VC dimensions généralisées pour les SVM multi-classes

Yann Guermeur

LORIA - CNRS

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

Plan de la présentation

Théorie statistique de la discrimination multi-classe

- Hypothèses fondamentales
- Marge multi-classe et risque à marge

Risque garanti pour les modèles multi-classes à grande marge

- Résultat de convergence uniforme de base
- Ψ-dimensions et dimension "fat-shattering"
- Ψ-dimensions à marge
- Lemme de Sauer-Shelah généralisé

Application aux SVM multi-classes

- Famille de fonctions réalisables
- Critère d'apprentissage et théorème de représentation
- Borne sur la dimension de Natarajan à marge

Hypothèses fondamentales

• **Caractérisation du problème**

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

• **Ce dont on dispose**

1. $s_m = \{(X_i, Y_i) : 1 \leq i \leq m\}$: m-échantillon constitué de copies de (X, Y) indépendantes
2. $\mathcal{F}$, famille de fonctions f, de $\mathcal{X}$ dans $\mathcal{Y}$, ou $\mathcal{G}$, famille de fonctions g, de $\mathcal{X}$ dans $\mathbb{R}^Q$

• **Ce que l'on cherche à faire**

1. ℓ, fonction de perte : $\ell\left(y, f(x)\right) = \mathbb{1}_{\{f(x) \neq y\}}$ ou $\ell\left(y, g(x)\right) = \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}}$
2. Risque : espérance de la fonction de perte
3. Sélection d'une fonction f^* ou g^* minimisant sur sa famille le risque

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

Marge multi-classe et risque à marge

Définition 1 (Marge multi-classe (Elisseeff *et al.*, 1999, Allwein *et al.*, 2000))

$$M(g,x,y) = \frac{1}{2}\left\{g_y(x) - \max_{k \neq y} g_k(x)\right\}$$

Opérateurs de marge

Définition 2 (Opérateurs 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 :

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

Soit $M(g,x) = \frac{1}{2}\max_k\{g_k(x) - \max_{l \neq k} g_l(x)\}$.

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

$$\forall k \in \{1, \dots, Q\}, \quad \Delta^* g_k(x) = \begin{cases} M(g,x) & si\ \frac{1}{2}\{g_k(x) - \max_{l \neq k} g_l(x)\} = M(g,x) \\ -M(g,x) & sinon \end{cases}$$

Marge multi-classe et risque à marge

- $\Delta^{\#}$ remplace indiféremment Δ ou Δ^{*}

Définition 3 (Risque empirique à marge) *Le risque empirique à marge $\gamma > 0$ de g sur un ensemble $s_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ est :*

$$R_{\gamma,s_m}(g) = \frac{1}{m} \cdot \# \left\{(x_i, y_i) \in s_m : \ \Delta^{\#} g_{y_i}(x_i) < \gamma \right\}$$

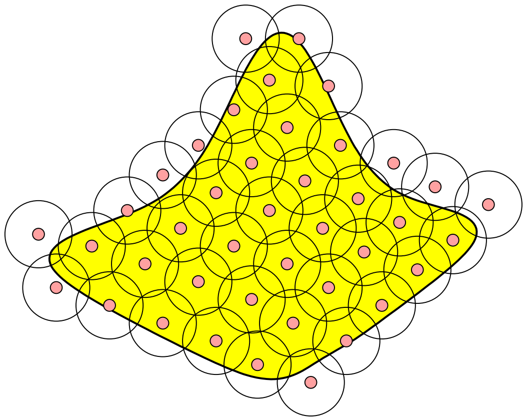
Famille de fonctions à étudier : $\Delta_\gamma^{\#} \mathcal{G}$

Pour $\gamma > 0$, soit $\pi_\gamma : \mathbb{R} \to [-\gamma, \gamma]$ la fonction linéaire saturée définie par

$$\pi_\gamma(t) = \begin{cases} \gamma.signe(t) & si \ |t| \geq \gamma \\ t & sinon \end{cases}$$

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

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



ϵ-*réseau* et ϵ-*couverture* d'un ensemble E' dans un espace pseudo-métrique (E, ρ)

Définition 4 (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' *(pas seulement dans* E*)*

Résultat de convergence uniforme de base

Définition 5 (Pseudo-métrique $d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^m})}$) *Pour $s_{\mathcal{X}^m} \in \mathcal{X}^m$, la pseudo-métrique $d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^m})}$ sur $\mathcal{G}$ est définie par :*

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

Soit $\mathcal{N}^{(p)}_{\infty,\infty}(\epsilon, \Delta^{\#}_\gamma \mathcal{G}, m) = \sup_{s_{\mathcal{X}^m} \in \mathcal{X}^m} \mathcal{N}^{(p)}(\epsilon, \Delta^{\#}_\gamma \mathcal{G}, d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^m})})$

Théorème 1 (Risque garanti) *Soit s_m un m-échantillon d'exemples i.i.d. suivant P. Avec une probabilité au moins $1 - \delta$, pour toute valeur de γ dans $]0, 1]$, le risque de toute fonction g dans $\mathcal{G}$ est borné supérieurement par :*

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

Ψ-dimensions

Définition 6 (Ψ-dimension (Ben-David *et al.*, 1995)) *Soit* $\mathcal{F}$ *une famille de fonctions sur un ensemble* $\mathcal{X}$ *prenant leurs valeurs dans un ensemble fini* $\{1, \ldots, Q\}$*. Soit* Ψ *une famille d'applications* ψ*, de* $\{1, \ldots, Q\}$ *dans* $\{-1, 1, *\}$*, où* $*$ *joue le rôle d'élément poubelle. Un sous-ensemble* $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ *de* $\mathcal{X}$ *est dit être* Ψ-pulvérisé *par* $\mathcal{F}$ *s'il existe une application* $\psi^m = \left\{\psi^{(1)}, \ldots, \psi^{(m)}\right\}$ *de* Ψ^m *telle que pour tout vecteur* v_y *de* $\{-1, 1\}^m$*, il existe une fonction* f_y *de* $\mathcal{F}$ *vérifiant* $\left(\psi^{(1)} \circ f_y(x_1), \ldots, \psi^{(i)} \circ f_y(x_i), \ldots, \psi^{(m)} \circ f_y(x_m)\right) = v_y$*.*

$\mathcal{F}$ *et* Ψ *étant définis comme précédemment, la* Ψ-dimension *de* $\mathcal{F}$*, notée* Ψ*-dim*$(\mathcal{F})$*, est le cardinal maximal d'un sous-ensemble de* $\mathcal{X}$ Ψ*-pulvérisé par* $\mathcal{F}$*, si ce cardinal est fini, l'infini dans le cas contraire.*

Principaux exemples de Ψ-dimensions

Définition 7 (Dimension graphique (Dudley, 1987, Natarajan, 1989)) *Soit $\mathcal{F}$ une famille de fonctions de $\mathcal{X}$ dans $\{1, \dots, Q\}$. La* dimension graphique *de $\mathcal{F}$, G-dim($\mathcal{F}$), est la Ψ-dimension de $\mathcal{F}$ dans le cas où Ψ est restreint à l'ensemble des Q applications ψ_k, $(1 \leq k \leq Q)$, telles que ψ_k prend la valeur 1 si son argument est égal à k, et la valeur -1 dans le cas contraire. Reformulé dans le contexte de la discrimination multi-classe, les fonctions ψ_k sont les fonctions indicatrices des catégories.*

Définition 8 (Dimension de Natarajan (Natarajan, 1989)) *Soit $\mathcal{F}$ une famille de fonctions de $\mathcal{X}$ dans $\{1, \dots, Q\}$. La* dimension de Natarajan *de $\mathcal{F}$ est la Ψ-dimension de $\mathcal{F}$ dans le cas où Ψ est restreint à l'ensemble des C_Q^2 applications $\psi_{k,l}$, $(1 \leq k < l \leq Q)$, telles que :*

$$\psi_{k,l}(t) = \begin{cases} 1 & si \;\; t = k \\ -1 & si \;\; t = l \\ * & sinon \end{cases}$$

Importance centrale de la dimension graphique

Lemme de Sauer

S'applique directement dans sa formulation de base, puisque la dimension graphique de $\mathcal{F}$ n'est autre que la dimension VC de l'*espace des graphes* de $\mathcal{F}$

Résultat de convergence uniforme

Théorème 2 (Anthony, 1997) *Soit $G_{G,\mathcal{F}}$ la fonction de croissance de $\mathcal{F}$ associée à la dimension graphique. Alors :*

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : \ (R_{s_m}(f) \leq (1-\delta)\epsilon \ \wedge \ R(f) > \epsilon)\right\} < 4G_{G,\mathcal{F}}(2m) \exp\left(-\frac{\delta^2 \epsilon m}{4}\right)$$

Dimension "fat-shattering" ou dimension γ

Définition 9 (Dimension "fat-shattering" (Kearns & Schapire, 1990)) *Soit $\mathcal{G}$ une famille de fonctions à valeurs réelles sur un ensemble $\mathcal{X}$. Pour $\gamma > 0$, un sous-ensemble $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ de $\mathcal{X}$ est dit être* γ-pulvérisé *par $\mathcal{G}$ s'il existe un vecteur $v_b = (b_i) \in \mathbb{R}^m$ tel que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^m$, il existe une fonction $g_y \in \mathcal{G}$ vérifiant*

$$y_i \left(g_{y_i}(x_i) - b_i\right) \geq \gamma, \ (1 \leq i \leq m)$$

On dit alors que le vecteur v_b témoigne *de la γ-pulvérisation de $s_{\mathcal{X}^m}$ par $\mathcal{G}$.*

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

Ψ-dimensions à marge

Définition 10 (γ-Ψ-pulvérisation) *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 $\{-1,1,*\}$. Pour $\gamma > 0$, un sous-ensemble $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ de $\mathcal{X}$ est dit être γ-Ψ-*pulvérisé *(Ψ-pulvérisé avec une marge γ) par $\Delta\mathcal{G}$ s'il existe une application $\psi^m = \left(\psi^{(1)},\dots,\psi^{(i)},\dots,\psi^{(m)}\right)$ de Ψ^m et un vecteur $v_b = (b_i)$ de $\mathbb{R}^m$ tels que, pour tout vecteur $v_y = (y_i)$ de $\{-1,1\}^m$, il existe une fonction g_y dans $\mathcal{G}$ satisfaisant*

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

Définition 11 (Ψ-dimension à marge γ) *$\mathcal{G}$, Ψ et γ sont définis comme précédemment. La Ψ-dimension de $\Delta\mathcal{G}$ à marge γ, notée Ψ-dim$(\Delta\mathcal{G},\gamma)$, est le cardinal maximal d'un sous-ensemble de $\mathcal{X}$ γ-Ψ-pulvérisé par $\Delta\mathcal{G}$, si ce cardinal est fini, ou l'infini dans le cas contraire.*

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 $\gamma > 0$, un sous-ensemble $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ de $\mathcal{X}$ est dit être* γ-N-pulvérisé *(N-pulvérisé avec une marge γ) par $\Delta\mathcal{G}$ s'il existe un ensemble*

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

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

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

La dimension de Natarajan à marge γ *de la famille $\Delta\mathcal{G}$, N-dim$(\Delta\mathcal{G}, \gamma)$, est le cardinal maximal d'un sous-ensemble de $\mathcal{X}$ γ-N-pulvérisé par $\Delta\mathcal{G}$, si ce cardinal est fini, ou l'infini dans le cas contraire.*

Lemme de Sauer-Shelah généralisé

Théorème 3 *Soit* $\mathcal{G}$ *une famille de fonctions d'un domaine* $\mathcal{X}$ *dans* $\mathbb{R}^Q$*. Pour toute valeur de* γ *dans* $]0,1]$ *et toute valeur entière de* m *satisfaisant* $2m \geq N\text{-}dim\left(\Delta_\gamma \mathcal{G}, \gamma/24\right)$*, on dispose de la borne suivante :*

$$\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/4, \Delta^*_\gamma \mathcal{G}, 2m) < 2\left(288\ m\ Q^2(Q-1)\right)^{\lceil d \log_2(23emQ(Q-1)/d)\rceil}$$

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

Espaces de Hilbert à noyau reproduisant

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

Noyau reproduisant (Aronszajn, 1950)

Une fonction $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ est appelée un *noyau reproduisant* pour H si et seulement si :

1. H contient toutes les fonctions de la forme

$$\forall x \in \mathcal{X}, \quad \kappa_x : t \mapsto \kappa(x, t)$$

2. $\forall x \in \mathcal{X}, \quad \forall h \in H, \quad h(x) = \langle h, \kappa_x \rangle_H$

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

Famille de fonctions de base

Soit κ un noyau symétrique semi-défini positif et $(H, \langle ., . \rangle_{H_\kappa})$ le RKHS correspondant

Soit $\mathcal{H} = \left((H, \langle ., . \rangle_{H_\kappa}) + \{1\} \right)^Q$

$\mathcal{H}$: famille des fonctions $h = (h_k)_{1 \leq k \leq Q}$ dont les fonctions composantes sont les fonctions de la forme

$$h_k(x) = \sum_{i=1}^{l_k} \beta_{ik} \kappa(x_{ik}, x) + b_k$$

où $s^{(k)} = \{x_{ik} : 1 \leq i \leq l_k\} \subset \mathcal{X}$ et les limites de ces fonctions lorsque $s^{(k)}$ devient dense dans $\mathcal{X}$

On note $\bar{\mathcal{H}} = (H, \langle ., . \rangle_{H_\kappa})^Q$

Machines à vecteurs support multi-classes (M-SVM)

$s_m = \{(x_i, y_i) : 1 \leq i \leq m\}$: échantillon d'apprentissage

ℓ_{SVM} : fonction de perte convexe

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

Problème 1

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

Théorème de représentation

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

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

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

Dimension de Natarajan à marge des M-SVMs

"Kernel trick"

$$\exists \Phi \ : \mathcal{X} \longrightarrow E_{\Phi(\mathcal{X})} \ / \ \forall (x, x') \in \mathcal{X}^2, \ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\ell_2}$$

$$H_\kappa = E^*_{\Phi(\mathcal{X})}$$

$$\forall k \in \{1, \dots, Q\}, \ h_k(x) = \langle w_k, \Phi(x) \rangle_{\ell_2} + b_k$$

Théorème 4 *Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une SVM à Q-catégories sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule de rayon $\Lambda_{\Phi(\mathcal{X})}$ de $E_{\Phi(\mathcal{X})}$, que les vecteurs w_k définissant les hyperplans séparateurs vérifient $1/2 \max_{1 \leq k < l \leq Q} \|w_k - w_l\|_{H_\kappa} \leq \Lambda_w$ et que $b = 0_Q$. Alors, pour toute valeur strictement positive d'ϵ, on dispose de la borne suivante :*

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

Spécificités du cas multi-classe

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

Lemme de Sauer-Shelah étendu

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

Dimension VC étendue

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

$\Longrightarrow$ l'astuce consiste à réaliser la transition entre Δ^* et Δ au niveau du lemme de Sauer-Shelah étendu

Conclusions et perspectives

Conclusions

- Les Ψ-dimensions à marge jouent pour les M-SVM (et les PMC !) le même rôle que la dimension fat-shattering pour les SVM bi-classes
- Ces travaux mettent en évidence la spécificité du cas multi-classe
- Des bornes sur le risque plus fines doivent résulter de l'extension au cas multi-classe d'inégalités de concentration plus récentes

Perspectives (et travail en cours)

- Achèvement et évaluation du travail sur les nombres d'entropie de l'opérateur d'évaluation (théorème de Maurey-Carl)
- Dérivation de bornes fondées sur des mesures empiriques et locales de la capacité (inégalités de concentration)
- Développement de méthodes de sélection de modèle dédiées aux M-SVM