

Introduction à l'apprentissage statistique

Fabrice Rossi

Projet AxIS, INRIA Paris Rocquencourt

Janvier 2008

Plan de la 1ère partie

Introduction

Apprentissage automatique

- Principes

- Exemples d'algorithmes d'apprentissage

Théorie de l'apprentissage

- Formalisation

- Construction d'un modèle

- Pas de repas gratuit

Plan de la 2ème partie

Le risque empirique

Concentration

- Inégalité de Hoeffding

- Bornes uniformes

Dimension de Vapnik-Chervonenkis

- Définition

- Application à la discrimination

- Preuve

Nombres de couverture

- Définition et résultat

- Calcul des nombres de couverture

Résumé

Contrôle de complexité

Contrôle implicite

Discrimination

Régression

Contrôle explicite

Minimisation du risque structurel

Validation

Régularisation

Au delà de la minimisation du risque structurel

Minimiser un autre coût

- Motivations

- Risque convexe

Retour sur la régularisation

- Régularisation dans les RKHS

Première partie I

Introduction

Plan

Apprentissage automatique

- Principes

- Exemples d'algorithmes d'apprentissage

Théorie de l'apprentissage

- Formalisation

- Construction d'un modèle

- Pas de repas gratuit

Apprentissage statistique

Apprentissage statistique

Apprentissage statistique

Apprentissage statistique =

Apprentissage automatique + statistique

Apprentissage statistique

Apprentissage statistique =

Apprentissage automatique + statistique

Apprentissage automatique :

1. observations d'un phénomène
2. construction d'un modèle de ce phénomène
3. prévisions et analyse du phénomène grâce au modèle

Apprentissage statistique

Apprentissage statistique =

Apprentissage automatique + statistique

Apprentissage automatique :

1. observations d'un phénomène
2. construction d'un modèle de ce phénomène
3. prévisions et analyse du phénomène grâce au modèle

... le tout automatiquement (sans intervention humaine)

Apprentissage statistique

Apprentissage statistique =

Apprentissage automatique + statistique

Apprentissage automatique :

1. observations d'un phénomène
2. construction d'un modèle de ce phénomène
3. prévisions et analyse du phénomène grâce au modèle

... le tout automatiquement (sans intervention humaine)

Statistique :

- ▶ formalisation du processus
- ▶ garanties sur sa qualité
- ▶ éventuellement suggestion de nouvelles techniques

Apprentissage automatique

- ▶ observations d'un phénomène \Rightarrow des données $\mathbf{z}_i \in \mathcal{Z}$
- ▶ deux grandes catégories de données :
 1. cas **non supervisé** :
 - ▶ pas de structure interne à \mathbf{z}
 - ▶ modélisation de la distribution de \mathbf{z} , recherche de classes homogènes, de règles d'association, etc.

Apprentissage automatique

- ▶ observations d'un phénomène \Rightarrow des données $\mathbf{z}_i \in \mathcal{Z}$
- ▶ deux grandes catégories de données :
 1. cas **non supervisé** :
 - ▶ pas de structure interne à \mathbf{z}
 - ▶ modélisation de la distribution de \mathbf{z} , recherche de classes homogènes, de règles d'association, etc.
 2. cas **supervisé**
 - ▶ $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$
 - ▶ modélisation du lien entre \mathbf{x} et \mathbf{y}
 - ▶ ... pour faire des prévisions : connaissant \mathbf{x} , on prédit \mathbf{y}

Apprentissage supervisé

- ▶ la difficulté dépend de la nature de \mathcal{Y}
 - ▶ $\mathcal{Y} = \mathbb{R}^q$: régression (multiple) de \mathbf{y} en \mathbf{x}
 - ▶ $\mathcal{Y} = \{1, \dots, q\}$: discrimination en q classes
 - ▶ $\mathcal{Y} = \{-1, 1\}$: discrimination en 2 classes
- ▶ ne pas en faire trop : inutile de modéliser la distribution de (\mathbf{x}, \mathbf{y}) pour de la discrimination
- ▶ remarque : *classification* (en anglais) \neq classification (en français)

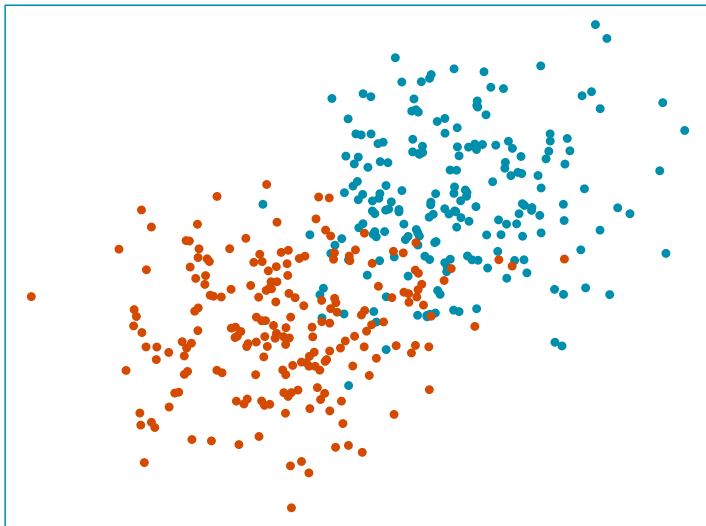
k plus proches voisins

- ▶ algorithme simple, relativement efficace, avec de bonnes propriétés théoriques
- ▶ données :
 - ▶ n observations $D_n = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, $y_i \in \{-1, 1\}$
 - ▶ \mathcal{X} est muni d'une distance d
 - ▶ un paramètre k entier strictement positif

k plus proches voisins

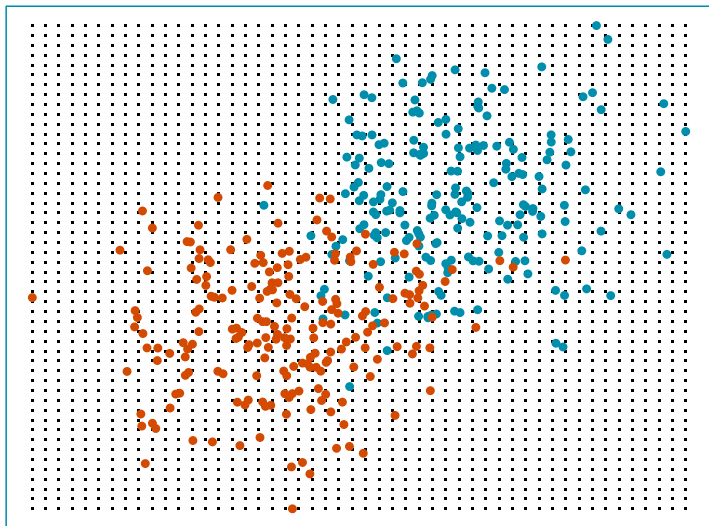
- ▶ algorithme simple, relativement efficace, avec de bonnes propriétés théoriques
- ▶ données :
 - ▶ n observations $D_n = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, $y_i \in \{-1, 1\}$
 - ▶ \mathcal{X} est muni d'une distance d
 - ▶ un paramètre k entier strictement positif
- ▶ algorithme pour une nouvelle observation \mathbf{x}
 - ▶ calcul des $d(\mathbf{x}, \mathbf{x}_i)$
 - ▶ tri par ordre croissant, $(j_i)_{i=1}^n$, $d(\mathbf{x}, \mathbf{x}_{j_i}) \leq d(\mathbf{x}, \mathbf{x}_{j_{i+1}})$
 - ▶ prédiction \mathbf{y} pour \mathbf{x} : la valeur majoritaire dans les k valeurs $\mathbf{y}_{j_1}, \dots, \mathbf{y}_{j_k}$
 - ▶ **remarque** : il faut éliminer les ambiguïtés (les cas $d(\mathbf{x}, \mathbf{x}_i) = d(\mathbf{x}, \mathbf{x}_k)$ pour $k \neq i$)
- ▶ applicable aussi pour q classes et pour $\mathcal{Y} = \mathbb{R}^q$

Exemple



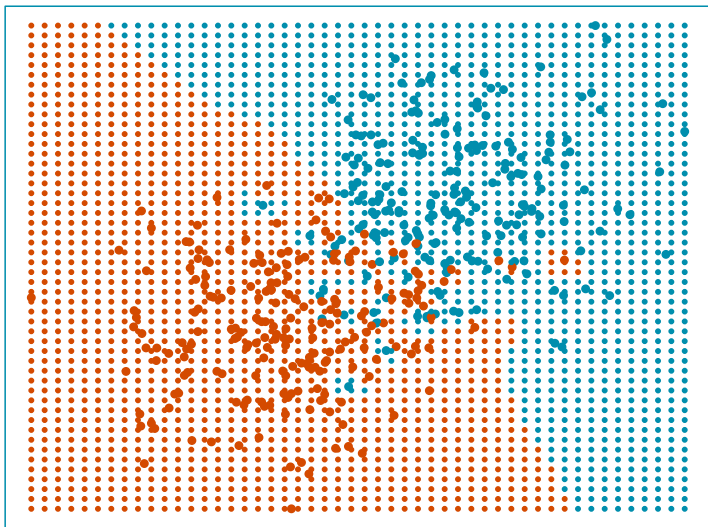
Données

Exemple



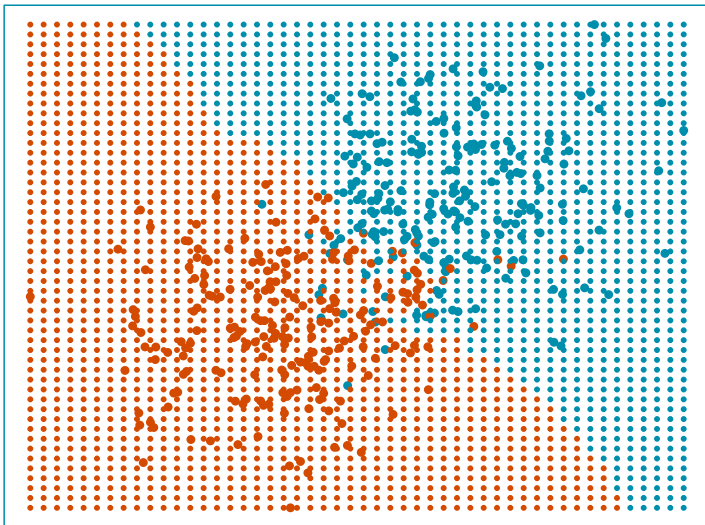
Nouveaux points

Exemple



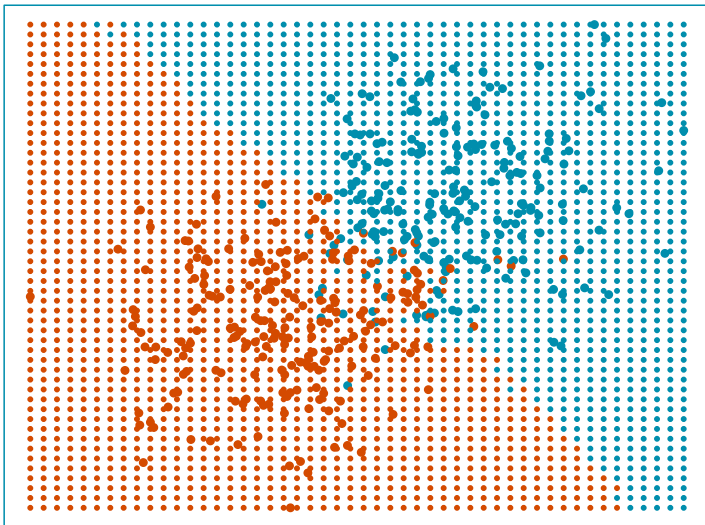
$$k = 1$$

Exemple



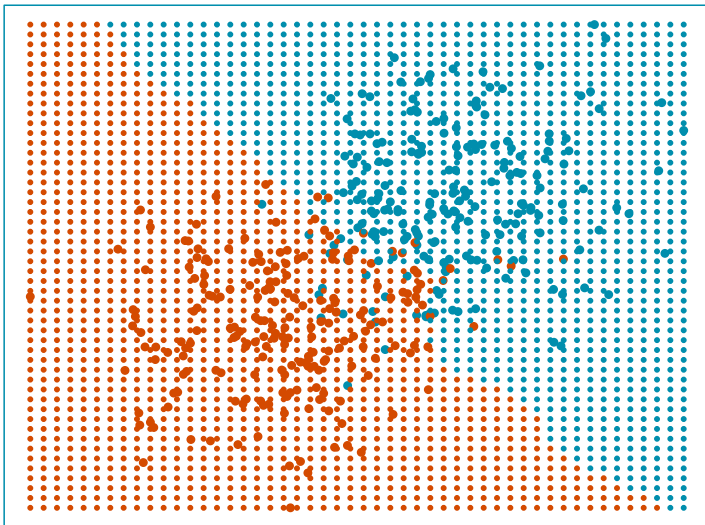
$$k = 5$$

Exemple



$k = 10$

Exemple



$k = 15$

Remarques

- ▶ petite valeur de k :
 - ▶ décision locale
 - ▶ « colle » aux données
 - ▶ modèle « complexe » (irrégulier)

Remarques

- ▶ petite valeur de k :
 - ▶ décision locale
 - ▶ « colle » aux données
 - ▶ modèle « complexe » (irrégulier)
- ▶ grande valeur de k :
 - ▶ décision plus globale
 - ▶ colle moins aux données
 - ▶ modèle plus simple

Remarques

- ▶ petite valeur de k :
 - ▶ décision locale
 - ▶ « colle » aux données
 - ▶ modèle « complexe » (irrégulier)
- ▶ grande valeur de k :
 - ▶ décision plus globale
 - ▶ colle moins aux données
 - ▶ modèle plus simple
- ▶ comment choisir k ?

Remarques

- ▶ petite valeur de k :
 - ▶ décision locale
 - ▶ « colle » aux données
 - ▶ modèle « complexe » (irrégulier)
- ▶ grande valeur de k :
 - ▶ décision plus globale
 - ▶ colle moins aux données
 - ▶ modèle plus simple
- ▶ comment choisir k ?
- ▶ quelles garanties sur la qualité du modèle ?

Remarques

- ▶ petite valeur de k :
 - ▶ décision locale
 - ▶ « colle » aux données
 - ▶ modèle « complexe » (irrégulier)
- ▶ grande valeur de k :
 - ▶ décision plus globale
 - ▶ colle moins aux données
 - ▶ modèle plus simple
- ▶ comment choisir k ?
- ▶ quelles garanties sur la qualité du modèle ?

L'apprentissage statistique cherche à donner des réponses à ce type de questions

Perceptrons multi-couches

- ▶ fonction sigmoïde σ de \mathbb{R} dans $[0, 1]$, croissante et telle que $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ par exemple : $\sigma(x) = 1/(1 + e^{-x})$

Perceptrons multi-couches

- ▶ fonction sigmoïde σ de \mathbb{R} dans $[0, 1]$, croissante et telle que $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ par exemple : $\sigma(x) = 1/(1 + e^{-x})$
- ▶ perceptrons à k neurones « cachés » :

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right\}$$

Perceptrons multi-couches

- ▶ fonction sigmoïde σ de \mathbb{R} dans $[0, 1]$, croissante et telle que $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ par exemple : $\sigma(x) = 1/(1 + e^{-x})$
- ▶ perceptrons à k neurones « cachés » :

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right\}$$

- ▶ choix du modèle en régression $\mathcal{Y} = \mathbb{R}$:
 - ▶ moindres carrés $g_k^* = \arg \min_{g \in \mathcal{G}(k)} \sum_{i=1}^n (\mathbf{y}_i - g(\mathbf{x}_i))^2$

Perceptrons multi-couches

- ▶ fonction sigmoïde σ de \mathbb{R} dans $[0, 1]$, croissante et telle que $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ par exemple : $\sigma(x) = 1/(1 + e^{-x})$
- ▶ perceptrons à k neurones « cachés » :

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right\}$$

- ▶ choix du modèle en régression $\mathcal{Y} = \mathbb{R}$:
 - ▶ moindres carrés $g_k^* = \arg \min_{g \in \mathcal{G}(k)} \sum_{i=1}^n (\mathbf{y}_i - g(\mathbf{x}_i))^2$
 - ▶ choix de k : ensemble de validation
 $k^* = \arg \min_{1 \leq k \leq K} \sum_{i=n+1}^{n+l} (\mathbf{y}_i - g_k^*(\mathbf{x}_i))^2$

Perceptrons multi-couches

- ▶ fonction sigmoïde σ de \mathbb{R} dans $[0, 1]$, croissante et telle que $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ par exemple : $\sigma(x) = 1/(1 + e^{-x})$
- ▶ perceptrons à k neurones « cachés » :

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right\}$$

- ▶ choix du modèle en régression $\mathcal{Y} = \mathbb{R}$:

- ▶ moindres carrés $g_k^* = \arg \min_{g \in \mathcal{G}(k)} \sum_{i=1}^n (\mathbf{y}_i - g(\mathbf{x}_i))^2$
- ▶ choix de k : ensemble de validation
 $k^* = \arg \min_{1 \leq k \leq K} \sum_{i=n+1}^{n+l} (\mathbf{y}_i - g_k^*(\mathbf{x}_i))^2$
- ▶ régularisation (*weight decay*) :

$$g_{k,\lambda}^* = \arg \min_{g \in \mathcal{G}(k)} \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - g(\mathbf{x}_i))^2 + \lambda \sum_{i=1}^k (c_i^2 + b_i^2 + \|\mathbf{a}_i\|^2)$$

Perceptrons multi-couches

- ▶ discrimination à 2 classes $\mathcal{Y} = \{0, 1\}$:

Perceptrons multi-couches

- ▶ discrimination à 2 classes $\mathcal{Y} = \{0, 1\}$:
 - ▶ perceptrons à valeurs dans $[0, 1]$:

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \tau \left(\sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right) \right\}$$

avec τ de \mathbb{R} dans $[0, 1]$

Perceptrons multi-couches

- ▶ discrimination à 2 classes $\mathcal{Y} = \{0, 1\}$:
 - ▶ perceptrons à valeurs dans $[0, 1]$:

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \tau \left(\sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right) \right\}$$

avec τ de \mathbb{R} dans $[0, 1]$

- ▶ maximum de vraisemblance : $\prod_{i=1}^n g(\mathbf{x}_i)^{y_i} (1 - g(\mathbf{x}_i))^{1-y_i}$

Perceptrons multi-couches

- ▶ discrimination à 2 classes $\mathcal{Y} = \{0, 1\}$:
 - ▶ perceptrons à valeurs dans $[0, 1]$:

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \tau \left(\sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right) \right\}$$

avec τ de \mathbb{R} dans $[0, 1]$

- ▶ maximum de vraisemblance : $\prod_{i=1}^n g(\mathbf{x}_i)^{y_i} (1 - g(\mathbf{x}_i))^{1-y_i}$
- ▶ quadratique en régression : maximum de vraisemblance pour un bruit gaussien

Perceptrons multi-couches

- ▶ discrimination à 2 classes $\mathcal{Y} = \{0, 1\}$:
 - ▶ perceptrons à valeurs dans $[0, 1]$:

$$\mathcal{G}(k) = \left\{ g(\mathbf{x}) = \tau \left(\sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0 \right) \right\}$$

avec τ de \mathbb{R} dans $[0, 1]$

- ▶ maximum de vraisemblance : $\prod_{i=1}^n g(\mathbf{x}_i)^{y_i} (1 - g(\mathbf{x}_i))^{1-y_i}$
 - ▶ quadratique en régression : maximum de vraisemblance pour un bruit gaussien
- ▶ aspects algorithmiques (généraux) :
 - ▶ problème optimisation complexe (non quadratique)
 - ▶ descente de gradient
 - ▶ calcul efficace du gradient par *rétro-propagation* (coût linéaire par rapport au nombre de paramètres du modèle)

Machines à vecteurs de support

► noyau :

- K de $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
- K symétrique
- K positive : $\sum_i^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- exemple $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$

Machines à vecteurs de support

► noyau :

► K de $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

► K symétrique

► K positive : $\sum_i^n \sum_j^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

► exemple $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$

► modèle choisi dans D_n

$$\mathcal{G}_{\mathbf{x}_1, \dots, \mathbf{x}_n} = \left\{ g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) \right\}$$

Machines à vecteurs de support

► noyau :

- K de $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
- K symétrique
- K positive : $\sum_i^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- exemple $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$

► modèle choisi dans D_n

$$\mathcal{G}_{\mathbf{x}_1, \dots, \mathbf{x}_n} = \left\{ g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) \right\}$$

► en discrimination à deux classes $\mathcal{Y} = \{-1, 1\}$:

$$g_\lambda^* = \arg \min_{g \in \mathcal{G}_{\mathbf{x}_1, \dots, \mathbf{x}_n}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \max(0, -\mathbf{y}_i(g(\mathbf{x}_i) + b)) + \lambda \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

Machines à vecteurs de support

- ▶ en fait, K engendre un **espace de Hilbert à noyau reproduisant**, \mathcal{H} , complété de

$$H = \left\{ g(\mathbf{x}) = \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \mathbf{x}); p \in \mathbb{N}, \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X} \right\}$$

muni du produit scalaire

$$\left\langle \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \cdot), \sum_{j=1}^m \beta_j K(\mathbf{x}'_j, \cdot) \right\rangle = \sum_{i=1}^p \sum_{j=1}^m \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}'_j)$$

Machines à vecteurs de support

- ▶ en fait, K engendre un **espace de Hilbert à noyau reproduisant**, \mathcal{H} , complété de

$$H = \left\{ g(\mathbf{x}) = \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \mathbf{x}); p \in \mathbb{N}, \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X} \right\}$$

muni du produit scalaire

$$\left\langle \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \cdot), \sum_{i=1}^m \beta_i K(\mathbf{x}'_i, \cdot) \right\rangle = \sum_{i=1}^p \sum_{j=1}^m \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}'_j)$$

- ▶ le choix de g est basé sur
 - ▶ le *hinge loss*, $c(u, v) = \max(0, -uv)$: majoration continue de $\mathbb{I}_{\{\text{signe}(u) \neq v\}}$
 - ▶ un terme de régularisation $\|g\|_{\mathcal{H}}^2$

Machines à vecteurs de support

Aspects algorithmiques

résoudre

$$\min_{g \in \mathcal{G}_{\mathbf{x}_1, \dots, \mathbf{x}_n}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \max(0, -\mathbf{y}_i(g(\mathbf{x}_i) + b)) + \lambda \|g\|_{\mathcal{H}}^2$$

revient à résoudre (pour $C = 1/(\lambda n)$)

$$\min_{g \in \mathcal{H}, \xi \in \mathbb{R}^n, b \in \mathbb{R}} \|g\|_{\mathcal{H}}^2 + C \sum_{i=1}^n \xi_i$$

avec $\mathbf{y}_i(g(\mathbf{x}_i) + b) \geq 1 - \xi_i$ et $\xi_i \geq 0$

qui est équivalent à

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{avec } \sum_{i=1}^n \alpha_i \mathbf{y}_i = 0, \quad 0 \leq \alpha_i \leq C$$

Boosting

Idée de base :

- ▶ s'appuyer sur des modèles très simples choisis dans \mathcal{C}
- ▶ les combiner linéairement, c.-à-d. choisir dans

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i f_i(\mathbf{x}); k \in \mathbb{N}, f_i \in \mathcal{C}, c_i \in \mathbb{R} \right\}$$

Boosting

Idée de base :

- ▶ s'appuyer sur des modèles très simples choisis dans \mathcal{C}
- ▶ les combiner linéairement, c.-à-d. choisir dans

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i f_i(\mathbf{x}); k \in \mathbb{N}, f_i \in \mathcal{C}, c_i \in \mathbb{R} \right\}$$

- ▶ en pratique, de façon itérative :
 - ▶ partir de $g_0 = f_0$
 - ▶ choisir $g_t = g_{t-1} + c_t f_t$ en optimisant un critère d'erreur par rapport à c_t et f_t

Ada Boost

Freund et Schapire, 1995

- ▶ discrimination à deux classes
- ▶ f_t est choisie de sorte à minimiser une erreur de classement pondérée $L_t(f) = \sum_{i=1}^n D_t(i) \mathbb{I}_{\{\text{signe}(f(\mathbf{x}_i)) \neq \mathbf{y}_i\}}$
- ▶ $c_t = \frac{1}{2} \ln \left(\frac{1 - L_t(f_t)}{L_t(f_t)} \right)$
- ▶ $D_1(i) = 1/n$ et

$$D_{t+1}(i) = \frac{D_t(i) \exp(-c_t \mathbf{y}_i f_t(\mathbf{x}_i))}{N_t}$$

- ▶ $g_t = \text{signe} \left(\sum_{t=1}^T c_t f_t \right)$
- ▶ en fait, Ada Boost cherche à optimiser $g = \text{signe}(h)$ par rapport à

$$\sum_{i=1}^n \exp(-\mathbf{y}_i h(\mathbf{x}_i))$$

Boosting régularisé

- ▶ même idée générale, mais on travaille dans l'enveloppe convexe de \mathcal{C}

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i f_i(\mathbf{x}); k \in \mathbb{N}, f_i \in \mathcal{C}, c_i \in [0, 1], \sum_i c_i = 1 \right\}$$

- ▶ g est choisi par minimisation de $\sum_{i=1}^n \phi(-\lambda \mathbf{y}_i g(\mathbf{x}_i))$:
 - ▶ ϕ est une fonction convexe comme par exemple \exp
 - ▶ λ est un paramètre de régularisation

Boosting régularisé

- ▶ même idée générale, mais on travaille dans l'enveloppe convexe de \mathcal{C}

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i f_i(\mathbf{x}); k \in \mathbb{N}, f_i \in \mathcal{C}, c_i \in [0, 1], \sum_i c_i = 1 \right\}$$

- ▶ g est choisi par minimisation de $\sum_{i=1}^n \phi(-\lambda \mathbf{y}_i g(\mathbf{x}_i))$:
 - ▶ ϕ est une fonction convexe comme par exemple \exp
 - ▶ λ est un paramètre de régularisation
- ▶ optimisation itérative :
 - ▶ choisir $g_1 = f_1$ dans \mathcal{C}
 - ▶ choisir f_t et c_t par optimisation de $g = (1 - c_t)g_{t-1} + c_t f_t$ par rapport à $\sum_{i=1}^n \phi(-\lambda \mathbf{y}_i g(\mathbf{x}_i))$
 - ▶ $g_t = (1 - c_t)g_{t-1} + c_t f_t$
 - ▶ en général, cette approche converge (Zhang, 2003)

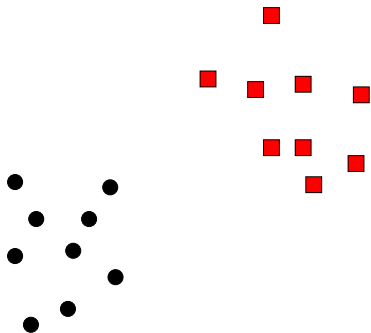
Marge en discrimination

- ▶ en discrimination à deux classes, le but devrait être d'éviter les erreurs : minimisation de $\sum_{i=1}^n \mathbb{I}_{\{\text{signe}(g(\mathbf{x}_i)) \neq y_i\}}$ pause
- ▶ mais souvent, on travaille sur $\phi(-\mathbf{y}_i g(\mathbf{x}_i))$ où $\mathbf{y}_i g(\mathbf{x}_i)$ est la **marge** du classifieur :
 - ▶ plus $\mathbf{y}_i g(\mathbf{x}_i)$ est grand, moins g est sensible à une erreur sur \mathbf{x}_i
 - ▶ plus $\mathbf{y}_i g(\mathbf{x}_i)$ est petit, plus g se trompe sur \mathbf{x}_i

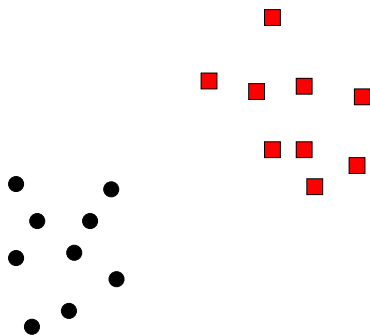
Marge en discrimination

- ▶ en discrimination à deux classes, le but devrait être d'éviter les erreurs : minimisation de $\sum_{i=1}^n \mathbb{I}_{\{\text{signe}(g(\mathbf{x}_i)) \neq y_i\}}$ pause
- ▶ mais souvent, on travaille sur $\phi(-\mathbf{y}_i g(\mathbf{x}_i))$ où $\mathbf{y}_i g(\mathbf{x}_i)$ est la **marge** du classifieur :
 - ▶ plus $\mathbf{y}_i g(\mathbf{x}_i)$ est grand, moins g est sensible à une erreur sur \mathbf{x}_i
 - ▶ plus $\mathbf{y}_i g(\mathbf{x}_i)$ est petit, plus g se trompe sur \mathbf{x}_i
- ▶ cas linéaire :
 - ▶ $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$
 - ▶ distance de \mathbf{x} à l'hyperplan $g(\mathbf{x}) = 0$: $\frac{|\langle \mathbf{w}, \mathbf{x} \rangle + b|}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$
 - ▶ distance orientée : $\frac{\mathbf{y}(\langle \mathbf{w}, \mathbf{x} \rangle + b)}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$
 - ▶ si on normalise $\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle} = 1$, la distance est $\mathbf{y}g(\mathbf{x})$ (la marge)

Marge maximale

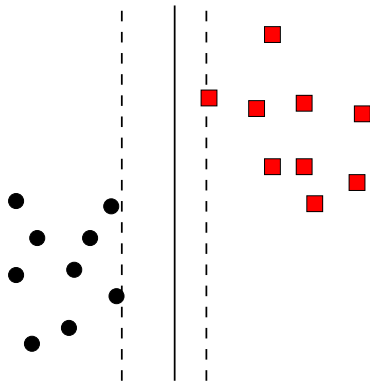


Marge maximale



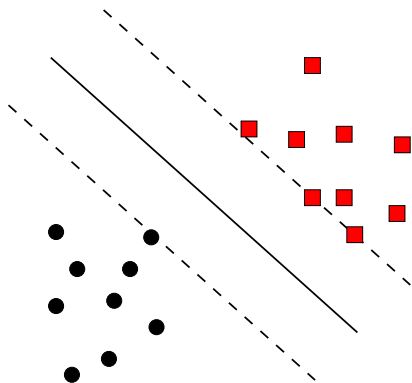
- ▶ Données linéairement séparables : une infinité de choix possibles

Marge maximale



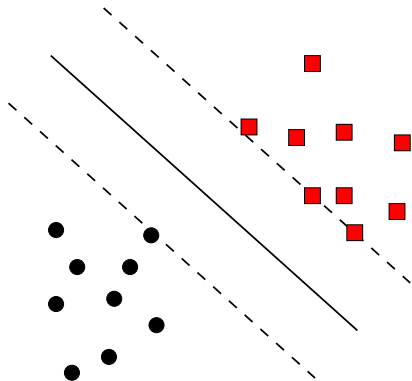
- ▶ Données linéairement séparables : une infinité de choix possibles
- ▶ Données proches du séparateur : petite « marge » \Rightarrow faible robustesse

Marge maximale



- ▶ Données linéairement séparables : une infinité de choix possibles
- ▶ Données proches du séparateur : petite « marge » \Rightarrow faible robustesse
- ▶ Un critère de choix possible : maximiser la marge

Marge maximale



- ▶ Données linéairement séparables : une infinité de choix possibles
- ▶ Données proches du séparateur : petite « marge » \Rightarrow faible robustesse
- ▶ Un critère de choix possible : maximiser la marge
- ▶ **Machine à vecteurs de support**

Marge maximale et MVS

- ▶ marge globale : $\min_i \frac{\mathbf{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle + b}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$

Marge maximale et MVS

- ▶ marge globale : $\min_i \frac{\mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$
- ▶ cas linéairement séparable :
 - ▶ normalisation par $\min_i \mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$ (linéairement séparable)
 - ▶ maximiser la marge revient alors à maximiser $\frac{1}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$ et donc à minimiser $\langle \mathbf{w}, \mathbf{w} \rangle$ (sous contrainte $\mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ pour tout i)

Marge maximale et MVS

- ▶ marge globale : $\min_i \frac{\mathbf{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle + b}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$
- ▶ cas linéairement séparable :
 - ▶ normalisation par $\min_i \mathbf{y}_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$ (linéairement séparable)
 - ▶ maximiser la marge revient alors à maximiser $\frac{1}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$ et donc à minimiser $\langle \mathbf{w}, \mathbf{w} \rangle$ (sous contrainte $\mathbf{y}_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ pour tout i)
- ▶ cas général :
 - ▶ $\mathbf{y}_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$ ($\xi_i \geq 0$)
 - ▶ minimiser $\langle \mathbf{w}, \mathbf{w} \rangle + C \sum_i \xi_i$ (compromis entre la marge et les erreurs)

Plan

Apprentissage automatique

- Principes

- Exemples d'algorithmes d'apprentissage

Théorie de l'apprentissage

- Formalisation

- Construction d'un modèle

- Pas de repas gratuit

Les données

Cas supervisé

- ▶ phénomène : couple de variables aléatoires
 $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$
- ▶ $Z = (X, Y)$ distribué selon P **inconnue**
- ▶ n observations :
 - ▶ $D_n = (X_i, Y_i)_{i=1}^n$
 - ▶ n copies **indépendantes** de (X, Y)
 - ▶ chaque copie est distribuée selon P
- ▶ but : construire une fonction g de \mathcal{X} dans \mathcal{Y} pour **prédire** Y à partir de X
- ▶ g_n : un modèle construit à partir de D_n

Qualité d'un modèle

- ▶ bon modèle : $\mathbf{y} \simeq g(\mathbf{x})$
- ▶ coût d'une prédiction, $c : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$
- ▶ **risque** d'un modèle

$$L(g) = \mathbb{E} \{c(g(X), Y)\},$$

espérance du coût par rapport à P

- ▶ risque optimal

$$L^* = \inf_g L(g)$$

- ▶ but : construire g_n de sorte que $L(g_n)$ soit proche de L^* (asymptotiquement)
- ▶ remarque : $L(g_n) = \mathbb{E} \{c(g_n(X), Y) \mid D_n\}$

Consistances

- ▶ **algorithme d'apprentissage** : méthode qui à une suite d'observations $(X_i, Y_i)_{i \geq 1}$ associe une suite de modèles $(g_i)_{i \geq 1}$ telle que g_n ne dépend que de $D_n = (X_i, Y_i)_{i=1}^n$

Consistances

- ▶ **algorithme d'apprentissage** : méthode qui à une suite d'observations $(X_i, Y_i)_{i \geq 1}$ associe une suite de modèles $(g_i)_{i \geq 1}$ telle que g_n ne dépend que de $D_n = (X_i, Y_i)_{i=1}^n$
- ▶ on cherche des résultats universels c.-à-d. indépendants de la loi P de (X, Y) :
 - ▶ **consistance universelle** : pour tout couple (X, Y) ,
$$\lim_{n \rightarrow \infty} \mathbb{E} \{L(g_n)\} = L^*$$
 - ▶ **consistance universelle forte** : pour tout couple (X, Y) ,
$$L(g_n) \xrightarrow{p.s.} L^*$$
 - ▶ **Probably Approximately Optimal** : $\forall \epsilon, \delta, \exists n(\epsilon, \delta)$ tel que
$$\forall n \geq n(\epsilon, \delta)$$

$$\mathbb{P} \{L(g_n) > L^* + \epsilon\} < \delta$$

Consistances

- ▶ **algorithme d'apprentissage** : méthode qui à une suite d'observations $(X_i, Y_i)_{i \geq 1}$ associe une suite de modèles $(g_i)_{i \geq 1}$ telle que g_n ne dépend que de $D_n = (X_i, Y_i)_{i=1}^n$
- ▶ on cherche des résultats universels c.-à-d. indépendants de la loi P de (X, Y) :
 - ▶ **consistance universelle** : pour tout couple (X, Y) ,
$$\lim_{n \rightarrow \infty} \mathbb{E} \{L(g_n)\} = L^*$$
 - ▶ **consistance universelle forte** : pour tout couple (X, Y) ,
$$L(g_n) \xrightarrow{p.s.} L^*$$
 - ▶ **Probably Approximately Optimal** : $\forall \epsilon, \delta, \exists n(\epsilon, \delta)$ tel que
$$\forall n \geq n(\epsilon, \delta) \quad \mathbb{P} \{L(g_n) > L^* + \epsilon\} < \delta$$
- ▶ remarque : si c est bornée, $\lim_{n \rightarrow \infty} \mathbb{E} \{L(g_n)\} = L^* \Leftrightarrow L(g_n) \xrightarrow{P} L^*$

Fonctions de coût

Les fonctions de coût dépendent de la nature du problème

- ▶ régression ($\mathcal{Y} = \mathbb{R}^q$) :
 - ▶ en général, coût quadratique $c(g(\mathbf{x}), \mathbf{y}) = \|g(\mathbf{x}) - \mathbf{y}\|^2$
 - ▶ modèle optimal $\eta(\mathbf{x}) = \mathbb{E}\{Y \mid X = \mathbf{x}\}$
 - ▶ d'autres solutions sont utilisées

Fonctions de coût

Les fonctions de coût dépendent de la nature du problème

- ▶ régression ($\mathcal{Y} = \mathbb{R}^q$) :
 - ▶ en général, coût quadratique $c(g(\mathbf{x}), \mathbf{y}) = \|g(\mathbf{x}) - \mathbf{y}\|^2$
 - ▶ modèle optimal $\eta(\mathbf{x}) = \mathbb{E}\{Y \mid X = \mathbf{x}\}$
 - ▶ d'autres solutions sont utilisées
- ▶ discrimination ($\mathcal{Y} = \{1, \dots, q\}$) :
 - ▶ coût 0/1, $c(g(\mathbf{x}), \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$
 - ▶ risque : probabilité de mauvais classement
 - ▶ modèle optimal (**classifieur de Bayes**) :
 - ▶ probabilités *a posteriori* : $\mathbb{P}\{Y = i \mid X = \mathbf{x}\}$
 - ▶ prédiction optimale : classe la plus probable
 - ▶ risque correspondant : **risque de Bayes**

Fonctions de coût

Les fonctions de coût dépendent de la nature du problème

- ▶ régression ($\mathcal{Y} = \mathbb{R}^q$) :
 - ▶ en général, coût quadratique $c(g(\mathbf{x}), \mathbf{y}) = \|g(\mathbf{x}) - \mathbf{y}\|^2$
 - ▶ modèle optimal $\eta(\mathbf{x}) = \mathbb{E}\{Y \mid X = \mathbf{x}\}$
 - ▶ d'autres solutions sont utilisées
- ▶ discrimination ($\mathcal{Y} = \{1, \dots, q\}$) :
 - ▶ coût 0/1, $c(g(\mathbf{x}), \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$
 - ▶ risque : probabilité de mauvais classement
 - ▶ modèle optimal (**classifieur de Bayes**) :
 - ▶ probabilités *a posteriori* : $\mathbb{P}\{Y = i \mid X = \mathbf{x}\}$
 - ▶ prédiction optimale : classe la plus probable
 - ▶ risque correspondant : **risque de Bayes**

P est inconnue : impossible de construire le modèle optimal

Construction d'un modèle

- ▶ méthode générique :
 1. choix d'une classe de modèles $\mathcal{G} = \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$
 2. choix de g_n dans \mathcal{G} par optimisation d'un critère
- ▶ \mathcal{G} peut dépendre de n ou même de D_n

Construction d'un modèle

- ▶ méthode générique :
 1. choix d'une classe de modèles $\mathcal{G} = \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$
 2. choix de g_n dans \mathcal{G} par optimisation d'un critère
- ▶ \mathcal{G} peut dépendre de n ou même de D_n
- ▶ exemples :
 - ▶ régression linéaire (\mathcal{X} Hilbert, $\mathcal{Y} = \mathbb{R}$)
 - ▶ $\mathcal{G} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle\}$
 - ▶ minimisation du critère des moindres carrés
$$\mathbf{w}_n = \arg \min_{\mathbf{w}} \sum_{i=1}^n (\mathbf{y}_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$
 - ▶ discrimination linéaire (\mathcal{X} Hilbert, $\mathcal{Y} = \{1, \dots, c\}$)
 - ▶ $\mathcal{G} = \{\mathbf{x} \mapsto \arg \max_k (W\mathbf{x})_k\}$
 - ▶ maximisation du rapport entre la variance inter classes et la variance intra classes (Fisher)

Construction d'un modèle

- ▶ méthode générique :
 1. choix d'une classe de modèles $\mathcal{G} = \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$
 2. choix de g_n dans \mathcal{G} par optimisation d'un critère
- ▶ \mathcal{G} peut dépendre de n ou même de D_n
- ▶ exemples :
 - ▶ régression linéaire (\mathcal{X} Hilbert, $\mathcal{Y} = \mathbb{R}$)
 - ▶ $\mathcal{G} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle\}$
 - ▶ minimisation du critère des moindres carrés
$$\mathbf{w}_n = \arg \min_{\mathbf{w}} \sum_{i=1}^n (\mathbf{y}_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$
 - ▶ discrimination linéaire (\mathcal{X} Hilbert, $\mathcal{Y} = \{1, \dots, c\}$)
 - ▶ $\mathcal{G} = \{\mathbf{x} \mapsto \arg \max_k (W\mathbf{x})_k\}$
 - ▶ maximisation du rapport entre la variance inter classes et la variance intra classes (Fisher)
- ▶ mais aussi des méthodes *ad hoc* :
 - ▶ plus proches voisins
 - ▶ arbres
 - ▶ noyaux

Critère à optimiser

- ▶ idéalement, on aimerait choisir g_n à partir de $L(g_n)$ mais comme P est inconnue, c'est impossible

Critère à optimiser

- ▶ idéalement, on aimerait choisir g_n à partir de $L(g_n)$ mais comme P est inconnue, c'est impossible
- ▶ une solution possible, le **risque empirique** : moyenne du coût sur les données

$$L_n(g) = \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i)$$

- ▶ pour g **fixé**, $\lim_{n \rightarrow \infty} L_n(g) = L(g)$ (loi des grands nombres)

Critère à optimiser

- ▶ idéalement, on aimerait choisir g_n à partir de $L(g_n)$ mais comme P est inconnue, c'est impossible
- ▶ une solution possible, le **risque empirique** : moyenne du coût sur les données

$$L_n(g) = \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i)$$

- ▶ pour g fixé, $\lim_{n \rightarrow \infty} L_n(g) = L(g)$ (loi des grands nombres)
- ▶ **mais ça ne suffit pas** : g_n change avec n
- ▶ questions fondamentales
 - ▶ peut on évaluer la qualité de g à partir de $L_n(g)$?
 - ▶ peut on choisir g_n à partir de L_n ?
 - ▶ peut on choisir g avec autre chose que L_n ?

Minimisation du risque empirique

- ▶ une méthode générique d'apprentissage :
 - ▶ choix d'une classe de modèles $\mathcal{G} = \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$
 - ▶ choix de g_n^* dans \mathcal{G} par

$$g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$$

- ▶ risque minimal sur \mathcal{G} , $L_{\mathcal{G}}^* = \inf_{g \in \mathcal{G}} L(g)$
- ▶ est-ce que cela fonctionne ?
 - ▶ $\mathbb{E} \{L(g_n^*)\} \rightarrow L_{\mathcal{G}}^*$?
 - ▶ $\mathbb{E} \{L(g_n^*)\} \rightarrow L^*$?

Minimisation du risque empirique

- ▶ une méthode générique d'apprentissage :
 - ▶ choix d'une classe de modèles $\mathcal{G} = \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$
 - ▶ choix de g_n^* dans \mathcal{G} par

$$g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$$

- ▶ risque minimal sur \mathcal{G} , $L_{\mathcal{G}}^* = \inf_{g \in \mathcal{G}} L(g)$
- ▶ est-ce que cela fonctionne ?
 - ▶ $\mathbb{E} \{L(g_n^*)\} \rightarrow L_{\mathcal{G}}^*$?
 - ▶ $\mathbb{E} \{L(g_n^*)\} \rightarrow L^*$?
- ▶ stratégie d'analyse, borner les quantités suivantes :

$L(g) - L_{\mathcal{G}}^*$	erreur dans la classe
$ L(g) - L_n(g) $	erreur d'estimation
$L_{\mathcal{G}}^* - L^*$	erreur d'approximation

Difficulté

- ▶ nécessite un contrôle de la « puissance » du modèle (de la « taille » de \mathcal{G})
- ▶ exemple de cas pathologique :
 - ▶ discrimination en c classes
 - ▶ \mathcal{G} : fonctions constantes par morceaux sur une partition de \mathcal{X}
 - ▶ $L_n(g_n^*) = 0$ (par exemple avec le classifieur du plus proche voisin)
 - ▶ mais $L(g_n^*) > 0$ (dès que $L^* > 0$!)
- ▶ conséquence classique : $L_{\mathcal{G}}^* > L^*$

$$\inf_{g \in \mathcal{G}} L(g) > \inf_g L(g)$$

Contrôle de complexité

Adapter la complexité du modèle aux données

- ▶ soit au niveau de la classe de modèles :
 - ▶ \mathcal{G}_n pour D_n
 - ▶ « puissance » croissante avec n

Contrôle de complexité

Adapter la complexité du modèle aux données

- ▶ soit au niveau de la classe de modèles :
 - ▶ \mathcal{G}_n pour D_n
 - ▶ « puissance » croissante avec n
- ▶ soit au niveau du critère à optimiser :
 - ▶ risque empirique (ou autre) plus un terme pénalisant les modèles complexes
 - ▶ **minimisation structurelle du risque** : \mathcal{G} est choisie comme l'union de classes de plus en plus complexes, avec une valeur de pénalité pour chaque classe
 - ▶ **régularisation** : la pénalisation s'appuie sur une propriété du modèle (dérivée d'ordre 2 par exemple)

Pas de repas gratuit

- ▶ nombreux résultats négatifs : beaucoup de propriétés ne peuvent pas être universelles

Pas de repas gratuit

- ▶ nombreux résultats négatifs : beaucoup de propriétés ne peuvent pas être universelles
- ▶ il y a toujours **une mauvaise distribution** :
 - ▶ cadre de la discrimination à deux classes
 - ▶ algorithme d'apprentissage fixé
 - ▶ pour tout $\epsilon > 0$ et tout n , il existe (X, Y) telle que $L^* = 0$ et

$$\mathbb{E} \{L(g_n)\} \geq \frac{1}{2} - \epsilon$$

Pas de repas gratuit

- ▶ nombreux résultats négatifs : beaucoup de propriétés ne peuvent pas être universelles
- ▶ il y a toujours **une mauvaise distribution** :
 - ▶ cadre de la discrimination à deux classes
 - ▶ algorithme d'apprentissage fixé
 - ▶ pour tout $\epsilon > 0$ et tout n , il existe (X, Y) telle que $L^* = 0$ et

$$\mathbb{E} \{L(g_n)\} \geq \frac{1}{2} - \epsilon$$

- ▶ on peut converger **arbitrairement lentement** :
 - ▶ (a_n) une suite décroissante de limite nulle $a_1 \leq 1/16$
 - ▶ algorithme d'apprentissage fixé (toujours en discrimination)
 - ▶ il existe (X, Y) telle que $L^* = 0$ et

$$\mathbb{E} \{L(g_n)\} \geq a_n$$

Pas de repas gratuit

- ▶ estimer le risque optimal est **difficile** :
 - ▶ discrimination
 - ▶ estimateur de L^* , \hat{L}_n
 - ▶ pour tout $\epsilon > 0$ et tout n , il existe (X, Y) telle que

$$\mathbb{E} \left\{ |\hat{L}_n - L^*| \right\} \geq \frac{1}{4} - \epsilon$$

Pas de repas gratuit

- ▶ estimer le risque optimal est **difficile** :
 - ▶ discrimination
 - ▶ estimateur de L^* , \hat{L}_n
 - ▶ pour tout $\epsilon > 0$ et tout n , il existe (X, Y) telle que

$$\mathbb{E} \left\{ |\hat{L}_n - L^*| \right\} \geq \frac{1}{4} - \epsilon$$

- ▶ la **régression est plus difficile que la discrimination** :
 - ▶ η_n estimateur de $\eta = \mathbb{E} \{ Y|X \}$ telle que

$$\lim_{n \rightarrow \infty} \mathbb{E} \left\{ (\eta_n(X) - \eta(X))^2 \right\} = 0$$

- ▶ pour $g_n(\mathbf{x}) = \mathbb{I}_{\{\eta_n(\mathbf{x}) > 1/2\}}$, on a

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E} \{ L(g_n) \} - L^*}{\sqrt{\mathbb{E} \{ (\eta_n(X) - \eta(X))^2 \}}} = 0$$

Deuxième partie II

Le risque empirique

Plan

Concentration

Inégalité de Hoeffding

Bornes uniformes

Dimension de Vapnik-Chervonenkis

Définition

Application à la discrimination

Preuve

Nombres de couverture

Définition et résultat

Calcul des nombres de couverture

Résumé

Risque empirique

- ▶ la qualité du risque empirique est liée à la **concentration** de la moyenne de $c(g(X), Y)$ autour de son espérance

$$L_n(g) - L(g) = \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i) - \mathbb{E} \{c(g(X), Y)\}$$

Risque empirique

- ▶ la qualité du risque empirique est liée à la **concentration** de la moyenne de $c(g(X), Y)$ autour de son espérance

$$L_n(g) - L(g) = \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i) - \mathbb{E} \{c(g(X), Y)\}$$

- ▶ de nombreux résultats de concentration existent, par exemple

- ▶ Markov $\mathbb{P} \{|X| \geq t\} \leq \frac{\mathbb{E}\{|X|\}}{t}$
- ▶ Bienaymé-Tchebychev $\mathbb{P} \{|X - \mathbb{E}\{X\}| \geq t\} \leq \frac{\text{Var}(X)}{t^2}$ et si les X_i sont indépendantes et à valeurs réelles, avec $S_n = \sum_{i=1}^n X_i$

$$\mathbb{P} \{|S_n - \mathbb{E}\{S_n\}| \geq t\} \leq \frac{\sum_{i=1}^n \text{Var}(X_i)}{t^2}$$

Inégalité de Hoeffding

Hoeffding, 1963

hypothèses

- ▶ X_1, \dots, X_n n v.a. indépendantes
- ▶ $X_i \in [a_i, b_i]$
- ▶ $S_n = \sum_{i=1}^n X_i$

résultat

$$\mathbb{P}\{S_n - \mathbb{E}\{S_n\} \geq \epsilon\} \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$$
$$\mathbb{P}\{S_n - \mathbb{E}\{S_n\} \leq -\epsilon\} \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$$

Loi des grands nombres quantitative

Application

- ▶ coût borné $c(u, v) \in [a, b]$ pour tout (u, v) :
 - ▶ régression avec Y bornée
 - ▶ discrimination $[a, b] = [0, 1]$
- ▶ $U_i = \frac{1}{n}c(g(X_i), Y_i)$, $U_i \in [\frac{a}{n}, \frac{b}{n}]$
- ▶ g doit être indépendant des (X_i, Y_i) (sinon les U_i ne sont plus indépendantes !)
- ▶ $\sum_{i=1}^n (b_i - a_i)^2 = \frac{(b-a)^2}{n}$
- ▶ on a donc

$$\mathbb{P} \{ |L_n(g) - L(g)| \geq \epsilon \} \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

- ▶ soit $L_n(g) \xrightarrow{P} L(g)$ et aussi $L_n(g) \xrightarrow{p.s.} L(g)$ (par Borel Cantelli)

Limitation

- ▶ g doit être indépendant des (X_i, Y_i)
- ▶ intuitivement :
 - ▶ $\delta = 2e^{-2n\epsilon^2/(b-a)^2}$
 - ▶ pour chaque g , la probabilité de tomber sur ensemble D_n sur lequel

$$|L_n(g) - L(g)| \leq (b - a) \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

est d'au moins $1 - \delta$

- ▶ mais les D_n « corrects » sont différents pour chaque g
- ▶ pour D_n fixé, la borne n'est valide que pour certains g
- ▶ pas de justification direct de $g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$

Preuve de l'inégalité de Hoeffding

- ▶ méthode de majoration de Chernoff (application de l'inégalité de Markov) :

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}\{e^{sX}\}}{e^{st}}$$

on contrôle le majorant grâce à s

- ▶ lemme : si $\mathbb{E}\{X\} = 0$ et $X \in [a, b]$, alors pour tout s ,

$$\mathbb{E}\{e^{sX}\} \leq e^{\frac{s^2(b-a)^2}{8}}$$

- ▶ convexité de $e \Rightarrow \mathbb{E}\{e^{sX}\} \leq e^{\phi(s(b-a))}$
- ▶ puis majoration de ϕ
- ▶ enfin on applique la méthode de Chernoff à $X = S_n - \mathbb{E}\{S_n\}$

Preuve de l'inégalité de Hoeffding

$$\begin{aligned}\mathbb{P} \{ \mathbf{S}_n - \mathbb{E} \{ \mathbf{S}_n \} \geq \epsilon \} &\leq e^{-s\epsilon} \mathbb{E} \left\{ e^{s \sum_{i=1}^n (X_i - \mathbb{E}\{X_i\})} \right\} \\ &= e^{-s\epsilon} \prod_{i=1}^n \mathbb{E} \left\{ e^{s(X_i - \mathbb{E}\{X_i\})} \right\} \text{ (indépendance)} \\ &\leq e^{-s\epsilon} \prod_{i=1}^n e^{\frac{s^2(b_i - a_i)^2}{8}} \text{ (lemme)} \\ &= e^{-s\epsilon} e^{\frac{s^2 \sum_{i=1}^n (b_i - a_i)^2}{8}} \\ &= e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2} \text{ (minimisation)}\end{aligned}$$

avec $s = 4\epsilon / \sum_{i=1}^n (b_i - a_i)^2$

Erreur uniforme

- ▶ $g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$
- ▶ $L(g_n^*) - L^* = [L(g_n^*) - \inf_{g \in \mathcal{G}} L(g)] + [\inf_{g \in \mathcal{G}} L(g) - L^*]$
erreur d'estimation + erreur d'approximation
- ▶ erreur uniforme

$$|L(g_n^*) - L_n(g_n^*)| \leq \sup_{g \in \mathcal{G}} |L_n(g) - L(g)|$$

$$\begin{aligned} L(g_n^*) - \inf_{g \in \mathcal{G}} L(g) &= L(g_n^*) - L_n(g_n^*) + L_n(g_n^*) - \inf_{g \in \mathcal{G}} L(g) \\ &\leq L(g_n^*) - L_n(g_n^*) + \sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \\ &\leq 2 \sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \end{aligned}$$

- ▶ il « suffit » donc d'une loi des grands nombres **uniforme** pour annuler asymptotiquement l'erreur d'estimation

Ensemble fini de modèles

- *union bound* : $\mathbb{P}\{A \text{ ou } B\} \leq \mathbb{P}\{A\} + \mathbb{P}\{B\}$ et donc

$$\mathbb{P}\left\{\max_{1 \leq i \leq m} U_i \geq \epsilon\right\} \leq \sum_{i=1}^m \mathbb{P}\{U_i \geq \epsilon\}$$

Ensemble fini de modèles

- ▶ *union bound* : $\mathbb{P}\{A \text{ ou } B\} \leq \mathbb{P}\{A\} + \mathbb{P}\{B\}$ et donc

$$\mathbb{P}\left\{\max_{1 \leq i \leq m} U_i \geq \epsilon\right\} \leq \sum_{i=1}^m \mathbb{P}\{U_i \geq \epsilon\}$$

- ▶ donc pour \mathcal{G} fini

$$\mathbb{P}\left\{\sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \geq \epsilon\right\} \leq 2|\mathcal{G}|e^{-2n\epsilon^2/(b-a)^2}$$

et donc $L(g_n^*) \xrightarrow{p.s.} \inf_{g \in \mathcal{G}} L(g)$ (mais en général $\inf_{g \in \mathcal{G}} L(g) > L^*$)

Puissance et borne

avec une probabilité au moins $1 - \delta$

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 2(b - a) \sqrt{\frac{\log |\mathcal{G}| + \log \frac{2}{\delta}}{2n}}$$

- ▶ $\inf_{g \in \mathcal{G}} L(g)$ décroît avec $|\mathcal{G}|$ (plus de choix)
- ▶ mais le second terme augmente avec $|\mathcal{G}|$: compromis puissance/qualité de l'estimation
- ▶ remarque : $\log |\mathcal{G}|$ est le nombre de bits pour coder le choix du modèle

Plan

Concentration

Inégalité de Hoeffding

Bornes uniformes

Dimension de Vapnik-Chervonenkis

Définition

Application à la discrimination

Preuve

Nombres de couverture

Définition et résultat

Calcul des nombres de couverture

Résumé

Ensemble infini de modèles

- ▶ les résultats obtenus ne s'appliquent qu'à \mathcal{G} fini
- ▶ or :
 - ▶ $\mathcal{G}_{\text{lin}} = \{\mathbf{x} \mapsto \arg \max_k (W\mathbf{x})_k\}$ (modèle linéaire) est infini
 - ▶ $\inf_{g \in \mathcal{G}_{\text{lin}}} L(g) > 0$ en général, même quand $L^* = 0$
(problèmes non linéaires)

Ensemble infini de modèles

- ▶ les résultats obtenus ne s'appliquent qu'à \mathcal{G} fini
- ▶ or :
 - ▶ $\mathcal{G}_{\text{lin}} = \{\mathbf{x} \mapsto \arg \max_k (W\mathbf{x})_k\}$ (modèle linéaire) est infini
 - ▶ $\inf_{g \in \mathcal{G}_{\text{lin}}} L(g) > 0$ en général, même quand $L^* = 0$ (problèmes non linéaires)
- ▶ une solution : mesurer la **capacité** d'un ensemble de modèles plutôt que son cardinal
 - ▶ problème le plus simple : discrimination à deux classes
 - ▶ un modèle : un ensemble (mesurable) $A \subset \mathcal{X}$
 - ▶ capacité : richesse des « formes » disponibles dans \mathcal{G}
 - ▶ au sens des données : si $A \cap D_n = B \cap D_n$, alors en terme de capacité A et B comptent pour un !

Fonction de croissance

- ▶ cadre abstrait : \mathcal{F} un ensemble de fonctions mesurables de \mathbb{R}^d dans $\{0, 1\}$ (i.e., \sim un ensemble de parties mesurables de \mathbb{R}^d)

Fonction de croissance

- ▶ cadre abstrait : \mathcal{F} un ensemble de fonctions mesurables de \mathbb{R}^d dans $\{0, 1\}$ (i.e., \sim un ensemble de parties mesurables de \mathbb{R}^d)
- ▶ pour $\mathbf{z}_1 \in \mathbb{R}^d, \dots, \mathbf{z}_n \in \mathbb{R}^d$, on pose

$$\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n} = \{u \in \{0, 1\}^n \mid \exists f \in \mathcal{F}, u = (f(\mathbf{z}_1), \dots, f(\mathbf{z}_n))\}$$

- ▶ interprétation : chaque u décrit une partition de $\mathbf{z}_1, \dots, \mathbf{z}_n$ en deux classes et $\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}$ est l'ensemble des partitions réalisables par \mathcal{F}

Fonction de croissance

- ▶ cadre abstrait : \mathcal{F} un ensemble de fonctions mesurables de \mathbb{R}^d dans $\{0, 1\}$ (i.e., \sim un ensemble de parties mesurables de \mathbb{R}^d)
- ▶ pour $\mathbf{z}_1 \in \mathbb{R}^d, \dots, \mathbf{z}_n \in \mathbb{R}^d$, on pose

$$\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n} = \{u \in \{0, 1\}^n \mid \exists f \in \mathcal{F}, u = (f(\mathbf{z}_1), \dots, f(\mathbf{z}_n))\}$$

- ▶ interprétation : chaque u décrit une partition de $\mathbf{z}_1, \dots, \mathbf{z}_n$ en deux classes et $\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}$ est l'ensemble des partitions réalisables par \mathcal{F}
- ▶ fonction de croissance

$$S_{\mathcal{F}}(n) = \sup_{\mathbf{z}_1 \in \mathbb{R}^d, \dots, \mathbf{z}_n \in \mathbb{R}^d} |\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}|$$

- ▶ interprétation : nombre maximal de partitions distinctes réalisables par \mathcal{F}

Dimension de Vapnik-Chervonenkis

- ▶ $\mathcal{S}_{\mathcal{F}}(n) \leq 2^n$
- ▶ vocabulaire :
 - ▶ $\mathcal{S}_{\mathcal{F}}(n)$ est le n -ième **coefficient de pulvérisation** (*shatter coefficient*) de \mathcal{F}
 - ▶ si $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = 2^n$, \mathcal{F} **pulvérise** $\mathbf{z}_1, \dots, \mathbf{z}_n$ (*shatters*)

Dimension de Vapnik-Chervonenkis

- ▶ $\mathcal{S}_{\mathcal{F}}(n) \leq 2^n$
- ▶ vocabulaire :
 - ▶ $\mathcal{S}_{\mathcal{F}}(n)$ est le n -ième **coefficient de pulvérisation** (*shatter coefficient*) de \mathcal{F}
 - ▶ si $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = 2^n$, \mathcal{F} **pulvérise** $\mathbf{z}_1, \dots, \mathbf{z}_n$ (*shatters*)
- ▶ **dimension de Vapnik-Chervonenkis**

$$VCdim(\mathcal{F}) = \sup\{n \in \mathbb{N}^+ \mid \mathcal{S}_{\mathcal{F}}(n) = 2^n\}$$

Dimension de Vapnik-Chervonenkis

- ▶ $\mathcal{S}_{\mathcal{F}}(n) \leq 2^n$
- ▶ vocabulaire :
 - ▶ $\mathcal{S}_{\mathcal{F}}(n)$ est le n -ième **coefficient de pulvérisation** (*shatter coefficient*) de \mathcal{F}
 - ▶ si $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = 2^n$, \mathcal{F} **pulvérise** $\mathbf{z}_1, \dots, \mathbf{z}_n$ (*shatters*)
- ▶ **dimension de Vapnik-Chervonenkis**

$$VCdim(\mathcal{F}) = \sup\{n \in \mathbb{N}^+ \mid \mathcal{S}_{\mathcal{F}}(n) = 2^n\}$$

- ▶ interprétation :
 - ▶ si $\mathcal{S}_{\mathcal{F}}(n) < 2^n$:
 - ▶ **pour tout** $\mathbf{z}_1, \dots, \mathbf{z}_n$, il existe une partition $u \in \{0, 1\}^n$, telle que $u \notin \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}$
 - ▶ pour tout ensemble d'observations, \mathcal{F} peut être prise en défaut
 - ▶ si $\mathcal{S}_{\mathcal{F}}(n) = 2^n$: il existe au moins un ensemble $\mathbf{z}_1, \dots, \mathbf{z}_n$ que \mathcal{F} pulvérise

Exemple

- ▶ points dans \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$

Exemple

▶ points dans \mathbb{R}^2

▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$

×

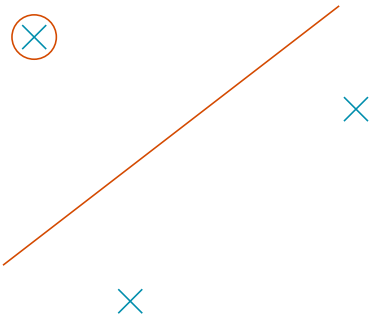
▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$

×

×

Exemple

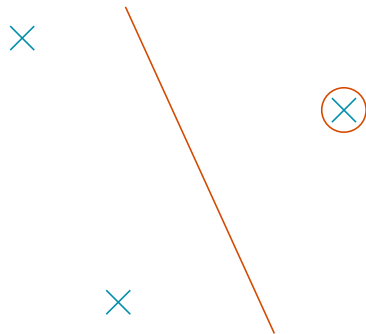
- ▶ points dans \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$



▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$

Exemple

- ▶ points dans \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$



▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$

Exemple

▶ points dans \mathbb{R}^2

▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$

×

▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$

×

⊗

Exemple

- ▶ points dans \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$



- ▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$
- ▶ même si on a parfois $|\mathcal{F}_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3}| < 2^3$

Exemple

- ▶ points dans \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$

×

×

×

×

- ▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$
- ▶ même si on a parfois $|\mathcal{F}_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3}| < 2^3$
- ▶ $\mathcal{S}_{\mathcal{F}}(4) < 2^4$

Exemple

- ▶ points dans \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$



- ▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$
- ▶ même si on a parfois $|\mathcal{F}_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3}| < 2^3$
- ▶ $\mathcal{S}_{\mathcal{F}}(4) < 2^4$
- ▶ $VCdim(\mathcal{F}) = 3$

Modèles linéaires (généralisés)

résultat

si \mathcal{G} est un espace vectoriel de fonctions définies sur \mathbb{R}^d de dimension p alors

$$VCdim \left(\left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(z) = \mathbb{I}_{\{g(z) \geq 0\}} \right\} \right) \leq p$$

Modèles linéaires (généralisés)

résultat

si \mathcal{G} est un espace vectoriel de fonctions définies sur \mathbb{R}^d de dimension p alors

$$VCdim \left(\left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(z) = \mathbb{I}_{\{g(z) \geq 0\}} \right\} \right) \leq p$$

preuve

- ▶ soit z_1, \dots, z_{p+1} , on considère F de \mathcal{G} dans \mathbb{R}^m défini par $F(g) = (g(z_1), \dots, g(z_{p+1}))$

Modèles linéaires (généralisés)

résultat

si \mathcal{G} est un espace vectoriel de fonctions définies sur \mathbb{R}^d de dimension p alors

$$VCdim \left(\left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(z) = \mathbb{I}_{\{g(z) \geq 0\}} \right\} \right) \leq p$$

preuve

- ▶ soit z_1, \dots, z_{p+1} , on considère F de \mathcal{G} dans \mathbb{R}^m défini par $F(g) = (g(z_1), \dots, g(z_{p+1}))$
- ▶ $\dim(F(\mathcal{G})) \leq p$ donc il existe un $\gamma = (\gamma_1, \dots, \gamma_{p+1})$ non nul, tel que $\sum_{i=1}^{p+1} \gamma_i g(z_i) = 0$

Modèles linéaires (généralisés)

résultat

si \mathcal{G} est un espace vectoriel de fonctions définies sur \mathbb{R}^d de dimension p alors

$$VCdim \left(\left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(z) = \mathbb{I}_{\{g(z) \geq 0\}} \right\} \right) \leq p$$

preuve

- ▶ soit z_1, \dots, z_{p+1} , on considère F de \mathcal{G} dans \mathbb{R}^m défini par $F(g) = (g(z_1), \dots, g(z_{p+1}))$
- ▶ $\dim(F(\mathcal{G})) \leq p$ donc il existe un $\gamma = (\gamma_1, \dots, \gamma_{p+1})$ non nul, tel que $\sum_{i=1}^{p+1} \gamma_i g(z_i) = 0$
- ▶ mais si z_1, \dots, z_{p+1} est pulvérisé, il existe g_j tel que $g_j(z_i) = \delta_{ij}$ et donc $\gamma_j = 0$ pour tout j

Modèles linéaires (généralisés)

résultat

si \mathcal{G} est un espace vectoriel de fonctions définies sur \mathbb{R}^d de dimension p alors

$$VCdim \left(\left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(z) = \mathbb{I}_{\{g(z) \geq 0\}} \right\} \right) \leq p$$

preuve

- ▶ soit z_1, \dots, z_{p+1} , on considère F de \mathcal{G} dans \mathbb{R}^m défini par $F(g) = (g(z_1), \dots, g(z_{p+1}))$
- ▶ $\dim(F(\mathcal{G})) \leq p$ donc il existe un $\gamma = (\gamma_1, \dots, \gamma_{p+1})$ non nul, tel que $\sum_{i=1}^{p+1} \gamma_i g(z_i) = 0$
- ▶ mais si z_1, \dots, z_{p+1} est pulvérisé, il existe g_j tel que $g_j(z_i) = \delta_{ij}$ et donc $\gamma_j = 0$ pour tout j
- ▶ aucun z_1, \dots, z_{p+1} n'est pulvérisé

VC dimension \neq nombre de paramètres

- ▶ linéaire généralisé $\mathcal{F} = \{f(z) = \mathbb{I}_{\{\sum_{i=1}^p w_i \phi_i(z) \geq 0\}}\}$
 $VCdim(\mathcal{F}) = p$

VC dimension \neq nombre de paramètres

- ▶ linéaire généralisé $\mathcal{F} = \{f(z) = \mathbb{I}_{\{\sum_{i=1}^p w_i \phi_i(z) \geq 0\}}\}$

$$VCdim(\mathcal{F}) = p$$

- ▶ c'est faux en général :

- ▶ $VCdim(\{f(z) = \mathbb{I}_{\{\sin(tz) \geq 0\}}\}) = \infty$

- ▶ réseau de neurones :

$$\mathcal{G} = \left\{ g(z) = T \left(\beta_0 + \sum_{k=1}^h \beta_k T \left(\alpha_{k0} + \sum_{j=1}^d \alpha_{kj} z_j \right) \right) \right\}$$

- ▶ $T(a) = \mathbb{I}_{\{a \geq 0\}}$

- ▶ $VCdim(\mathcal{G}) \geq W \log_2(h/4)/32$ avec $W = dh + 2h + 1$
(nombre de paramètres)

Convergence uniforme

Vapnik et Chervonenkis (1971)

hypothèses

- ▶ n v.a. i.i.d. Z_1, \dots, Z_n à valeurs dans \mathbb{R}^d
- ▶ $Pf = \mathbb{E} \{f(Z_1)\}$ et $P_n f = \frac{1}{n} \sum_{i=1}^n f(Z_i)$

résultat

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 8S_{\mathcal{F}}(n) e^{-n\epsilon^2/8}$$

si $S_{\mathcal{F}}(n)$ croît polynomialement avec n , $P_n f$ converge vers Pf uniformément sur \mathcal{F}

Comportement de $\mathcal{S}_{\mathcal{F}}(n)$

Lemme de Sauer 1972 (et indépendamment de Vapnik et Chervonenkis, 1971)

résultat

Si $VCdim(\mathcal{F}) < \infty$ alors pour tout n

$$\mathcal{S}_{\mathcal{F}}(n) \leq \sum_{i=0}^{VCdim(\mathcal{F})} \binom{n}{i}$$

majorations

$$\mathcal{S}_{\mathcal{F}}(n) \leq n^{VCdim(\mathcal{F})} + 1$$

$$\mathcal{S}_{\mathcal{F}}(n) \leq \left(\frac{en}{VCdim(\mathcal{F})} \right)^{VCdim(\mathcal{F})} \quad \text{pour } n \geq VCdim(\mathcal{F})$$

Application à la discrimination

- ▶ problème le plus simple de l'apprentissage :
 - ▶ discrimination à deux classes ($\mathcal{Y} = \{-1, 1\}$ et $\mathcal{X} = \mathbb{R}^d$)
 - ▶ fonction de coût : $c(g(\mathbf{x}), \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$
 - ▶ ensemble de modèles \mathcal{G} (*classifieurs*)
- ▶ *loss class*

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(\mathbf{x}, \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}} \right\}$$

- ▶ pour f associée à g

$$Pf = L(g)$$

$$P_n f = L_n(g)$$

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8\mathcal{S}_{\mathcal{F}}(n) e^{-n\epsilon^2/8}$$

Attention, \mathcal{G} est **fixe** !

VC-dimension

- ▶ comme g est à valeurs dans $\{-1, 1\}$, on peut considérer $\mathcal{S}_{\mathcal{G}}(n)$ et $VCdim(\mathcal{G})$
- ▶ $VCdim(\mathcal{G}) = VCdim(\mathcal{F})$
 - ▶ si $\mathbf{z}_1, \dots, \mathbf{z}_n$ est pulvérisé par \mathcal{F} , alors $\mathbf{x}_1, \dots, \mathbf{x}_n$ est pulvérisé par \mathcal{G} :
 - ▶ à $v \in \{-1, 1\}^n$ on associe $u_i = (v_i/y_i + 1)/2 \in \{0, 1\}$
 - ▶ $\exists f \in \mathcal{F}, f(\mathbf{z}_i) = u_i$
 - ▶ $f(\mathbf{z}_i) = u_i \Leftrightarrow g(\mathbf{x}_i) = (2u_i - 1)y_i$ et donc $g(\mathbf{x}_i) = v_i$
 - ▶ donc $v \in \mathcal{G}_{\mathbf{x}_1, \dots, \mathbf{x}_n}$
 - ▶ de même si $\mathbf{x}_1, \dots, \mathbf{x}_n$ est pulvérisé par \mathcal{G} , il existe $\mathbf{z}_1, \dots, \mathbf{z}_n$ pulvérisé par \mathcal{F}
 - ▶ donc $\mathcal{S}_{\mathcal{F}}(n) = 2^n \Leftrightarrow \mathcal{S}_{\mathcal{G}}(n) = 2^n$ et $\mathcal{S}_{\mathcal{F}}(n) < 2^n \Leftrightarrow \mathcal{S}_{\mathcal{G}}(n) < 2^n$
 - ▶ et donc $VCdim(\mathcal{G}) = VCdim(\mathcal{F})$
- ▶ en fait, on a même $\mathcal{S}_{\mathcal{F}}(n) = \mathcal{S}_{\mathcal{G}}(n)$

Bornes sur le risque

- ▶ avec une probabilité au moins $1 - \delta$

$$\begin{aligned} |L(g) - L_n(g)| &\leq 2\sqrt{2\frac{\log \mathcal{S}_{\mathcal{F}}(n) + \log \frac{8}{\delta}}{n}} \\ &\leq 2\sqrt{2\frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}} \end{aligned}$$

quand $3 \leq VCdim(\mathcal{G}) \leq n < \infty$

- ▶ on a aussi

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 4\sqrt{2\frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}}$$

ce qui justifie la minimisation du risque empirique

Compléments

- ▶ on peut déduire qu'il existe une **constante universelle** c telle que

$$\mathbb{E} \{L(g_n^*)\} - \inf_{g \in \mathcal{G}} L(g) \leq c \sqrt{\frac{VCdim(\mathcal{G}) \log n}{n}}$$

- ▶ les meilleures bornes connues conduisent à

$$\mathbb{E} \{L(g_n^*)\} - \inf_{g \in \mathcal{G}} L(g) \leq c' \sqrt{\frac{VCdim(\mathcal{G})}{n}}$$

à partir de majorations de la forme

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq \frac{\gamma}{\epsilon \sqrt{n}} \left(\frac{\gamma n \epsilon^2}{VCdim(\mathcal{F})} \right)^{VCdim(\mathcal{F})} e^{-2n\epsilon^2}$$

Preuve (résultat simplifié)

Lemme fondamental (symétrisation)

hypothèses

- ▶ n v.a. i.i.d. Z_1, \dots, Z_n à valeurs dans \mathbb{R}^d
- ▶ une copie (le *ghost sample*) indépendante Z'_1, \dots, Z'_n
- ▶ $P'_n f = \frac{1}{n} \sum_{i=1}^n f(Z'_i)$
- ▶ $n\epsilon^2 \geq 2$

résultat

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \right\} \leq 2 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| \geq \frac{\epsilon}{2} \right\}$$

Preuve du lemme

- ▶ soit f^* qui dépasse la borne ϵ ($|Pf^* - P_n f^*| > \epsilon$)

$$\begin{aligned}\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{I}_{\{|Pf^* - P'_n f^*| < \epsilon/2\}} &= \mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon \wedge |Pf^* - P'_n f^*| < \epsilon/2\}} \\ &\leq \mathbb{I}_{\{|P'_n f^* - P_n f^*| > \epsilon/2\}}\end{aligned}$$

Preuve du lemme

- ▶ soit f^* qui dépasse la borne ϵ ($|Pf^* - P_n f^*| > \epsilon$)

$$\begin{aligned}\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{I}_{\{|Pf^* - P'_n f^*| < \epsilon/2\}} &= \mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon \wedge |Pf^* - P'_n f^*| < \epsilon/2\}} \\ &\leq \mathbb{I}_{\{|P'_n f^* - P_n f^*| > \epsilon/2\}}\end{aligned}$$

- ▶ espérance par rapport au *ghost sample*

$$\begin{aligned}\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{P} \{ |Pf^* - P'_n f^*| < \epsilon/2 \mid Z_1, \dots, Z_n \} \\ \leq \mathbb{P} \{ |P'_n f^* - P_n f^*| > \epsilon/2 \mid Z_1, \dots, Z_n \}\end{aligned}$$

Preuve du lemme

- ▶ soit f^* qui dépasse la borne ϵ ($|Pf^* - P_n f^*| > \epsilon$)

$$\begin{aligned}\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{I}_{\{|Pf^* - P'_n f^*| < \epsilon/2\}} &= \mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon \wedge |Pf^* - P'_n f^*| < \epsilon/2\}} \\ &\leq \mathbb{I}_{\{|P'_n f^* - P_n f^*| > \epsilon/2\}}\end{aligned}$$

- ▶ espérance par rapport au *ghost sample*

$$\begin{aligned}\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{P} \{ |Pf^* - P'_n f^*| < \epsilon/2 \mid Z_1, \dots, Z_n \} \\ \leq \mathbb{P} \{ |P'_n f^* - P_n f^*| > \epsilon/2 \mid Z_1, \dots, Z_n \}\end{aligned}$$

- ▶ Bienaymé-Tchebychev

$$\begin{aligned}\mathbb{P}' \{ |Pf^* - P'_n f^*| \geq \epsilon/2 \} &\leq \frac{4\text{Var}f^*(Z_1)}{n\epsilon^2} \leq \frac{1}{n\epsilon^2} \\ \mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \left(1 - \frac{1}{n\epsilon^2} \right) &\leq \mathbb{P} \{ |P'_n f^* - P_n f^*| > \epsilon/2 \mid Z_1, \dots, Z_n \}\end{aligned}$$

Preuve du lemme (suite)

- ▶ espérance par rapport à l'échantillon

$$\mathbb{P} \{ |Pf^* - P_n f^*| > \epsilon \} \left(1 - \frac{1}{n\epsilon^2} \right) \leq \mathbb{P} \{ |P'_n f^* - P_n f^*| > \epsilon/2 \}$$

- ▶ et comme $1 - \frac{1}{n\epsilon^2} \leq \frac{1}{2}$, on obtient la conclusion en passant au supremum
- ▶ l'intérêt est qu'il n'y a qu'un ensemble fini de valeurs possibles pour $P'_n f^* - P_n f^*$ (au maximum 2^{2n})
- ▶ il s'agit des valeurs obtenues en balayant les $\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n, \mathbf{z}'_1, \dots, \mathbf{z}'_n}$

Preuve

- ▶ par Hoeffding appliquée aux $U_i = \frac{1}{n}(f(Z_i) - f(Z'_i))$ pour tout f fixée

$$\mathbb{P} \{ |P'_n f - P_n f| > \epsilon \} \leq 2e^{-n\epsilon^2/2}$$

Preuve

- ▶ par Hoeffding appliquée aux $U_i = \frac{1}{n}(f(Z_i) - f(Z'_i))$ pour tout f fixée

$$\mathbb{P} \left\{ |P'_n f - P_n f| > \epsilon \right\} \leq 2e^{-n\epsilon^2/2}$$

- ▶ finalement

$$\begin{aligned} \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} &\leq 2\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| \geq \frac{\epsilon}{2} \right\} \\ &= 2\mathbb{P} \left\{ \sup_{f \in \mathcal{F}_{z_1, \dots, z_n, z'_1, \dots, z'_n}} |P'_n f - P_n f| \geq \frac{\epsilon}{2} \right\} \\ &\leq 2S_{\mathcal{F}}(2n) \mathbb{P} \left\{ |P'_n f - P_n f| \geq \frac{\epsilon}{2} \right\} \\ &\leq 4S_{\mathcal{F}}(2n) e^{-n\epsilon^2/8} \end{aligned}$$

Plan

Concentration

Inégalité de Hoeffding

Bornes uniformes

Dimension de Vapnik-Chervonenkis

Définition

Application à la discrimination

Preuve

Nombres de couverture

Définition et résultat

Calcul des nombres de couverture

Résumé

Et la régression ?

- ▶ mesure de capacité :
 - ▶ étudier les $(f(\mathbf{z}_1), \dots, f(\mathbf{z}_n))$ quand $f \in \mathcal{F}$
 - ▶ mais en régression $f(\mathbf{x}) \in [0, B]$ (et non $f(\mathbf{x}) \in \{0, 1\}$)
 - ▶ donc $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = \infty$, en général
 - ▶ mais si $f(\mathbf{x}) \simeq h(\mathbf{x})$, pas de réel apport de h
- ▶ principe : considérer les fonctions suffisamment différentes sur les données

Et la régression ?

- ▶ mesure de capacité :
 - ▶ étudier les $(f(\mathbf{z}_1), \dots, f(\mathbf{z}_n))$ quand $f \in \mathcal{F}$
 - ▶ mais en régression $f(\mathbf{x}) \in [0, B]$ (et non $f(\mathbf{x}) \in \{0, 1\}$)
 - ▶ donc $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = \infty$, en général
 - ▶ mais si $f(\mathbf{x}) \simeq h(\mathbf{x})$, pas de réel apport de h
- ▶ principe : considérer les fonctions suffisamment différentes sur les données
- ▶ **Nombre de couverture**
 - ▶ dans \mathbb{R}^d , on pose $d(u, v) = \frac{1}{d} \sum_{i=1}^d |u_i - v_i|$
 - ▶ $A \subset \mathbb{R}^d$: une ϵ -couverture de A est un ensemble fini z_1, \dots, z_q tel que $A \subset \bigcup_{i=1}^q B(z_i, \epsilon)$ avec $B(u, \epsilon) = \{v \in \mathbb{R}^d \mid d(u, v) \leq \epsilon\}$
 - ▶ Nombre de couverture $N(\epsilon, A)$: cardinal de la plus petite ϵ -couverture de A

Convergence uniforme

Pollard, 1984

hypothèses

- ▶ Z_1, \dots, Z_n n v.a. indépendantes
- ▶ \mathcal{F} ensemble de fonctions à valeurs dans $[0, B]$
- ▶ $Pf = \mathbb{E} \{f(Z_1)\}$ et $P_n f = \frac{1}{n} \sum_{i=1}^n f(Z_i)$

résultat

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 8 \mathbb{E} \left\{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \right\} e^{-n\epsilon^2/(128B^2)}$$

avec

$$\mathcal{F}_{Z_1, \dots, Z_n} = \{u \in [0, B]^n \mid \exists f \in \mathcal{F}, u = (f(Z_1), \dots, f(Z_n))\}$$

Preuve

Toujours par symétrisation

hypothèses

- ▶ n v.a. i.i.d. Z_1, \dots, Z_n à valeurs dans \mathbb{R}^d
- ▶ \mathcal{F} ensemble de fonctions à valeurs dans $[0, B]$
- ▶ n v.a. i.i.d. de *Rademacher* $\sigma_1, \dots, \sigma_n$ à valeurs dans $\{-1, 1\}$, avec $\mathbb{P}\{\sigma_i = 1\} = 1/2$
- ▶ $n\epsilon^2 \geq 2B^2$

résultat

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 4 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \right\}$$

Preuve du lemme

- ▶ première étape symétrisation classique qui donne

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 2\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| > \epsilon/2 \right\}$$

- ▶ la condition $n\epsilon^2 \geq 2B^2$ correspond à $\text{Var}(f(Z_i)) \leq B^2/4$

Preuve du lemme

- ▶ première étape symétrisation classique qui donne

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 2\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| > \epsilon/2 \right\}$$

- ▶ la condition $n\epsilon^2 \geq 2B^2$ correspond à $\text{Var}(f(Z_i)) \leq B^2/4$
- ▶ deuxième étape : v.a. de *Rademacher*.

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| > \frac{\epsilon}{2} \right\} = \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - f(Z'_i)) \right| > \frac{\epsilon}{2} \right\}$$

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - f(Z'_i)) \right| \leq \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| + \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z'_i) \right|$$

Preuve du lemme (suite)

- ▶ et donc, par *union bound*

$$\begin{aligned} & \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - f(Z'_i)) \right| > \frac{\epsilon}{2} \right\} \\ & \leq \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \right\} + \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z'_i) \right| > \frac{\epsilon}{4} \right\} \end{aligned}$$

- ▶ d'où le résultat

Preuve

- ▶ g_1, \dots, g_M une $\epsilon/8$ couverture minimale de $\mathcal{F}_{Z_1, \dots, Z_n}$
($M = N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})$)

Preuve

- ▶ g_1, \dots, g_M une $\epsilon/8$ couverture minimale de $\mathcal{F}_{Z_1, \dots, Z_n}$
($M = N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})$)
- ▶ pour $f \in \mathcal{F}$, on a $g^* \in \{g_1, \dots, g_M\}$ tel que

$$\frac{1}{n} \sum_{i=1}^n |f(Z_i) - g^*(Z_i)| \leq \frac{\epsilon}{8}$$

Preuve

- ▶ g_1, \dots, g_M une $\epsilon/8$ couverture minimale de $\mathcal{F}_{Z_1, \dots, Z_n}$
($M = N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})$)
- ▶ pour $f \in \mathcal{F}$, on a $g^* \in \{g_1, \dots, g_M\}$ tel que

$$\frac{1}{n} \sum_{i=1}^n |f(Z_i) - g^*(Z_i)| \leq \frac{\epsilon}{8}$$

- ▶ soit

$$\begin{aligned} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| &\leq \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g^*(Z_i) \right| + \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - g^*(Z_i)) \right| \\ &\leq \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g^*(Z_i) \right| + \frac{\epsilon}{8} \end{aligned}$$

Preuve (suite)

► donc

$$\begin{aligned} & \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \\ & \leq \mathbb{P} \left\{ \max_{1 \leq j \leq M} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\} \end{aligned}$$

Preuve (suite)

- ▶ donc

$$\begin{aligned} \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \\ \leq \mathbb{P} \left\{ \max_{1 \leq j \leq M} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\} \end{aligned}$$

- ▶ on applique Hoeffding aux $U_i = \frac{1}{n} \sigma_i g_j(Z_i)$ (pour Z_1, \dots, Z_n fixé) d'espérances nulles :

$$\mathbb{P} \left\{ \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\} \leq 2e^{-n\epsilon^2/(128B^2)}$$

Preuve (suite)

► donc

$$\begin{aligned} \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \\ \leq \mathbb{P} \left\{ \max_{1 \leq j \leq M} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\} \end{aligned}$$

► on applique Hoeffding aux $U_i = \frac{1}{n} \sigma_i g_j(Z_i)$ (pour Z_1, \dots, Z_n fixé) d'espérances nulles :

$$\mathbb{P} \left\{ \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\} \leq 2e^{-n\epsilon^2/(128B^2)}$$

► soit

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \leq N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) 2e^{-n\epsilon^2/(128B^2)}$$

Application en régression

- ▶ problème assez général :

- ▶ $\mathcal{Y} = \mathbb{R}^p$
- ▶ fonction de coût c bornée par B : par exemple erreur quadratique et fonctions bornées (ainsi que Y)
- ▶ ensemble de modèles \mathcal{G}

- ▶ *loss class*

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \rightarrow [0, B] \mid \exists g \in \mathcal{G}, f(\mathbf{x}, \mathbf{y}) = c(g(\mathbf{x}), \mathbf{y}) \right\}$$

- ▶ pour f associée à g

$$Pf = L(g)$$

$$P_n f = L_n(g)$$

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8 \mathbb{E} \{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \} e^{-n\epsilon^2/(128B^2)}$$

Remarque

- ▶ résultat *a priori* limité :
 - ▶ $\mathbb{E} \{N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})\}$ dépend de la distribution de $Z = (X, Y)$
 - ▶ est difficile à calculer
- ▶ on peut montrer en discrimination

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8 \mathbb{E} \{ |\mathcal{F}_{Z_1, \dots, Z_n}| \} e^{-n\epsilon^2/8}$$

- ▶ idée de base : majorer $\mathbb{E} \{N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})\}$ de façon géométrique et combinatoire (comme pour la VC dimension)

Pseudo dimension

- ▶ à $\mathcal{F} = \{f : \mathbb{R}^d \rightarrow [0, B]\}$ on associe

$$\mathcal{F}^+ = \{f^+ : \mathbb{R}^d \times [0, B] \rightarrow \{0, 1\} \mid \exists f \in \mathcal{F}, f^+(x, t) = \mathbb{I}_{\{t \leq f(x)\}}\}$$

- ▶ on a (Pollard, 1984)

$$N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq \left(\frac{4eB}{\epsilon} \log \frac{2eB}{\epsilon} \right)^{VCdim(\mathcal{F}^+)}$$

- ▶ lien avec le *packing number* : $M(\epsilon, \mathcal{F}, \mu)$ cardinal de la plus grande collection de fonctions f de \mathcal{F} telles que

$$\int_{\mathbb{R}^d} |f_i(x) - f_j(x)| \mu(dx) \leq \epsilon$$

$$M(2\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}, 1/n) \leq N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq M(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}, 1/n)$$

Pulvérisation avec marge (*fat shattering*)

$VCdim(\mathcal{F}^+) < \infty$ n'est pas nécessaire pour obtenir un bon comportement de $\mathbb{E} \{N(\epsilon/8, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n})\}$

pulvérisation avec marge γ

$\mathbf{z}_1, \dots, \mathbf{z}_n$ est γ -pulvérisé par \mathcal{F} si pour tout $u \in \{-1, 1\}^n$, il existe $t \in [0, B]^n$ et $f \in \mathcal{F}$ tels que

$$(f(\mathbf{z}_i) - t_i)u_i \geq \gamma$$

la **dimension de pulvérisation avec marge γ** de \mathcal{F} , $\text{fat}_\gamma(\mathcal{F})$, est le cardinal du plus grand ensemble γ -pulvérisé par \mathcal{F} si $d = \text{fat}_\gamma(\mathcal{F})$ alors (Alon et al., 1993)

$$N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) < 2 \left(\frac{4nB^2}{\epsilon^2} \right)^{d \log_2(4eBn/(d\epsilon))}$$

En pratique

- ▶ on se contente généralement de passer par la pseudo dimension
- ▶ quelques propriétés :
 - ▶ si $\mathcal{H} = \{h + f \mid f \in \mathcal{F}\}$ pour h fixée, $VCdim(\mathcal{H}^+) = VCdim(\mathcal{F}^+)$
 - ▶ si h est une fonction croissante de $[0, B]$ dans \mathbb{R} et $\mathcal{H} = \{h \circ f \mid f \in \mathcal{F}\}$, $VCdim(\mathcal{H}^+) \leq VCdim(\mathcal{F}^+)$
 - ▶ pour k classes de fonctions $\mathcal{F}_1, \dots, \mathcal{F}_k$, on considère la classe $\mathcal{F} = \{f_1 + \dots + f_k \mid f_i \in \mathcal{F}_i\}$

$$N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq \prod_{j=1}^k N(\epsilon/k, \mathcal{F}_j, \mathbf{z}_1, \dots, \mathbf{z}_n)$$

- ▶ pour \mathcal{F}_i ($i = 1, 2$) fonctions à valeurs dans $[-B_i, B_i]$, on considère la classe $\mathcal{H} = \{f_1 f_2 \mid f_i \in \mathcal{F}_i\}$

$$N(\epsilon, \mathcal{H}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq N(\epsilon/(2B_2), \mathcal{F}_1, \mathbf{z}_1, \dots, \mathbf{z}_n) N(\epsilon/(2B_1), \mathcal{F}_2, \mathbf{z}_1, \dots, \mathbf{z}_n)$$

Exemple

- ▶ cadre général (Lugosi et Zegler, 1995) :

- ▶ $|Y| \leq L$

- ▶ $c(u, v) = |u - v|^p$

- ▶ $\mathcal{G} = \left\{ \sum_{j=1}^k w_j \phi_j; \sum_{j=1}^k |w_j| \leq \beta \right\}, \beta \geq L, |\phi_j| \leq 1$

- ▶ $\mathcal{F} = \left\{ f(x, y) = \left| \sum_{j=1}^k w_j \phi_j(x) - y \right|^p; \sum_{j=1}^k |w_j| \leq \beta \right\}$

- ▶ $f(x, y) \leq 2^p \max(\beta^p, L^p) \leq 2^p \beta^p$

- ▶ comme $||a|^p - |b|^p| \leq p|a - b| \max(a, b)^{p-1}$,

$$\int |f_1(x, y) - f_2(x, y)| \nu(dx, dy) \leq p(2\beta)^{p-1} \int |g_1(x) - g_2(x)| \mu(dx)$$

donc avec $\nu = 1/n$, $N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq N\left(\frac{\epsilon}{p(2\beta)^{p-1}}, \mathcal{G}_{\mathbf{z}_1, \dots, \mathbf{z}_n}\right)$

Exemple (suite)

- ▶ comme \mathcal{G} est une partie d'un e.v. de dimension k ,
 $VCdim(\mathcal{G}^+) \leq k$
- ▶ donc

$$\begin{aligned} N\left(\frac{\epsilon}{p(2\beta)^{p-1}}, \mathcal{G}_{\mathbf{z}_1, \dots, \mathbf{z}_n}\right) \\ \leq 2 \left(\frac{e2^{p+1}\beta^p}{\epsilon/(p(2\beta)^{p-1})} \log \frac{e2^{p+1}\beta^p}{\epsilon/(p(2\beta)^{p-1})} \right)^k \\ \leq 2 \left(\frac{ep2^{2p}\beta^{2p-1}}{\epsilon} \right)^k \end{aligned}$$

Résumé

- ▶ si le modèle est choisi dans \mathcal{G}

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq C(n, \mathcal{G}, \epsilon) e^{-cn\epsilon^2}$$

- ▶ c est liée uniquement à $\sup_{g \in \mathcal{G}} \|g\|_\infty$
- ▶ $C(n, \mathcal{G}, \epsilon)$ mesure la complexité de la classe \mathcal{G} :
 - ▶ nombre de couverture ou coefficient de pulvérisation
 - ▶ dimension de Vapnik-Chervonenkis
 - ▶ pseudo-dimension et dimension avec marge
 - ▶ on peut obtenir des majorations **universelles**

Conditions nécessaires

- ▶ discrimination à deux classes :

- ▶ $N(\mathcal{F}, Z_1, \dots, Z_n) = |\mathcal{F}_{Z_1, \dots, Z_n}|$
- ▶ $H_{\mathcal{F}}(n) = \log \mathbb{E} \{N(\mathcal{F}, Z_1, \dots, Z_n)\}$: entropie de VC
- ▶ C.N.S. de convergence uniforme du risque

$$\frac{H_{\mathcal{F}}(n)}{n} \rightarrow 0$$

- ▶ C.N.S. de convergence **universelle** uniforme du risque :
 $VCdim(\mathcal{F}) < \infty$

- ▶ régression :

- ▶ résultats similaires
- ▶ la dimension avec marge γ doit être finie pour tout γ
- ▶ la cible doit être bornée

Limitations

- ▶ \mathcal{G} est fixé et de puissance limitée
 - ▶ en général $\inf_{g \in \mathcal{G}} L(g) > L^*$
 - ▶ incompatible avec certaines techniques :
 - ▶ \mathcal{G} dépend de X_1, \dots, X_n
 - ▶ si on considère la classe globale (union de toutes les classes), la puissance est trop grande
- ▶ en régression, le risque doit être borné :
 - ▶ hypothèse assez forte
 - ▶ facile à vérifier dans \mathcal{G}
 - ▶ mais délicat pour les données
- ▶ solution :
 - ▶ adapter la puissance aux données
 - ▶ utiliser des bornes dépendant des données

Quelques résultats négatifs en discrimination

- ▶ $VCdim(\mathcal{G}) = \infty$ est gênant :
 - ▶ algorithme d'apprentissage fixé avec choix dans \mathcal{G} de dimension de VC infinie
 - ▶ pour tout $\epsilon > 0$ et tout n , il existe (X, Y) telle que $L_{\mathcal{G}}^* = 0$ et

$$\mathbb{E} \{L(g_n)\} \geq \frac{1}{2e} - \epsilon$$

- ▶ pulvériser un ensemble infini est plus gênant :
 - ▶ algorithme d'apprentissage fixé avec choix dans \mathcal{G}
 - ▶ il existe un ensemble A infini tel que pour tout $B \subset A$, il existe $g \in \mathcal{G}$ tel que $g(x) = 1$ sur B et $g(x) = 0$ sur $A \setminus B$
 - ▶ (a_n) une suite décroissante de limite nulle $a_1 \leq 1/16$
 - ▶ il existe (X, Y) telle que $L_{\mathcal{G}}^* = 0$ et pour tout n

$$\mathbb{E} \{L(g_n)\} \geq a_n$$

Quelques résultats négatifs en discrimination

- ▶ $VCdim(\mathcal{G}) < \infty$ limite la puissance des modèles :
 - ▶ $VCdim(\mathcal{G}) < \infty$
 - ▶ pour tout $\epsilon > 0$, il existe (X, Y) telle que

$$\inf_{g \in \mathcal{G}} L(g) - L^* > \frac{1}{2} - \epsilon$$

- ▶ même en augmentant la puissance avec n :
 - ▶ suite de classes avec $VCdim(\mathcal{G}^{(j)}) < \infty$ pour tout j
 - ▶ pour toute (a_n) suite positive de limite nulle, il existe (X, Y) telle qu'à partir d'un certain k

$$\inf_{g \in \mathcal{G}^{(k)}} L(g) - L^* > a_k$$

- ▶ l'ensemble de tous les classifieurs (les fonctions mesurables de \mathcal{X} dans $\{-1, 1\}$) ne peut pas se représenter comme l'union dénombrables de classes de dimensions de VC finies

Troisième partie III

Contrôle de complexité

Plan

Contrôle implicite

- Discrimination

- Régression

Contrôle explicite

- Minimisation du risque structurel

- Validation

- Régularisation

Complexité croissante

pour la discrimination

hypothèses

- ▶ classes $(\mathcal{G}^{(j)})_j$ de plus en plus puissantes avec $VCdim(\mathcal{G}^{(j)}) < \infty$
- ▶ asymptotiquement optimal : $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$
- ▶ $k_n \rightarrow \infty$ et $\frac{VCdim(\mathcal{G}^{(j)}) \log n}{n} \rightarrow 0$

résultat

le classifieur $g_n^* = \arg \min_{g \in \mathcal{G}^{(k_n)}} L_n(g)$ est universellement fortement consistant :

$$L(g_n^*) \xrightarrow{p.s.} L^*$$

Preuve élémentaire

▶ $L(g_n^*) - L^* = \left[L(g_n^*) - \inf_{g \in \mathcal{G}^{(k_n)}} L(g) \right] + \left[\inf_{g \in \mathcal{G}^{(k_n)}} L(g) - L^* \right]$

- ▶ contrôle du premier terme

$$L(g_n^*) - \inf_{g \in \mathcal{G}^{(k_n)}} L(g) \leq 2 \sup_{g \in \mathcal{G}^{(k_n)}} |L_n(g) - L(g)|$$

$$\begin{aligned} \mathbb{P} \left\{ L(g_n^*) - \inf_{g \in \mathcal{G}^{(k_n)}} L(g) \geq \epsilon \right\} &\leq \mathbb{P} \left\{ \sup_{g \in \mathcal{G}^{(k_n)}} |L(g) - L_n(g)| > \epsilon/2 \right\} \\ &\leq 8 \mathcal{S}_{\mathcal{F}^{(k_n)}}(n) e^{-n\epsilon^2/32} \\ &\leq 8(n^{\text{VCdim}(\mathcal{G}^{(k_n)})} + 1) e^{-n\epsilon^2/32} \end{aligned}$$

- ▶ Borel Cantelli donne la convergence presque sûre
- ▶ le deuxième terme est contrôlé par les hypothèses

Régression

- ▶ exactement le même principe :
 - ▶ asymptotiquement optimal : $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$
 - ▶ puissance croissante mais contrôlée, par exemple par la pseudo-dimension $VCdim(\mathcal{G}^{(j)+})$
 - ▶ fonctions bornées $\sup_{g \in \mathcal{G}^{(j)}} \|g\|_{\infty} < \infty$, mais borne évolutive
 - ▶ fonction cible bornée $|Y| < \infty$
- ▶ suppression de la contrainte de borne :
 - ▶ Lugosi et Zeger, 1995
 - ▶ pour $c(u, v) = |u - v|^p$ et $\mathbb{E}\{|Y|^p\} < \infty$
 - ▶ idée de base : cible tronquée $Y_L = \text{signe}(Y) \min(|Y|, L)$

Réseau de neurones

Perceptrons multi-couches (Lugosi et Zeger, 1995)

- ▶ fonction sigmoïde σ de \mathbb{R} dans $[0, 1]$, croissante et telle que $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ par exemple $\sigma(x) = 1/(1 + e^{-x})$
- ▶ classe correspondante

$$\mathcal{G}(k, \beta) = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0; \sum_{i=1}^k |c_i| \leq \beta \right\}$$

- ▶ si $k_n \rightarrow \infty$ et $\beta_n \rightarrow \infty$, $\bigcup_n \mathcal{G}(k_n, \beta_n)$ est dense dans $L^p(\mu)$ (pour toute probabilité μ , Hornik, Stinchcombe, White, 1989) et donc $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}(k_n, \beta_n)} L(g) = L^*$
($c(u, v) = |u - v|^p$)
- ▶ contrôle de complexité : $\frac{k_n \beta_n^{2p} \log(k_n \beta_n)}{n} \rightarrow 0$

Minimisation du risque structurel

- ▶ le contrôle de complexité ne dépend pas des données
- ▶ équilibre
$$L(g_n^*) - L^* = [L(g_n^*) - \inf_{g \in \mathcal{G}} L(g)] + [\inf_{g \in \mathcal{G}} L(g) - L^*]$$
- ▶ idée : ajouter à $L_n(g)$ une mesure de la complexité de \mathcal{G}
- ▶ **minimisation du risque structurel** (discrimination) :
 - ▶ $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$
 - ▶ $\sum_{j=1}^{\infty} e^{-VCdim(\mathcal{G}^{(j)})} < \infty$
 - ▶ $r(j, n) = \sqrt{\frac{8}{n} VCdim(\mathcal{G}^{(j)}) \log(en)}$
 - ▶ g_n^* minimise $\tilde{L}_n(g) = L_n(g) + r(j(g), n)$, où $j(g) = \inf\{k \mid g \in \mathcal{G}^{(k)}\}$
 - ▶ alors $L(g_n^*) \xrightarrow{p.s.} L^*$

Preuve

- ▶ $g_{n,j} = \arg \min_{g \in \mathcal{G}^{(j)}} L_n(g)$ (et donc $g_n^* = \arg \min_j \tilde{L}_n(g_{n,j})$)
- ▶ décomposition

$$L(g_n^*) - L^* = (L(g_n^*) - \inf_j \tilde{L}_n(g_{n,j})) + (\inf_j \tilde{L}_n(g_{n,j}) - L^*)$$

- ▶ le premier terme est en fait $L(g_n^*) - \tilde{L}_n(g_n^*)$ et

$$\begin{aligned} \mathbb{P} \left\{ L(g_n^*) - \tilde{L}_n(g_n^*) > \epsilon \right\} &\leq \mathbb{P} \left\{ \sup_j (L(g_{n,j}) - \tilde{L}_n(g_{n,j})) > \epsilon \right\} \\ &\leq \mathbb{P} \left\{ \sup_j (L(g_{n,j}) - L_n(g_{n,j}) - r(j, n)) > \epsilon \right\} \\ &\leq \sum_{j=1}^{\infty} \mathbb{P} \left\{ |L(g_{n,j}) - L_n(g_{n,j})| > \epsilon + r(j, n) \right\} \\ &\leq \sum_{j=1}^{\infty} 8n^{\text{VCdim}(\mathcal{G}^{(j)})} e^{-n(\epsilon+r(j,n))^2/8} \end{aligned}$$

Preuve

► or

$$\begin{aligned} \sum_{j=1}^{\infty} 8n^{\text{VCdim}(\mathcal{G}^{(j)})} e^{-n(\epsilon+r(j,n))^2/8} &\leq \sum_{j=1}^{\infty} 8n^{\text{VCdim}(\mathcal{G}^{(j)})} e^{-n\epsilon^2/8} e^{-r(j,n)^2/8} \\ &\leq 8e^{-n\epsilon^2/8} \sum_{j=1}^{\infty} e^{-\text{VCdim}(\mathcal{G}^{(j)})} \end{aligned}$$

- on obtient donc par Borel Cantelli la convergence presque sûre de $L(g_n^*)$ vers $\inf_j \tilde{L}_n(g_{n,j})$
- pour la seconde partie on fixe $\epsilon > 0$ et on trouve k tel que $\inf_{g \in \mathcal{G}^{(k)}} L(g) - L^* \leq \epsilon$:
- pour tout n assez grand $n r(k, n) \leq \frac{\epsilon}{2}$ (car $r(k, n)$ tend vers 0 avec n pour k fixé)

- ▶ on a alors

$$\begin{aligned} \mathbb{P} \left\{ \inf_j \tilde{L}_n(g_{n,j}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \epsilon \right\} &\leq \mathbb{P} \left\{ \tilde{L}_n(g_{n,k}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \epsilon \right\} \\ &\leq \mathbb{P} \left\{ L_n(g_{n,k}) + r(k, n) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \epsilon \right\} \\ &\leq \mathbb{P} \left\{ L_n(g_{n,k}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \frac{\epsilon}{2} \right\} \\ &\leq \mathbb{P} \left\{ \sup_{g \in \mathcal{G}^{(k)}} |L_n(g) - L(g)| > \frac{\epsilon}{4} \right\} \\ &\leq 8n^{\text{VCdim}(\mathcal{G}^{(k)})} e^{-n\epsilon^2/128} \end{aligned}$$

- ▶ donc $\mathbb{P} \left\{ \limsup_{n \rightarrow \infty} \inf_j \tilde{L}_n(g_{n,j}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) = 0 \right\} = 1$
- ▶ d'où le résultat

Choix de modèles

- ▶ méthode de l'ensemble de validation
- ▶ on coupe D_n en $D_m = (X_i, Y_i)_{1 \leq i \leq m}$ et $D_l = (X_i, Y_i)_{m+1 \leq i \leq n}$ ($l = n - m$)
- ▶ on construit à partir de D_m une classe \mathcal{G}_m
- ▶ on choisit $g_n = \arg \min_{g \in \mathcal{G}_m} \frac{1}{l} \sum_{i=m+1}^n c(g(X_i), Y_i)$
- ▶ les bornes dérivées pour le MRE s'appliquent à $L(g_n)$ en considérant la classe \mathcal{G}_m
- ▶ applications :
 - ▶ choix des paramètres d'une méthode (nombre de voisins)
 - ▶ MRE pour différentes classes sur D_m puis choix du meilleur modèle parmi ceux ainsi obtenus

Régularisation

- ▶ généralisation du principe de la minimisation du risque structurel
- ▶ idée fondamentale :
 - ▶ choisir g_n en minimisant sur une classe \mathcal{G}

$$A_n(g) + \lambda \mathbf{R}(g)$$

- ▶ $A_n(g)$ est une mesure de performance empirique (pas nécessairement le risque empirique)
 - ▶ $\mathbf{R}(g)$ est une mesure de la complexité du modèle
- ▶ exemples :
 - ▶ machine à vecteurs de support en discrimination :
 - ▶ $A_n(g) = \frac{1}{n} \sum_{i=1}^n \max(0, -\mathbf{y}_i g(\mathbf{x}_i))$ (*hinge loss*)
 - ▶ $\mathbf{R}(g) = \|g\|_{\mathcal{H}}^2$
 - ▶ perceptrons multi-couches en régression :
 - ▶ $A_n(g)$: risque empirique
 - ▶ $\mathbf{R}(g)$: somme des carrés de tous les paramètres numériques du réseau (les poids)

Régularisation

Plusieurs difficultés :

- ▶ si A_n n'est pas le risque empirique, il faut montrer que sa minimisation conduit bien à minimiser le risque
- ▶ il faut choisir λ :
 - ▶ soit en donnant des règles de comportement de λ avec n
 - ▶ soit en intégrant un choix automatique (de type minimisation du risque structurel)
- ▶ la classe est souvent de dimension de VC infinie, ce qui complique l'analyse

Quatrième partie IV

Au delà de la minimisation du risque structurel

Plan

Minimiser un autre coût

- Motivations

- Risque convexe

Retour sur la régularisation

- Régularisation dans les RKHS

Problème algorithmique

- ▶ en régression, optimiser L_n est « facile »
 - ▶ par exemple, $L_n(g) = \frac{1}{n} \sum_{i=1}^n \|g(\mathbf{x}_i) - \mathbf{y}_i\|^2$: moindres carrés
 - ▶ si \mathcal{G} est une classe paramétrique

$$\mathcal{G} = \{g(\mathbf{x}) = h(\mathbf{x}, w); w \in W\},$$

on utilise des techniques classiques (descente de gradient)

- ▶ en discrimination, la situation est plus délicate :
 - ▶ risque idéal $\mathbb{P}\{g(X) \neq Y\}$
 - ▶ donc le risque empirique est « discret » et ne peut pas être optimiser par descente de gradient
 - ▶ problème NP difficile dans certains cas particulier

Minimiser un autre coût

- ▶ solution classique : minimiser un autre coût
- ▶ par exemple, le coût quadratique : $a(u, v) = (u - v)^2$, ce qui revient à remplacer $L_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{g(X_i) \neq Y_i\}}$ par

$$A_n(g) = \frac{1}{n} \sum_{i=1}^n a(g(X_i), Y_i) = \frac{1}{n} \sum_{i=1}^n (g(X_i) - Y_i)^2$$

- ▶ résultats en régression \Rightarrow bornes sur $|A_n(g) - A(g)|$, où

$$A(g) = \mathbb{E} \{ a(g(X), Y) \}$$

puis convergence vers $A_G^* = \inf_{g \in \mathcal{G}} A(g)$

- ▶ mais on veut des informations sur $L(g)$!
- ▶ en général, on trouve g à valeurs dans \mathbb{R} qu'on transforme en classifieur par $h(\mathbf{x}) = \text{signe}(g(\mathbf{x}))$

Erreur quadratique

- ▶ cas simple

- ▶ $\eta(\mathbf{x}) = \mathbb{E} \{ Y \mid X = \mathbf{x} \}$

- ▶ algorithme fortement consistant :

$$\mathbb{E} \{ (g_n(X) - Y)^2 \mid D_n \} \xrightarrow{p.s.} \mathbb{E} \{ (\eta(X) - Y)^2 \} \text{ c.-à-d.,}$$

$$\mathbb{E} \{ (g_n(X) - \eta(X))^2 \mid D_n \} \xrightarrow{p.s.} 0$$

- ▶ or

$$\mathbb{P} \{ h_n(X) \neq Y \mid D_n \} - \mathbb{P} \{ h^*(X) \neq Y \} \leq \mathbb{E} \{ (g_n(X) - \eta(X))^2 \mid D_n \}$$

où h^* est le classifieur de Bayes

- ▶ donc $L(h_n) \xrightarrow{p.s.} L^*$

Cas plus général

- ▶ plus généralement, tout se passe bien si on peut garantir que $L(g) - L^*$ tend vers 0 quand $A(g) - A^*$ tend vers 0
- ▶ approche de Steinwart (2005) :
 - ▶ $a: \mathbb{R} \times \{-1, 1\} \rightarrow \mathbb{R}$
 - ▶ $C(t, \alpha) = \alpha a(t, 1) + (1 - \alpha)a(t, -1)$
 - ▶ $A(g) = \int C(\mathbb{P}\{Y = 1 \mid X = x\}, x) P_X(dx)$
 - ▶ a est **admissible** si :
 - ▶ a est continue
 - ▶ $\alpha < 1/2 \Rightarrow \arg \min_t C(t, \alpha) < 0$
 - ▶ $\alpha > 1/2 \Rightarrow \arg \min_t C(t, \alpha) > 0$
 - ▶ c'est-à-dire si un minimiseur de A est du même signe que la fonction de régression !
 - ▶ dans ce cas, pour tout ϵ , il existe δ tel que $A(g) - A^* \leq \delta$ implique $L(g) - L^* \leq \epsilon$

Exemple : le *hinge loss*

- ▶ *hinge loss*

$$a(x, y) = \max(1 - yx, 0)$$

- ▶ $a(g(\mathbf{x}), \mathbf{y}) \geq \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$

- ▶ $C(t, \alpha) = \alpha \max(1 - t, 0) + (1 - \alpha) \max(1 + t, 0)$:

- ▶ si $t \geq 1$, $C(t, \alpha) = (1 - \alpha)(1 + t) \geq 2(1 - \alpha)$

- ▶ si $t \leq -1$, $C(t, \alpha) = \alpha(1 - t) \geq 2\alpha$

- ▶ si $t \in [-1, 1]$,

- $C(t, \alpha) = \alpha(1 - t) + (1 - \alpha)(1 + t) = 1 + (1 - 2\alpha)t$

- ▶ si $\alpha < 1/2$, le minimum sur $[-1, 1]$ est atteint en -1 et vaut $2\alpha \leq 2(1 - \alpha)$

- ▶ symétriquement, pour $\alpha > 1/2$, le minimum est atteint en 1 et vaut $2(1 - \alpha) \leq 2\alpha$

- ▶ le *hinge loss* est donc admissible

Minimisation d'un risque convexe

- ▶ cas particulier très général $a(u, v) = \phi(uv)$ pour ϕ positive :
 - ▶ comme précédemment, $C(t, \alpha) = \alpha\phi(t) + (1 - \alpha)\phi(-t)$
 - ▶ $H(\alpha) = \inf_t C(t, \alpha)$ et $H^-(\alpha) = \inf_{t(2\alpha-1) \leq 0} C(t, \alpha)$
 - ▶ ϕ est calibrée si $H^-(\alpha) > H(\alpha)$ pour tout α
- ▶ pour tout ϕ il existe une fonction ψ telle que

$$\psi(L(g) - L^*) \leq A(g) - A^*$$

- ▶ ϕ calibrée est équivalent à

$$\lim_{i \rightarrow \infty} \psi(\alpha_i) = 0 \Leftrightarrow \lim_{i \rightarrow \infty} \alpha_i = 0$$

- ▶ si ϕ est convexe, ϕ calibrée est équivalent à ϕ dérivable en 0 et $\phi'(0) < 0$

Exemples

- ▶ $\phi(a) = \max(1 - a, 0)$ ($a(x, y) = \max(1 - yx, 0)$)
- ▶ $\phi(a) = e^{-a}$ ($a(x, y) = e^{-yx}$)
- ▶ $\phi(a) = |1 - a|^p$ ($a(x, y) = |1 - yx|^p$)
- ▶ remarque importante : si ϕ est calibrée, il existe γ tel que

$$\gamma\phi(a) \geq \mathbb{I}_{\{a \leq 0\}}$$

et donc

$$\gamma a(g(x), y) \geq \mathbb{I}_{\{g(x)y \leq 0\}} = \mathbb{I}_{\{\text{signe}(g(x)) \neq y\}}$$

Plan

Minimiser un autre coût

Motivations

Risque convexe

Retour sur la régularisation

Régularisation dans les RKHS

Machines à vecteurs de support

- ▶ on se donne un noyau K et on travaille dans le complété \mathcal{H} de

$$H = \left\{ g(\mathbf{x}) = \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \mathbf{x}); p \in \mathbb{N}, \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X} \right\}$$

- ▶ l'algorithme standard cherche g en résolvant

$$g_n = \arg \min_{g \in \mathcal{H}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \max(0, -\mathbf{y}_i(g(\mathbf{x}_i) + b)) + \lambda_n \|g\|_{\mathcal{H}}^2$$

- ▶ on peut généraliser en cherchant

$$g_n = \arg \min_{g \in \mathcal{H}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n a(g(\mathbf{x}_i) + b, \mathbf{y}_i) + \Omega(\lambda_n, \|g\|_{\mathcal{H}})$$

Consistance des MVS

Analyse de Steinwart 2005

- ▶ deux difficultés :
 - ▶ \mathcal{H} n'est généralement pas de dimension de VC finie
 - ▶ on optimise une grandeur complexe qui n'est ni le risque empirique, ni une somme de celui-ci et d'un terme simple
- ▶ stratégie d'analyse :
 - ▶ montrer quand λ tend vers 0 l'influence de la régularisation disparaît : asymptotiquement, on optimise A
 - ▶ utiliser l'analyse précédente sur A et A_n pour montrer que si on optimise A tout se passe comme si on optimisait asymptotiquement L
 - ▶ montrer enfin que la régularisation est suffisante pour contrôler $A_n - A$

Quelques hypothèses

- ▶ les X_i sont à valeurs dans \mathcal{X} un espace métrique **compact**
- ▶ le noyau K est tel que :
 - ▶ $\phi(\mathbf{x}) = K(\mathbf{x}, \cdot)$ (de \mathcal{X} dans \mathcal{H}) est continue
 - ▶ l'ensemble $\{g \in C(\mathcal{X}) \mid g(\cdot) = K(w, \phi(\cdot)); w \in \mathcal{H}\}$ est dense dans $C(\mathcal{X})$ (noyau **universel**)
 - ▶ par exemple $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$ pour \mathcal{X} un compact de \mathbb{R}^d
- ▶ Steinwart étudie le cas général pour a **admissible** et pour Ω vérifiant des hypothèses de régularité assez techniques
- ▶ ici MVS classiques :
 - ▶ a est le *hinge loss*
 - ▶ $\Omega(\lambda_n, \|g\|_{\mathcal{H}}) = \lambda_n \|g\|_{\mathcal{H}}^2$
 - ▶ pas de terme b (pour simplifier les hypothèses)

Régularisation

- ▶ on pose

$$A_\lambda(g) = \mathbb{E} \{ \max(0, -Y(g(X))) \} + \lambda \|g\|_{\mathcal{H}}^2$$

$$A_{\lambda,n}(g) = \frac{1}{n} \sum_{i=1}^n \max(0, -\mathbf{y}_i(g(\mathbf{x}_i) + b)) + \lambda \|g\|_{\mathcal{H}}^2$$

- ▶ on montre que tout minimiseur de A_λ sur \mathcal{H} vérifie $\|g\|_{\mathcal{H}}^2 \leq \frac{2}{\lambda}$
- ▶ de même, tout minimiseur de $A_{\lambda,n}$ vérifie $\|g\|_{\mathcal{H}}^2 \leq \frac{2}{\lambda}$
- ▶ de ce fait, la régularisation impose une borne sur le minimiseur qui va permettre de contrôler $|A(g) - A_n(g)|$
- ▶ on montre aussi que si $g_\lambda = \arg \min_{g \in \mathcal{H}} A_\lambda(g)$

$$\lim_{\lambda \rightarrow 0} A_\lambda(g_\lambda) = \inf_g A(g)$$

Concentration

- ▶ on majore $|A(g) - A_n(g)|$ en utilisant un nombre de couverture
- ▶ pour une classe \mathcal{F} de fonctions à valeurs dans $[0, B]$

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \right\} \leq 2N_\infty \left(\frac{\epsilon}{3}, \mathcal{F} \right) e^{-2n\epsilon^2/(9B^2)}$$

- ▶ le nombre de couverture est défini ici par rapport à la norme infinie
- ▶ il suffit de considérer une $\epsilon/3$ -couverture de \mathcal{F} , f_1, \dots, f_m :
 - ▶ $|Pf - P_n f| = |Pf - Pf_i + Pf_i - P_n f_i + P_n f_i - P_n f| \leq \frac{2}{3}\epsilon + |Pf_i - P_n f_i|$
 - ▶ $\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \right\} \leq \mathbb{P} \left\{ \sup_{1 \leq i \leq m} |Pf_i - P_n f_i| \geq \frac{\epsilon}{3} \right\}$
 - ▶ par Hoeffding $\mathbb{P} \left\{ |Pf_i - P_n f_i| \geq \frac{\epsilon}{3} \right\} \leq 2e^{-2n\epsilon^2/(9B^2)}$

Concentration

- ▶ on note $\|K\|_\infty = \sup_{\mathbf{x} \in \mathcal{X}} \sqrt{K(\mathbf{x}, \mathbf{x})}$
- ▶ $\|g\|_{\mathcal{H}} \leq \delta$ implique $\|g\|_\infty \leq \delta \|K\|_\infty$:
 - ▶ noyau « reproduisant » : $g(\mathbf{x}) = \langle g, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$
 - ▶ $|\langle g, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}| \leq \|g\|_{\mathcal{H}} \|K(\mathbf{x}, \cdot)\|_{\mathcal{H}}$
 - ▶ $\|K(\mathbf{x}, \cdot)\|_{\mathcal{H}} = \sqrt{K(\mathbf{x}, \mathbf{x})} \leq \|K\|_\infty$
- ▶ on considère la classe suivante

$$\mathcal{F}_\delta = \{a(g(\cdot), \cdot); \|g\|_{\mathcal{H}} \leq \delta\}$$

pour $a(u, v) = \max(0, 1 - uv)$

- ▶ pour tout $f \in \mathcal{F}_\delta$, $f(\mathbf{x}, \mathbf{y}) \in [0, \delta \|K\|_\infty]$ et donc

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}_\delta} |Pf - P_n f| \geq \epsilon \right\} \leq 2N_\infty \left(\frac{\epsilon}{3}, \mathcal{F}_\delta \right) e^{-2n\epsilon^2 / (9\delta^2 \|K\|_\infty^2)}$$

Concentration

- ▶ soit g_1, \dots, g_m une ϵ -couverture de $\mathcal{H}_\delta = \{g \in \mathcal{H} \mid \|g\|_{\mathcal{H}} \leq \delta\}$
- ▶ $\|g_i - g_j\|_\infty \leq \epsilon$ implique $\|a(g_i(\cdot), \cdot) - a(g_j(\cdot), \cdot)\|_\infty \leq \epsilon$
- ▶ donc les $(a(g_i(\cdot), \cdot))_{1 \leq i \leq m}$ forment une ϵ -couverture de \mathcal{F}_δ
- ▶ donc $N_\infty(\epsilon, \mathcal{F}_\delta) \leq N_\infty(\epsilon, \mathcal{H}_\delta)$
- ▶ et donc

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{H}_\delta} |A(g) - A_n(g)| \geq \epsilon \right\} \leq 2N_\infty \left(\frac{\epsilon}{3}, \mathcal{H}_\delta \right) e^{-2n\epsilon^2 / (9\delta^2 \|K\|_\infty^2)}$$

- ▶ donc si $g_{n,\lambda} = \arg \min_{g \in \mathcal{H}} A_{\lambda,n}(g)$

$$\mathbb{P} \{ |A(g_{n,\lambda}) - A_n(g_{n,\lambda})| \geq \epsilon \} \leq 2N_\infty \left(\frac{\epsilon}{3}, \mathcal{H}_{\sqrt{\frac{2}{\lambda}}} \right) e^{-n\epsilon^2 \lambda / (9\|K\|_\infty^2)}$$

Récapitulation

- ▶ $g_n = \arg \min_{g \in \mathcal{H}} A_{\lambda_n, n}(g)$ (λ_n tend vers 0)
- ▶ $g_n^* = \arg \min_{g \in \mathcal{H}} A_{\lambda_n}(g)$
- ▶ soit $\epsilon > 0$:
 - ▶ il existe $\delta < 0$ tel que $A(g) \leq A^* + \delta$ implique $L(g) \leq L^* + \epsilon$ (admissibilité)
 - ▶ pour $n \geq n_0$, $|A_{\lambda_n}(g_n^*) - A^*| \leq \delta/3$
 - ▶ si $|A(g_n) - A_n(g_n)| < \delta/3$ et $|A(g_n^*) - A_n(g_n^*)| < \delta/3$

$$\begin{aligned} A(g_n) &\leq A(g_n) + \lambda_n \|g_n\|_{\mathcal{H}}^2 \\ &\leq A_n(g_n) + \lambda_n \|g_n\|_{\mathcal{H}}^2 + \delta/3 \\ &\leq A_n(g_n^*) + \lambda_n \|g_n^*\|_{\mathcal{H}}^2 + \delta/3 \\ &\leq A(g_n^*) + \lambda_n \|g_n^*\|_{\mathcal{H}}^2 + 2\delta/3 \\ &\leq A^* + \delta \end{aligned}$$

- ▶ donc pour $n \geq n_0$, $\mathbb{P}\{L(g) \geq L^* + \epsilon\} \leq \mathbb{P}\{|A(g_n) - A_n(g_n)| \geq \delta/3\} + \mathbb{P}\{|A(g_n^*) - A_n(g_n^*)| \geq \delta/3\}$

Nombre de couverture

- ▶ reste donc à contrôler $N_\infty(\epsilon, \mathcal{H}_\delta)$
- ▶ résultats d'analyse sur l'approximation des opérateurs
- ▶ quand un noyau K est régulier, en général

$$\ln N_\infty(\epsilon, \mathcal{H}_1) \leq c\epsilon^{-\gamma}$$

pour des constantes $c > 0$ et $\gamma > 0$, et donc

$$\ln N_\infty(\epsilon, \mathcal{H}_\delta) \leq c \left(\frac{\delta}{\epsilon} \right)^\gamma$$

- ▶ on peut faire mieux pour certains noyaux spécifiques. Par exemple, le noyau gaussien sur \mathbb{R}^d donne

$$\ln N_\infty(\epsilon, \mathcal{H}_1) \leq c \left(\log \frac{1}{\epsilon} \right)^{d+1}$$

Consistance

- ▶ il suffit maintenant de contrôler $2N_\infty \left(\frac{\epsilon}{3}, \mathcal{H}_{\sqrt{\frac{2}{\lambda_n}}} \right) e^{-n\epsilon^2 \lambda_n / (9\|K\|_\infty^2)}$
- ▶ par exemple $\frac{\|K\|_\infty^2}{n\lambda_n} \ln N_\infty \left(\epsilon, \mathcal{H}_{\sqrt{\frac{2}{\lambda_n}}} \right) \rightarrow 0$ assure la consistance
- ▶ pour un noyau régulier, il suffit donc que

$$n\lambda_n^{1+\gamma/2} \rightarrow \infty$$

Cinquième partie V

Annexe

Plan

Références I



S. Boucheron, O. Bousquet, and G. Lugosi.

Theory of classification : a survey of some recent advances.

ESAIM ; Probability and Statistics, 9 :323–375, November 2005.

Résumé de résultats assez récents

<http://www.econ.upf.edu/~lugosi/esaimsurvey.pdf>



O. Bousquet, S. Boucheron, and G. Lugosi.

Advanced lectures in machine learning, volume 3176 of *LNAI*, chapter Introduction to statistical learning theory, pages 169–207.

Springer-Verlag, 2004.

Introduction très claire

http://www.econ.upf.edu/~lugosi/mlss_slit.pdf

Références II



L. Devroye, L. Györfi, and G. Lugosi.

A Probabilistic Theory of Pattern Recognition, volume 21 of *Applications of Mathematics*.

Springer, 1996.

Ouvrage de référence



S. Kulkarni, G. Lugosi, and S. Venkatesh.

Learning pattern classification – a survey.

IEEE Transactions on Information Theory,
44(6) :2178–2206, October 1998.

Etat de l'art en 1998

http://www.princeton.edu/~kulkarni/Papers/Journals/j1998_klv_transit.pdf

Références III



G. Lugosi and K. Zeger.

Nonparametric estimation via empirical risk minimization.
IEEE Transactions on Information Theory, 41(3) :677–687,
May 1995.

Régression et coût L^p

[http://www.code.ucsd.edu/~zeger/publications/journals/LuZe95-IT-Nonparametric/
LuZe95-IT-Nonparametric.pdf](http://www.code.ucsd.edu/~zeger/publications/journals/LuZe95-IT-Nonparametric/LuZe95-IT-Nonparametric.pdf)



G. Lugosi and K. Zeger.

Concept learning using complexity regularization.
IEEE Transactions on Information Theory, 42(1) :48–54,
January 1996.

Minimisation structurelle du risque

[http://www.code.ucsd.edu/~zeger/publications/journals/LuZe96-IT-Concept/
LuZe96-IT-Concept.pdf](http://www.code.ucsd.edu/~zeger/publications/journals/LuZe96-IT-Concept/LuZe96-IT-Concept.pdf)

Références IV



D. Pollard.

Convergence of Stochastic Processes.

Springer-Verlag, New York, 1984.

Ouvrage de référence



I. Steinwart.

Consistency of support vector machines and other regularized kernel machines.

IEEE Transactions on Information Theory,

51(1) :128—142, January 2005.

Consistance de différentes méthodes à noyau dont les SVM