilbert à noyau reproduisant* (RKHS).

Les noyaux symétriques semi-définis positifs sont des noyaux reproduisants.

$Q = 2$: machines à vecteurs support bi-classes (SVM)

$((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times \{-1, 1\})^m$: ensemble d'apprentissage

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

$\ell_{\text{SVM}}(y, \tilde{h}(x)) = \left(1 - y\tilde{h}(x)\right)_+$ (fonction de perte charnière)

SVM : solution d'un problème de programmation convexe (quadratique)

Problème 8

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

Théorème de représentation

Ce théorème établit que l'apprentissage (la résolution du problème 8) revient à trouver les valeurs des coefficients β_i dans :

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

(la valeur du "biais" b s'en déduit par application des conditions de Kuhn-Tucker)

M-SVM de Weston et Watkins

Algorithme d'apprentissage - formulation duale

α_{ik} : multiplicateur de Lagrange associé à la contrainte $\langle w_{y_i} - w_k, \Phi(x_i) \rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}$

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

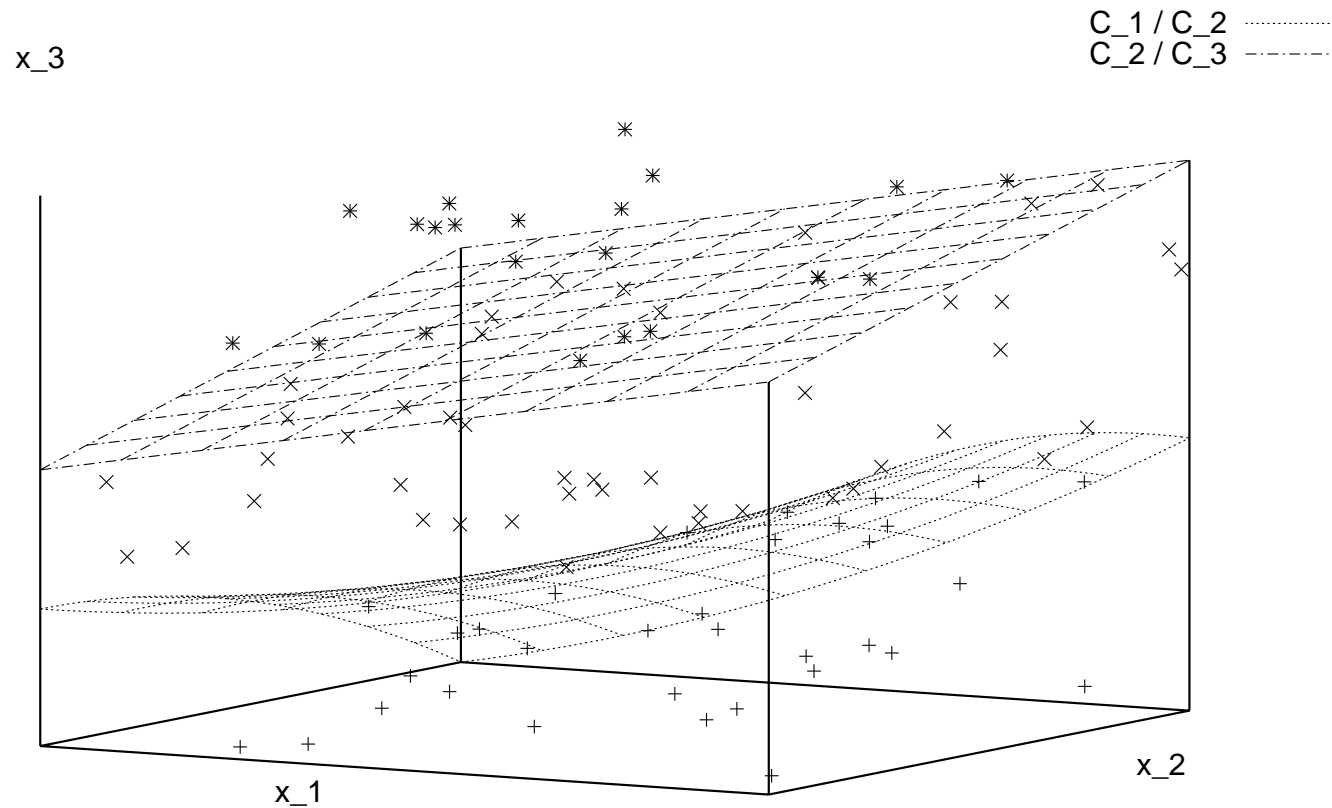
Problème 9 (M-SVM1)

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

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

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

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

FIG. 9 – 3 catégories non linéairement séparables dans \mathbb{R}^3

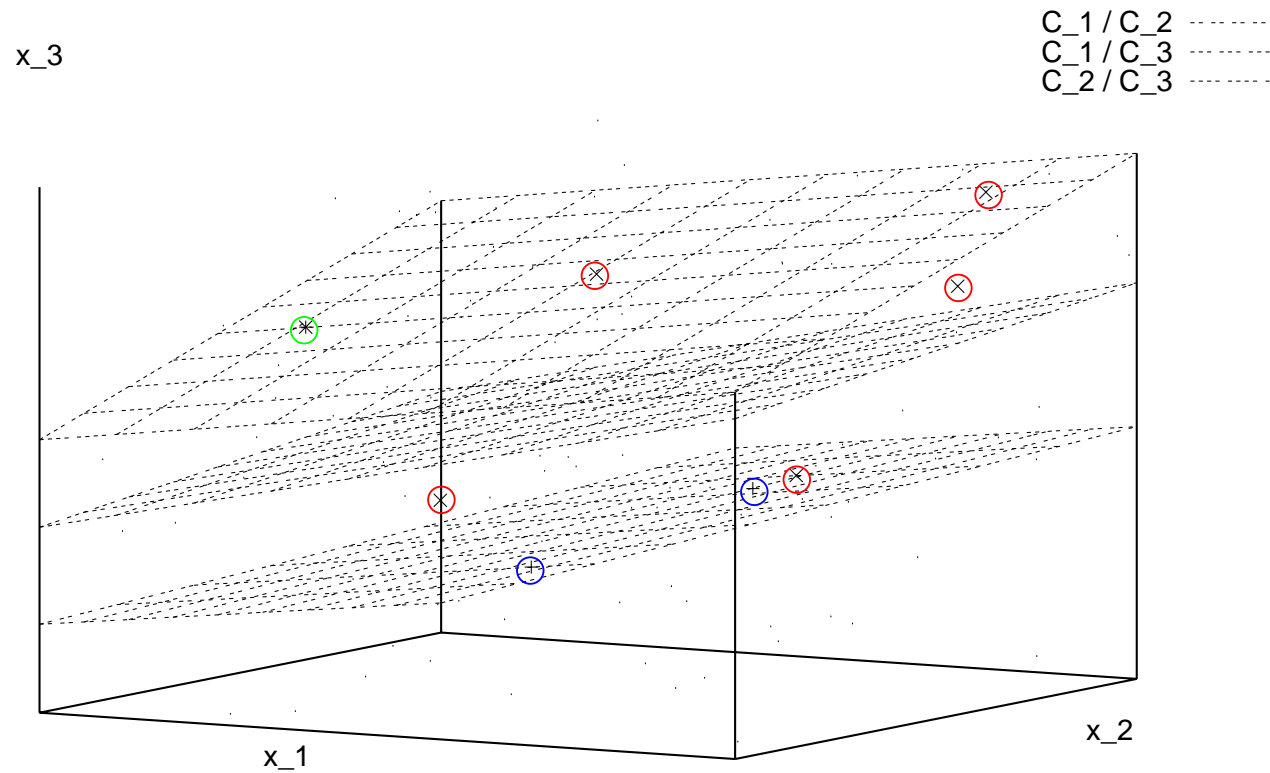


FIG. 10 – Hyperplans séparateurs et vecteurs support de la M-SVM linéaire

Théorie VC des classifieurs à grande marge - spécificités du cas multi-classe

$\Delta \neq \Delta^*$ pour $Q \geq 3$

Lemme de Sauer-Shelah généralisé

Opérateur Δ : fournit "trop" d'information pour établir le lien entre séparation et capacité à "pulvériser" un singleton

Dimension VC généralisée

Opérateur Δ^* : ne fournit pas assez d'information pour calculer une borne sur une dimension VC généralisée

\implies astuce : réaliser la transition entre Δ^* et Δ au niveau du lemme de Sauer-Shelah généralisé

Des nombres de couverture aux nombres d'entropie

Définition 14 (Nombres d'entropie d'un ensemble) Soient (E, ρ) un espace pseudo-métrique (ou $(E, \|\cdot\|_E)$ un espace de Banach) et E' un sous-ensemble borné de E . Alors, pour $n \in \mathbb{N}^*$, le n -ième nombre d'entropie de E' , $\epsilon_n(E')$, est :

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

Définition 15 (Nombres d'entropie d'un opérateur linéaire borné) Soient $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ deux espaces de Banach. Soit $\mathcal{L}(E, F)$ l'espace de Banach de tous les opérateurs (linéaires bornés) de $(E, \|\cdot\|_E)$ dans $(F, \|\cdot\|_F)$ muni de la norme :

$$\forall S \in \mathcal{L}(E, F), \|S\| = \sup_{e \in E: \|e\|_E=1} \|S(e)\|_F.$$

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

Des nombres de couverture aux nombres d'entropie

Définition 16 (Opérateur d'évaluation) Pour $n \in \mathbb{N}^*$, soit $x^n \in \mathcal{X}^n$. L'opérateur d'évaluation S_{x^n} sur $\bar{\mathcal{H}}$ est défini par :

$$S_{x^n} : \quad \bar{\mathcal{H}} \quad \longrightarrow \quad \ell_\infty^{Qn}$$

$$\bar{h} = (w_k)_{1 \leq k \leq Q} \quad \mapsto \quad S_{x^n}(\bar{h}) = (\langle w_k, \Phi(x_i) \rangle)_{1 \leq i \leq n, 1 \leq k \leq Q}$$

$$\mathcal{U} = \{\bar{h} \in \bar{\mathcal{H}} : \|\mathbf{w}\|_\infty \leq 1\}$$

Le lien entre $\mathcal{N}(\epsilon, \mathcal{U}, n)$ et les nombres d'entropie de S_{x^n} est fourni par la proposition suivante :

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

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

Majoration des nombres d'entropie Espace de représentation de dimension finie

Proposition 3 (Carl & Stephani, 1990) *Soient E et F des espaces de Banach et $S \in \mathcal{L}(E, F)$. Si S est de rang r , alors pour $n \in \mathbb{N}^*$,*

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

Théorème 11 *Soit \mathcal{H} la famille des fonctions réalisables par une M -SVM à Q catégories sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$, que le vecteur \mathbf{w} vérifie $\|\mathbf{w}\|_{\infty} \leq \Lambda_w$ et que $\mathbf{b} \in [-\beta, \beta]^Q$. Si la dimension de l'espace $E_{\Phi(\mathcal{X})}$ est finie et égale à d , alors, pour tout $\gamma \in \mathbb{R}_+^*$,*

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

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

Majoration des nombres d'entropie Espace de représentation de dimension infinie

Théorème 12 (Théorème de Maurey-Carl, Carl & Stephani, 1990) *Soient H un espace de Hilbert et S un opérateur appartenant à $\mathcal{L}(\ell_1^n, H)$ ou $\mathcal{L}(H, \ell_\infty^n)$. Alors, pour tout couple d'entiers (k, n) vérifiant $1 \leq k \leq n$, on a*

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

où le nombre d'entropie dyadique $e_k(S)$ est égal à $\epsilon_{2^{k-1}}(S)$ et c est une constante universelle.

Théorème 13 *Soit \mathcal{H} la famille des fonctions réalisables par une M -SVM à Q catégories sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$, que le vecteur \mathbf{w} vérifie $\|\mathbf{w}\|_\infty \leq \Lambda_w$ et que $\mathbf{b} \in [-\beta, \beta]^Q$. Alors, pour tout $\gamma \in \mathbb{R}_+^*$,*

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

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

Approche hybride pour la prédiction de la structure secondaire

