Estimation non-paramétrique et Apprentissage statistique

Sébastien Loustau, Université d'Angers

25 Février 2010, Université de La Rochelle

De la statistique paramétrique...

La statistique paramétrique remonte à Fisher, 1920 : estimation d'un nombre fini de paramètres $\theta \subset \mathbb{R}^k$.

Limites de l'approche :

1. modèles simplistes ne fournissant qu'une approximation de la réalité,
2. résultats très souvent asymptotiques.

La réalité est souvent plus complexe, le nombre d'observations limité, et les inconnues des fonctions possédant certaines propriétés de régularité.

... à la statistique non-paramétrique

La statistique non-paramétrique s'intéresse à l'estimation, à partir d'un nombre fini d'observations, d'une fonction inconnue $f \in \Theta$, où Θ est un espace fonctionnel assez large.

Ces 30 dernières années, la théorie de l'estimation non-paramétrique s'est développée autour des thèmes suivants :

1. Méthodes de constructions d'estimateurs,
2. Propriétés statistiques de ces estimateurs,
3. Optimalité de ces estimateurs,
4. Estimation adaptative.

Plan de l'exposé

1. Statistique non-paramétrique
   - Modèles statistiques
   - Estimateurs, risque, régularisation
   - Vitesses de convergence
   - Adaptation, inégalités oracles
2. Apprentissage statistique
   - Modèles d'apprentissage
   - Algorithmes d'apprentissage
   - Vitesses de convergence
   - Adaptation

Modèle non-paramétrique : l'estimation d'une densité

On dispose d'observations X_i, $i = 1, \dots n$ i.i.d. de loi inconnue P_f de densité f telle que :

- $f \in \{f(x, \theta), \theta \in \Theta\}$ où $\Theta \subset \mathbb{R}^k$ et $f(x, \theta)$ connue $\rightarrow$ estimation paramétrique de θ.
- $f \in \mathcal{F}$, où $\mathcal{F}$ espace fonctionnel $\rightarrow$ estimation non-paramétrique de f.

Construction d'un estimateur : l'estimateur à noyau (Rosenblatt, 1956).
D'après Glivenko-Cantelli, on a (uniformément en x) :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x) \xrightarrow{p.s.} F(x) = \mathbb{P}(X \leq x).$$

Or, pour h assez petit,

$$f(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

et donc l'estimateur à noyau de Rosenblatt $\hat{f}_n$ est défini par :

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}(-h < x - X_i \leq h).$$

ou plus généralement (Parzen, 1962) :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K_0(\frac{x - X_i}{h}),$$

où K_0 est un noyau (rectangulaire, gaussien, ...), et h est appelée la fenêtre.

Autres modèles classiques

- Régression non-paramétrique : on dispose d'observations (X_i, Y_i), $i = 1, \dots n$ i.i.d. telle que :

$$Y_i = f(X_i) + \epsilon_i,$$

où les variables ϵ_i vérifient $\mathbb{E}\epsilon_i = 0$, et $f \in \mathcal{F}$ est inconnue.

- Modèle du bruit blanc gaussien : on observe une trajectoire $\{Y(t),\ t \in [0,1]\}$ du processus Y défini par :

$$dY(t) = f(t)dt + \epsilon dW(t),\ t \in [0,1],$$

où W est le processus de Wiener standard sur $[0,1]$, f une fonction inconnue.

- Problème inverse statistique : $Af(t)dt$ au lieu $f(t)dt$, A compact est connu.

Décomposition en valeurs singulières de A

Pour $A : \mathcal{H} \to \mathcal{K}$ compact, on considère $(\varphi_k)_{k\in\mathbb{N}^*}$ b.o.n. de $\mathcal{H}$ de fonctions propres de A^*A et on note $(b_k^2)_{k\in\mathbb{N}^*}$ ses valeurs propres correspondantes. On peut construire $(\psi_k)_{k\in\mathbb{N}^*}$ b.o.n. de $\mathcal{K}$ et on a :

$$A\varphi_k = b_k\psi_k \text{ et } A^*\psi_k = b_k\varphi_k.$$

La suite $(b_k)_{k\in\mathbb{N}^*}$ est appelé suite des valeurs singulières de A et on a clairement $b_k \longrightarrow 0$.
Il nous reste à projeter Y dans la base $(\psi_k)_{k\in\mathbb{N}^*}$ et on obtient :

Modèle de suites gaussiennes

$$y_k := < Y, \psi_k > = b_k \theta_k + \epsilon \xi_k,\ k \in \mathbb{N}^*,$$

avec :

- (y_k) suite d'observations,
- $b_k \to 0$ connus (cas direct $b_k \equiv 1$),
- $(\theta_k) = (< f, \varphi_k >)$ coefficients de f à estimer,
- (ξ_k) suite i.i.d. de variables aléatoires $\mathcal{N}(0,1)$,
- $\epsilon > 0$ niveau de bruit.

But : estimer la suite $(\theta_k)_{k \geq 1}$ à l'aide des observations $(y_k)_{k \geq 1}$.

Estimateur, risque

Idée naturelle pour estimer θ_k : $\hat{\theta}_k = y_k b_k^{-1}$ puisque pour tout k, $\mathbb{E}\hat{\theta_k} = \theta_k$.
Pour $\hat{\theta}$ estimateur de θ, on considère le risque quadratique suivant :

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta \|\theta - \hat{\theta}\|^2 = \mathbb{E}_\theta \sum_{k \geq 1} (\hat{\theta}_k - \theta_k)^2.$$

On obtient pour notre estimateur :

$$R(\theta, \hat{\theta}) = \epsilon^2 \sum_{k \geq 1} b_k^{-2} = +\infty !!$$

Conclusion : on ne peut pas estimer tous les paramètres θ_k.

Estimateurs linéaires

Pour contrôler le risque, on peut introduire la famille des estimateurs linéaires :

$$\{\hat{\theta}(\lambda) = (\hat{\theta}_k = \lambda_k y_k b_k^{-1})_{k\in\mathbb{N}^*} : \lambda_k \in [0,1], k = 1\ldots\}.$$

On obtient alors :

$$\begin{aligned} R(\hat{\theta}(\lambda),\theta) &= \sum_{k\geq 1}(\lambda_k - 1)^2\theta_k^2 + \epsilon^2\sum_{k\geq 1}\lambda_k^2 b_k^{-2} \\ &= b(\lambda)^2 + \sigma^2(\lambda). \end{aligned}$$

Il faut choisir $(\lambda_k)_{k\geq 1}$ qui réalise **le compromis biais-variance**.

Exemple : régularisation par projection

Considérons la famille d'estimateurs par projection (ou spectral cut-off) $\{\hat{\theta}(N), N \geq 1\}$ définie par :

$$\hat{\theta}_k(N) = \mathbb{1}(k \leq N) y_k b_k^{-1}, \ k = 1, \ldots$$

Le risque quadratique de $\hat{\theta}(N)$ s'écrit :

$$R(\theta, N) = \sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} = b(N)^2 + \sigma^2(N).$$

$\Rightarrow$ Choix de N ?

Solution non-adaptative (cas légèrement mal-posé)

Hypothèse de régularité sur la suite θ :

$$\theta \in \Theta(s, Q) = \{\theta \in l^2(\mathbb{N}) : \sum k^{2s}\theta_k^2 \leq Q\}.$$

Alors on obtient, en supposant $b_k \sim k^{-\beta}$:

$$R(\theta, N) = \sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} \leq QN^{-2s} + N^{2\beta+1}\epsilon^2.$$

En prenant $N_s \sim \epsilon^{-\frac{2}{2s+2\beta+1}}$, on obtient :

$$R(\theta, N_s) \leq Q\epsilon^{\frac{4s}{2s+2\beta+1}}.$$

On dit que $\hat{\theta}(N_s)$ atteint la vitesse de convergence $\epsilon^{\frac{4s}{2s+2\beta+1}}$.

Le compromis biais-variance en estimation de densité

De la même manière, si on considère $\hat{f}_n$ vu précédemment :

$$\begin{aligned}\mathbb{E}(\hat{f}_n(x_0) - f(x_0))^2 &= (\mathbb{E}\hat{f}_n(x_0) - f(x_0))^2 + \mathbb{E}(\hat{f}_n(x_0) - \mathbb{E}\hat{f}_n(x_0))^2 \\ &:= b^2(x_0) + \sigma^2(x_0),\end{aligned}$$

appelée biais et variance de l'estimateur $\hat{f}_n$ au point x_0.
En supposant que $f \in \Sigma(s, L)$, et en prenant un noyau d'ordre $l = [s]$, on obtient :

$$\mathbb{E}(\hat{f}_n(x_0) - f(x_0))^2 \leq C_1 h^{2s} + \frac{C_2}{nh},$$

d'où l'importance de calibrer la fenêtre h de l'estimateur.

Le compromis biais-variance en estimation de densité

De la même manière, si on considère $\hat{f}_n$ vu précédemment :

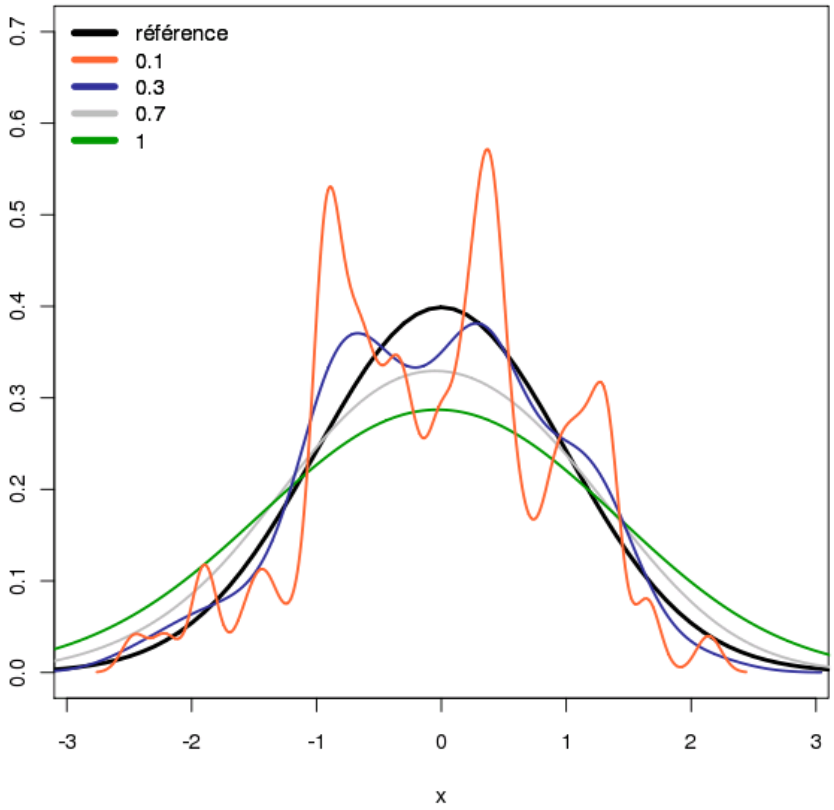
$$\begin{aligned}\mathbb{E}(\hat{f}_n(x_0) - f(x_0))^2 &= (\mathbb{E}\hat{f}_n(x_0) - f(x_0))^2 + \mathbb{E}(\hat{f}_n(x_0) - \mathbb{E}\hat{f}_n(x_0))^2 \\ &:= b^2(x_0) + \sigma^2(x_0),\end{aligned}$$

appelée biais et variance de l'estimateur $\hat{f}_n$ au point x_0.
En supposant que $f \in \Sigma(s, L)$, et en prenant un noyau d'ordre $l = [s]$, on obtient :

$$\mathbb{E}(\hat{f}_n(x_0) - f(x_0))^2 \leq C n^{-\frac{2s}{2s+1}}, \text{ pour } h \sim n^{-\frac{1}{2s+1}}.$$

d'où l'importance de calibrer la fenêtre h de l'estimateur.

Illustration



Vitesse minimax

Le **risque minimax** sur $\Theta(s)$ est définie par

$$r(\Theta(s)) = \inf_{\hat{f}} \sup_{f \in \Theta(s)} R(\hat{f}, f),$$

et on dira que $\hat{f}$ atteint la vitesse minimax sur $\Theta(s)$ s'il existe une constante $C \geq 1$ telle que :

$$\sup_{f \in \Theta(s)} R(\hat{f}, f) \leq C r(\Theta(s)).$$

Dans notre cadre, on a :

$$r(\Theta(s, Q)) \approx \epsilon^{\frac{4s}{2s+2\beta+1}} \text{ et } r(\Sigma(s, L)) \approx n^{-\frac{2s}{2s+1}},$$

et ainsi $\hat{\theta}(N_s)$ et $\hat{f}_n$ atteignent la vitesse minimax.

Approche minimax

- Le meilleur estimateur au sens minimax est celui dont le risque maximal sur $\Theta(s)$ est le plus petit.
- Approche pessimiste et *qui dépend de connaissances préalables sur la fonction à estimer*.

Ici, $\hat{\theta}(N_s)$ dépend de s, régularité de la fonction à estimer. On dit que cet estimateur est non-adaptatif.

Vitesse minimax adaptative : Vitesse atteinte quelquesoit la régularité de f.

L'approche oracle

Etant donnée une famille d'estimateurs $\{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ de f, on définit l'oracle f_{λ^*} par :

$$R(f, f_{\lambda^*}) = \inf_{\lambda \in \Lambda} R(f, \hat{f}_\lambda).$$

f_{λ^*} n'est pas un estimateur (dépend de f inconnue) !
On cherche un estimateur $\tilde{f} = \hat{f}_{\hat{\lambda}}$ vérifiant une inégalité oracle, i.e. :

$$R(f, \hat{f}_{\hat{\lambda}}) \leq C_\epsilon \inf_{\lambda \in \Lambda} R(f, \hat{f}_\lambda) + r_\epsilon,$$

où $C_\epsilon \geq 1$ (proche de 1) et r_ϵ négligeable.

Oracle vs Minimax

Deux approches différentes :

- minimax : on cherche la meilleure vitesse, étant donné $f \in \Theta(s)$.
- oracle : étant donné une famille d'estimateurs,on cherche le meilleur estimateur.

Ainsi :

- minimax : garantit une certaine performance (mais sous une hypothèse de régularité sur f) ;
- oracle : dépend de la famille d'estimateurs (mais aucune hypothèse sur f).

L'approche oracle pour choisir N

Etant donnée $\{\hat{\theta}(N),\, N \geq 1\}$, on cherche N qui s'approche de

$$N^* = \arg\min_N R(\theta, N).$$

On dira que $\hat{\theta}(\hat{N})$ satisfait une **inégalité oracle exacte** lorsque :

$$R(\hat{\theta}(\hat{N}), \theta) \leq (1 + \rho_\epsilon) R(\theta, N^*) + r_\epsilon,$$

avec $\rho_\epsilon \longrightarrow 0$ lorsque $\epsilon \longrightarrow 0$ et r_ϵ terme résiduel.

Exemple : la méthode du risque sans biais (URE)

On va estimer le risque $R(\theta, N)$ par un estimateur sans biais $U(y, N)$ en utilisant les observations $y_k = b_k\theta_k + \epsilon\xi_k$, $k = 1, \ldots$.
Puis on minimise sur $N \geq 1$ l'estimateur du risque.
On a

$$U(y, N) = \sum_{k>N} b_k^{-2}(y_k^2 - \epsilon^2) + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} \text{ e.s.b. de } R(\theta, N),$$

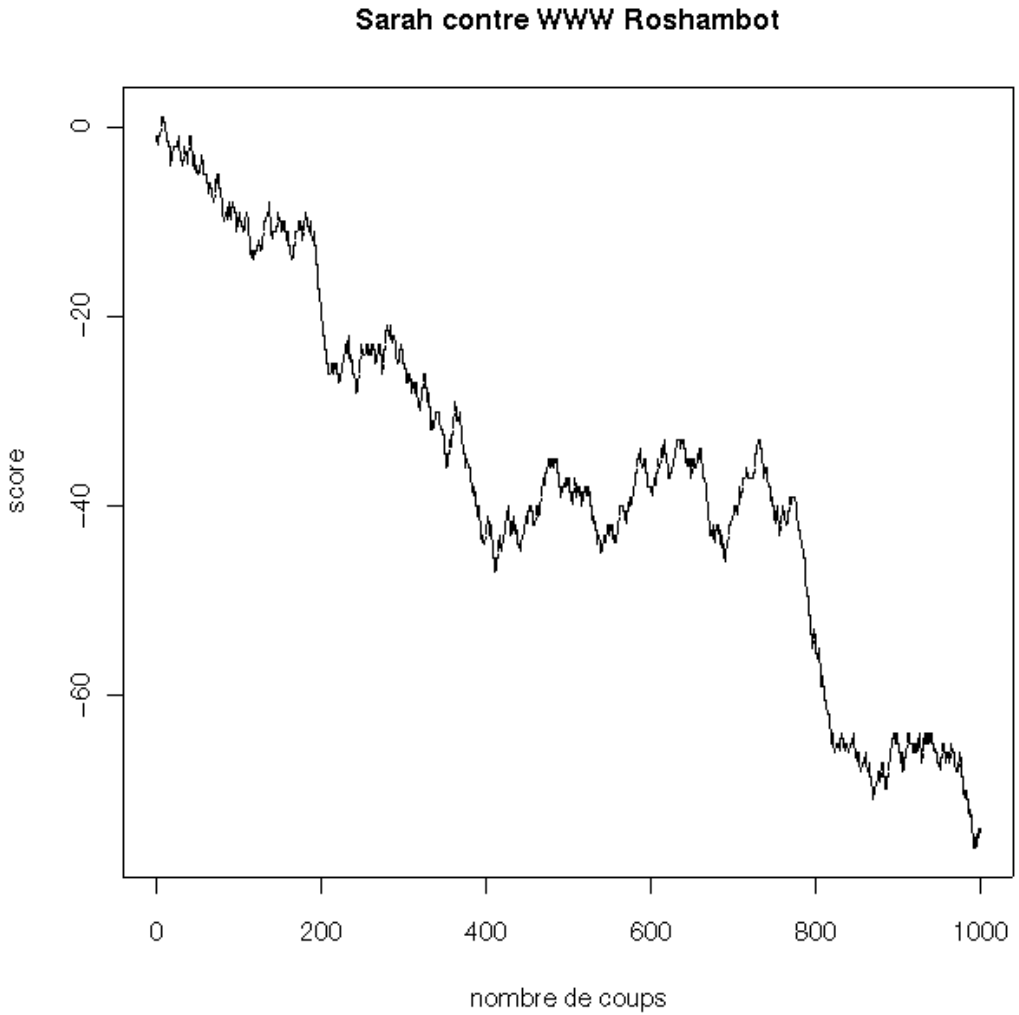
et on obtient le choix suivant de N :

$$N_{URE} = \arg\min_N \left(-\sum_{k=1}^{N} b_k^{-2} y_k^2 + 2\epsilon^2 \sum_{k=1}^{N} b_k^{-2} \right).$$

On a bien, $\forall \gamma > 0$,

$$R(\hat{\theta}(N_{URE}), \theta) \leq (1+\gamma) R(\theta, N^*) + C^* \frac{\epsilon^2}{\gamma}.$$

Apprentissage : une illustration



Apprentissage statistique

On observe (X_i, Y_i), $i = 1, \dots n$ ensemble d'apprentissage de loi P sur $\mathcal{X} \times \mathcal{Y}$ inconnue avec :

- $\mathcal{X}$ quelconque,
- $\mathcal{Y} \subset \mathbb{R}$.

But : on veut "apprendre", à partir des observations, la réponse Y d'une nouvelle observation X.
Exemples :

- Classification binaire : $\mathcal{Y} = \{-1, 1\}$.
- Régression : $\mathcal{Y} = \mathbb{R}$.
- Statistique fonctionnelle : $\mathcal{X}$ espace fonctionnel.

Applications

Dynamic Reconstruction of Chaotic Systems ⊕Protein Structure Prediction ⊕Identification of alternative exons using SVM ⊕Breast cancer diagnosis and prognosis ⊕Support Vector Machines Based Modeling of Seismic Liquefaction PotentialSVM for Geo- and Environmental Sciences ⊕SVM for Protein Fold and Remote Homology Detection ⊕Detecting Steganography in digital images ⊕Breast Cancer Prognosis : Chemotherapy Effect on Survival Rate ⊕Text Categorization ⊕Facial expression classification ⊕Application of The Kernel Method to the Inverse Geosounding Problem ⊕Support Vector Machine Classification of Microarray Gene Expression Data ⊕Intervals Using Least Squares Support Vector Machines ⊕Support Vector Machines For Texture Classification ⊕SVM application in E-learning ⊕Support vector machines-based generalized predictive control ⊕Isolated Handwritten Jawi Characters Categorization Using Support Vector Machines (SVM). ⊕Image Clustering ⊕ewsRec, a SVM-driven Personal Recommendation System for News Websites ⊕Equbits Foresight ⊕Speaker/speech recognition ⊕Student in AI ⊕Analysis and Applications of Support Vector Forecasting Model Based on Chaos Theory⊕Image classification⊕Object

Le modèle de classification binaire

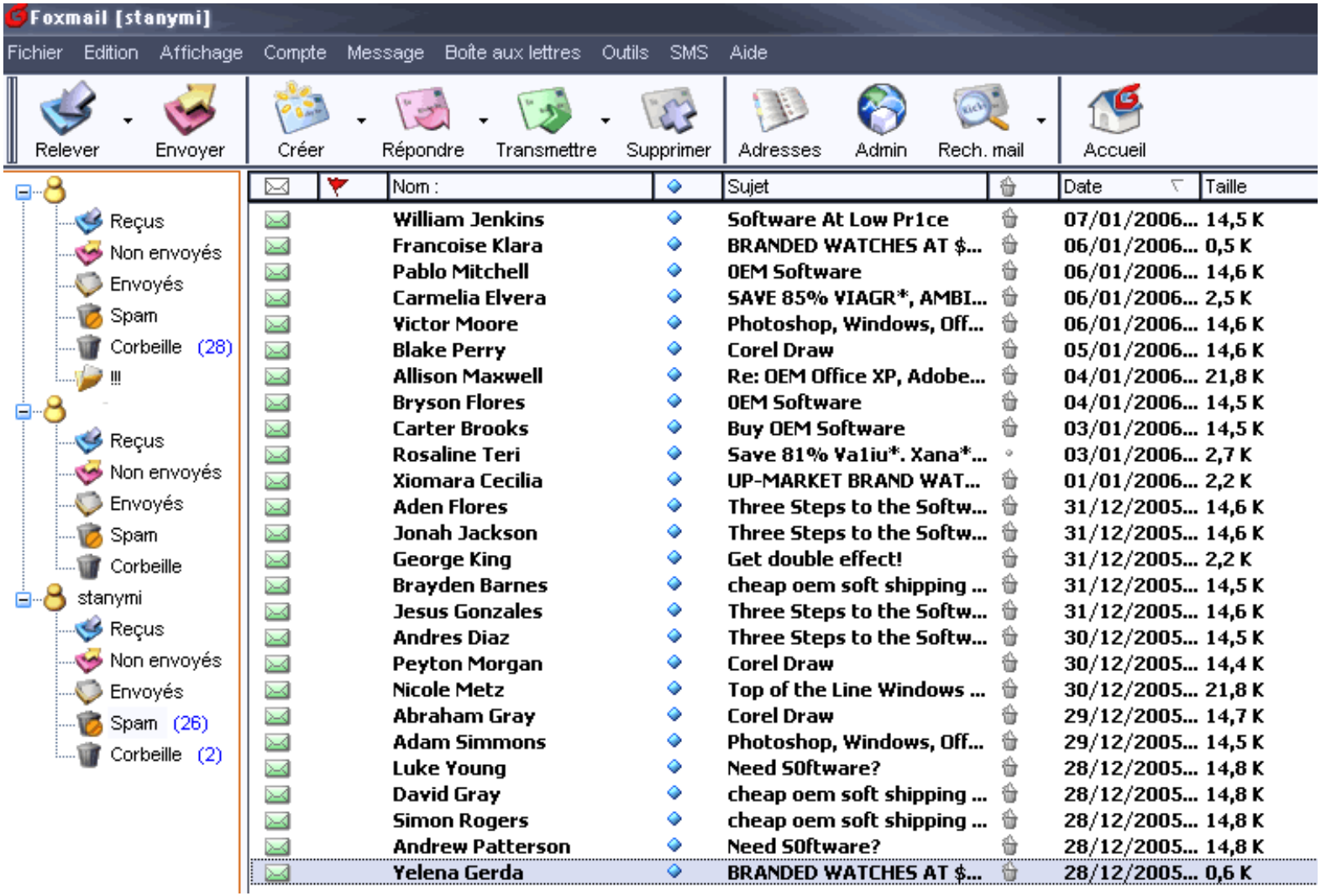
- On observe $(X_1, Y_1), \dots (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ i.i.d. de loi π, où :
  - $\mathcal{X} = \mathbb{R}^d$,
  - $\mathcal{Y} = \{-1, +1\}$, classe correspondante.
- But : $X \to Y$? avec $\hat{f} : \mathbb{R}^d \to \{-1, +1\}$ classifieur.
- On définit le risque de $\hat{f}$ par $R(\hat{f}) = \mathbb{P}(\hat{f}(X) \neq Y)$ et on a :

$$f^* = \arg\min R(f) = \operatorname{sign}(2\eta - 1),$$

où $\eta(x) = \mathbb{P}(Y = 1 | X = x)$.

- On veut contrôler l'excès de risque $R(\hat{f}, f^*) = R(\hat{f}) - R(f^*)$.

Exemple


Foxmail [stanymi]
Fichier Edition Affichage Compte Message Boîte aux lettres Outils SMS Aide
Relever Envoyer Créer Répondre Transmettre Supprimer Adresses Admin Rech. mail Accueil
Reçus
Non envoyés
Envoyés
Spam
Corbeille (28)
!!!
Reçus
Non envoyés
Envoyés
Spam
Corbeille
stanymi
Reçus
Non envoyés
Envoyés
Spam (26)
Corbeille (2)
Nom : Sujet Date Taille
William Jenkins Software At Low Pr1ce 07/01/2006... 14,5 K
Francoise Klara BRANDED WATCHES AT $... 06/01/2006... 0,5 K
Pablo Mitchell OEM Software 06/01/2006... 14,6 K
Carmelia Elvera SAVE 85% VIAGR*, AMBI... 06/01/2006... 2,5 K
Victor Moore Photoshop, Windows, Off... 06/01/2006... 14,6 K
Blake Perry Corel Draw 05/01/2006... 14,6 K
Allison Maxwell Re: OEM Office XP, Adobe... 04/01/2006... 21,8 K
Bryson Flores OEM Software 04/01/2006... 14,5 K
Carter Brooks Buy OEM Software 03/01/2006... 14,5 K
Rosaline Teri Save 81% Va1iu*. Xana*... 03/01/2006... 2,7 K
Xiomara Cecilia UP-MARKET BRAND WAT... 01/01/2006... 2,2 K
Aden Flores Three Steps to the Softw... 31/12/2005... 14,6 K
Jonah Jackson Three Steps to the Softw... 31/12/2005... 14,6 K
George King Get double effect! 31/12/2005... 2,2 K
Brayden Barnes cheap oem soft shipping ... 31/12/2005... 14,5 K
Jesus Gonzales Three Steps to the Softw... 31/12/2005... 14,6 K
Andres Diaz Three Steps to the Softw... 30/12/2005... 14,5 K
Peyton Morgan Corel Draw 30/12/2005... 14,4 K
Nicole Metz Top of the Line Windows ... 30/12/2005... 21,8 K
Abraham Gray Corel Draw 29/12/2005... 14,7 K
Adam Simmons Photoshop, Windows, Off... 29/12/2005... 14,5 K
Luke Young Need SOftware? 28/12/2005... 14,8 K
David Gray cheap oem soft shipping ... 28/12/2005... 14,8 K
Simon Rogers cheap oem soft shipping ... 28/12/2005... 14,8 K
Andrew Patterson Need SOftware? 28/12/2005... 14,8 K
Yelena Gerda BRANDED WATCHES AT $... 28/12/2005... 0,6 K

Idée naturelle : Minimisation du risque empirique

On considère $\mathcal{C}$ ensemble de classifieurs.

$$f_{\mathcal{C}}^{*} \xrightarrow{\text{minimise}} \mathbb{P}(f(X) \neq Y)$$

Idée naturelle : Minimisation du risque empirique

On considère $\mathcal{C}$ ensemble de classifieurs.

$$f_{\mathcal{C}}^* \xrightarrow{\text{minimise}} \mathbb{P}(f(X) \neq Y)$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(f(X_i) \neq Y_i) \quad \uparrow \text{estime}$$

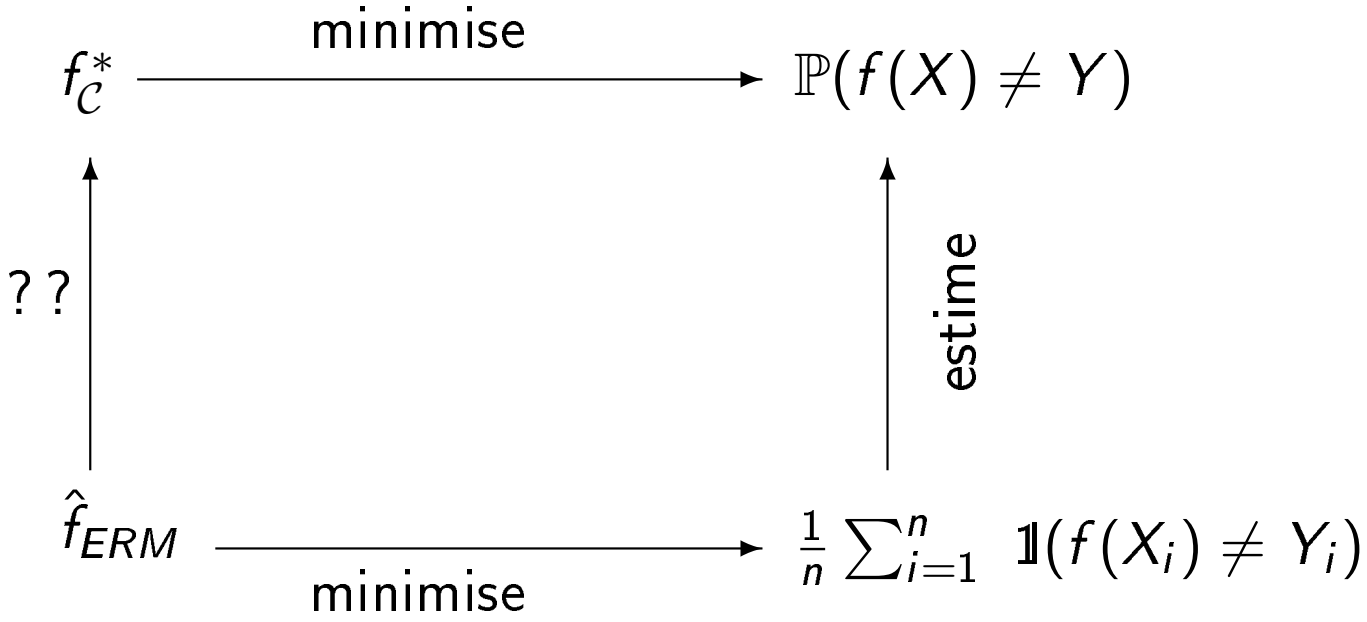
Idée naturelle : Minimisation du risque empirique

On considère $\mathcal{C}$ ensemble de classifieurs.

$$
\begin{array}{ccc}
f_{\mathcal{C}}^{*} & \xrightarrow{\text{minimise}} & \mathbb{P}(f(X) \neq Y) \\
 & & \uparrow \text{estime} \\
\hat{f}_{ERM} & \xrightarrow[\text{minimise}]{} & \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(f(X_i) \neq Y_i)
\end{array}
$$

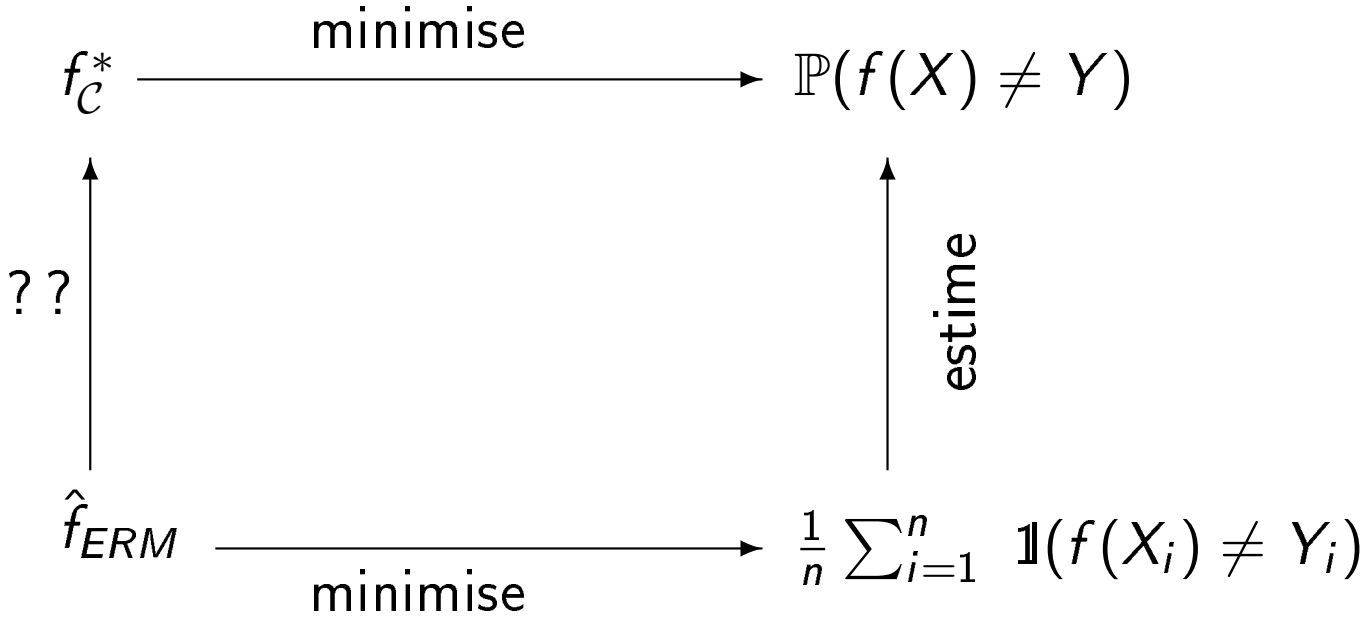
Idée naturelle : Minimisation du risque empirique

On considère $\mathcal{C}$ ensemble de classifieurs.



Idée naturelle : Minimisation du risque empirique

On considère $\mathcal{C}$ ensemble de classifieurs.



Si $f^* \in \mathcal{C}$, on a :

$$\mathbb{E}_{\pi^n} R(\hat{f}_{ERM}, f^*) \underset{n \to +\infty}{\longrightarrow} 0.$$

Vitesses des ERM

Si $f^* \in \mathcal{C}$, on a aussi des vitesses de convergence :

- $\mathbb{E}R(\hat{f}_{ERM}, f^*) \leq Cn^{-\frac{1}{2}}$ lorsque la dimension de Vapnik de f^* est finie (Vapnik et Chervonenkis 1982)
- Vitesse rapide lorsque $R(f^*) = 0$: $\mathbb{E}R(\hat{f}_{ERM}, f^*) \leq Cn^{-1}$.
- Plus récemment, vitesse minimax $n^{-\frac{\kappa}{2\kappa+\rho-1}}$ avec :
  - $0 < \rho < 1$ est la complexité de $\mathcal{C}$.
  - $\kappa \geq 1$ paramétre de marge, i.e. :

$$\mathbb{P}(|2\eta(x) - 1| \leq t) \leq ct^{\frac{1}{\kappa-1}}, \text{ pour } t \to 0.$$

Vitesses des ERM

Si $f^* \in \mathcal{C}$, on a aussi des vitesses de convergence :

- $\mathbb{E}R(\hat{f}_{ERM}, f^*) \leq Cn^{-\frac{1}{2}}$ lorsque la dimension de Vapnik de f^* est finie (Vapnik et Chervonenkis 1982)
- Vitesse rapide lorsque $R(f^*) = 0$: $\mathbb{E}R(\hat{f}_{ERM}, f^*) \leq Cn^{-1}$.
- Plus récemment, vitesse minimax $n^{-\frac{\kappa}{2\kappa+\rho-1}}$ avec :
  - $0 < \rho < 1$ est la complexité de $\mathcal{C}$.
  - $\kappa \geq 1$ paramétre de marge, i.e. :

$$\mathbb{P}(|2\eta(x) - 1| \leq t) \leq ct^{\frac{1}{\kappa-1}}, \text{ pour } t \to 0.$$

A-t'on $f^* \in \mathcal{C}$? ?

Problème des ERM : le choix de $\mathcal{C}$!

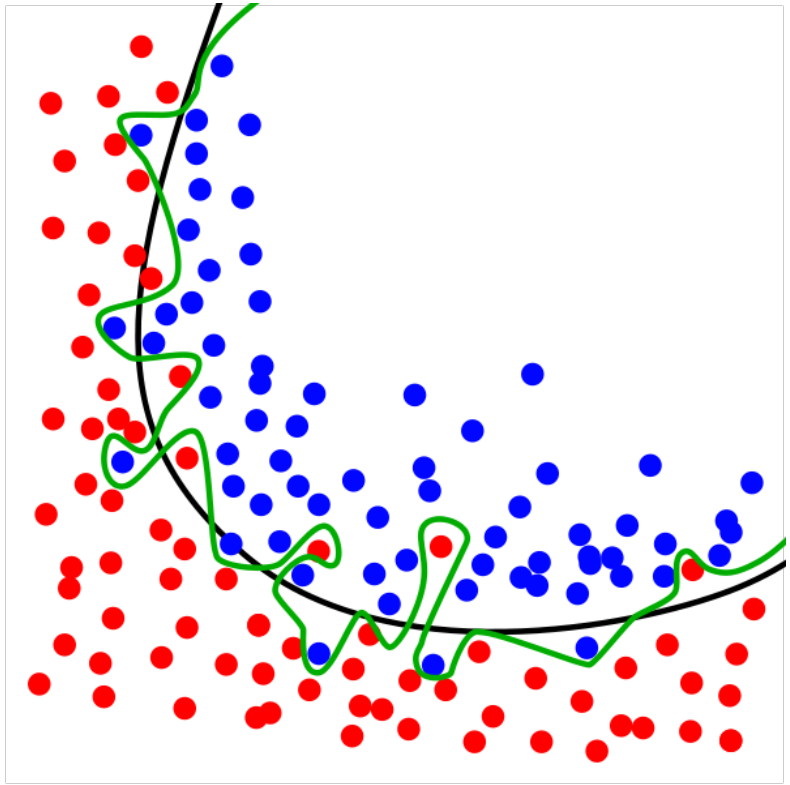
On peut écrire dans le cas général

$$R(\hat{f}_{ERM}, f^*) = \left(R(\hat{f}_{ERM}) - \inf_{\mathcal{C}} R(f) \right) + \left(\inf_{\mathcal{C}} R(f) - R(f^*) \right).$$

$\Rightarrow$ La taille de $\mathcal{C}$ doit réaliser un compromis :

- $\mathcal{C}$ trop grand : l'erreur d'estimation est trop grande.
- $\mathcal{C}$ trop petit : l'erreur d'approximation est trop grande.

Le sur-apprentissage



$\mathcal{C}$ trop grand $\Rightarrow$ solution très instable.

ERM pénalisé

Si $\mathcal{C}$ est suffisamment riche,

$$\min_{f \in \mathcal{C}} R_n(f) = 0 \Rightarrow \text{sur} - \text{apprentissage}.$$

On tient compte de la complexité de la solution. Par exemple :

$$\min_{f \in \mathcal{C}} \left[R_n(f) + \alpha \Omega(f) \right],$$

où $\Omega(f)$ mesure la complexité de f et α est un paramètre de régularisation.

Exemple :

- $\Omega(f) = \|f\|^2_{\mathcal{H}_K}$ (SVM).
- $\Omega(f) = \|f\|_1$ (LASSO).

Le classifieur SVM (Support Vector Machines)

classifieur SVM

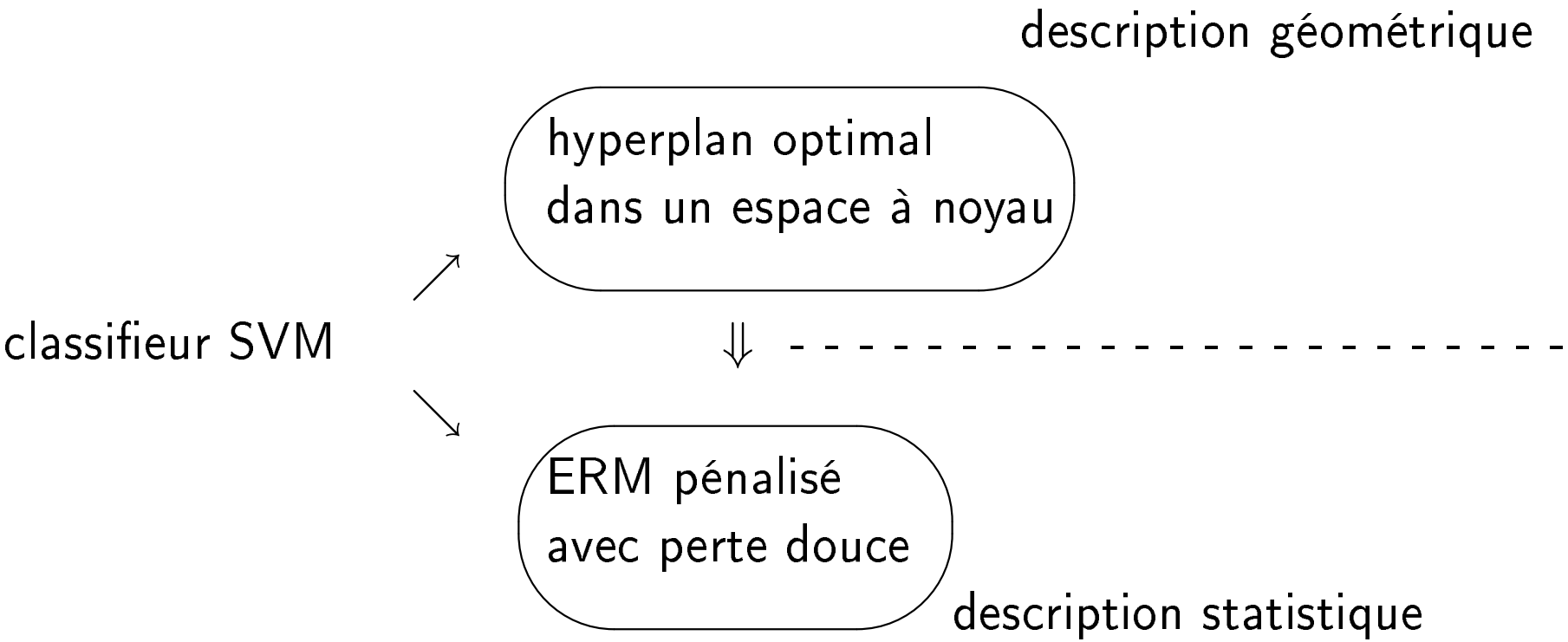
Le classifieur SVM (Support Vector Machines)

description géométrique

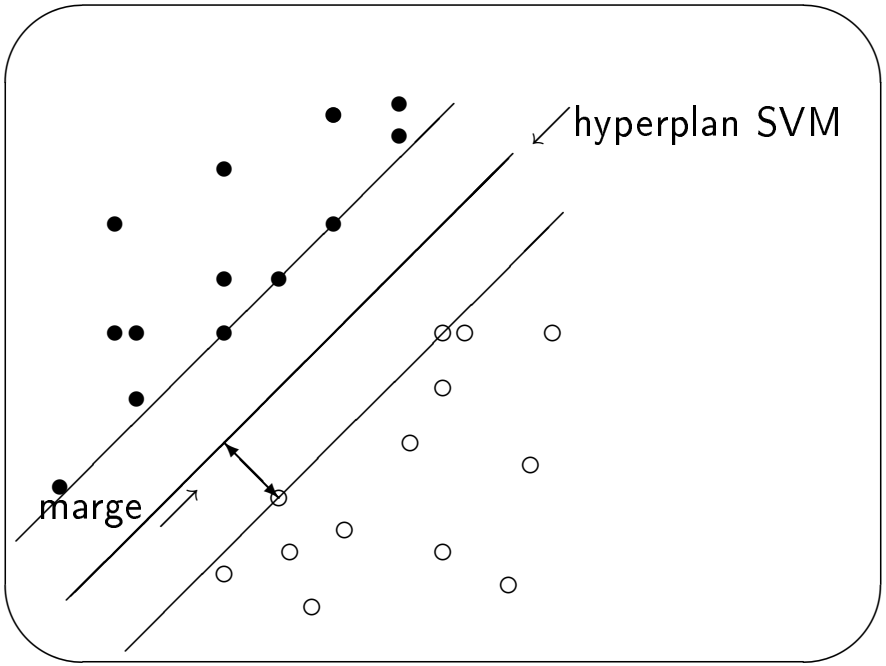
hyperplan optimal
dans un espace à noyau

classifieur SVM

Le classifieur SVM (Support Vector Machines)



Support Vector Machines : cas linéaire



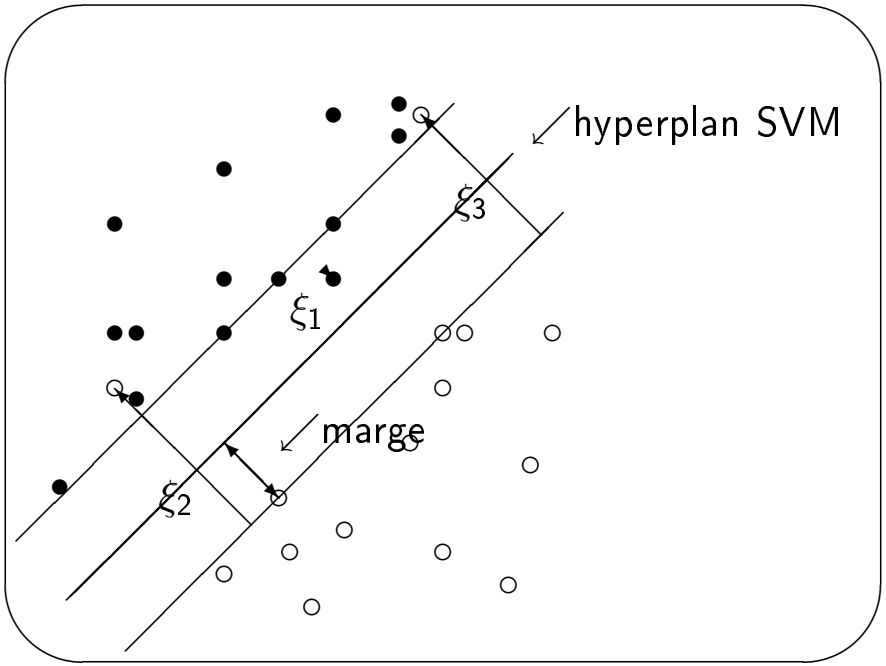
Cas linéaire sans bruit, $\mathcal{X} = \mathbb{R}^2$.

Hyperplan maximisant la marge :

$$\begin{cases} \max\limits_{w,b} \ m \\ \\ \forall\, i = 1, .., n \ y_i f(x_i) \geq m, \end{cases}$$

où $f(x) =< w, x > +b$.

SVM : variables ressorts dans le cas bruit



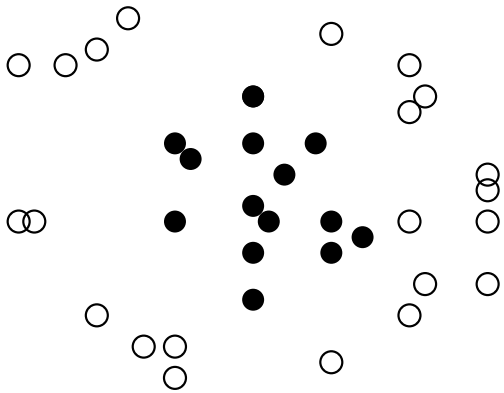
Cas linéaire bruit, $\mathcal{X} = \mathbb{R}^2$.

On rajoute des variables ressorts ξ :

$$(*) \begin{cases} \max\limits_{w,b} \ (m - C \sum_{i=1}^{n} \xi_i) \\ \\ y_i f_{w,b}(x_i) \geq 1 - \xi_i \, , \ \xi_i \geq 0, \end{cases}$$

- $f(x) =< w, x > +b$.

Problème non linéaire

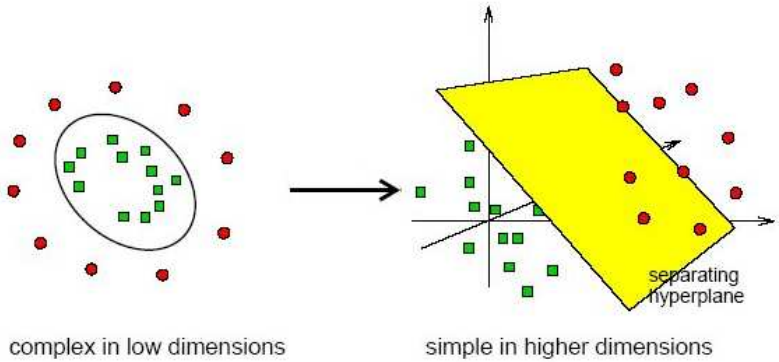


Pas d'hyperplan qui sépare...

$\Rightarrow$ méthode à noyau

Le "Kernel trick"

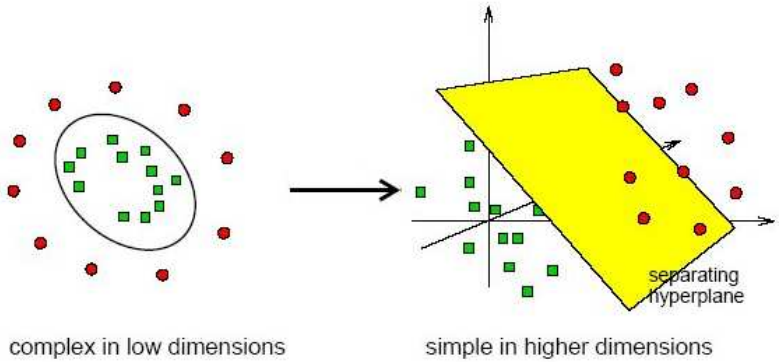
Du linéaire au non-linéaire avec $\Phi : \mathcal{X} \to \Phi(\mathcal{X})$



$$\max_{v:0\leq v_i\leq C} L_D = \max_{v:0\leq v_i\leq C} \left(\sum_{i=1}^{n} v_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j Y_i Y_j < X_i, X_j > \right).$$

Le "Kernel trick"

Du linéaire au non-linéaire avec $\Phi : \mathcal{X} \to \Phi(\mathcal{X})$



$$\max_{v:0\leq v_i\leq C} L_D = \max_{v:0\leq v_i\leq C} \left(\sum_{i=1}^{n} v_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j Y_i Y_j K(X_i, X_j) \right) .$$

Définition Un noyau est une application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que :

$$K(x, y) = < \Phi(x), \Phi(y) >_{\Phi(\mathcal{X})} .$$

Espace de Hilbert à noyau reproduisant (EHNR)

Définitions

- On appelle noyau une application $K : \mathcal{X}^2 \to \mathbb{R}$ symétrique définie-positive.
- L'EHNR $\mathcal{H}_K$ est un espace de Hilbert de fonction $f : \mathcal{X} \to \mathbb{R}$ vérifiant :
  - $K(x, \cdot) \in \mathcal{H}_K$, $\forall x \in \mathcal{X}$,
  - $< f, K(x, \cdot) >_K = f(x)$, $\forall f \in \mathcal{H}_K$.

  K est appelé le noyau reproduisant de $\mathcal{H}_K$.

Exemples pour $\mathcal{X} = \mathbb{R}^d$:

- noyau gaussien $K(x, y) = \exp(-\sigma^2 \|x - y\|^2)$.
- noyau Laplace $K(x, y) = \exp(-\sigma \|x - y\|)$.

L'algorithme des SVM

L'algorithme SVM peut s'écrire :

$$\min_{f \in \mathcal{H}_K} \left[\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \alpha \|f\|^2_{\mathcal{H}_K} \right],$$

où

- $l(y, f(x)) = (1 - yf(x))_+$ est la perte SVM,
- α est un paramètre de régularisation,
- $\mathcal{H}_K$ est un espace de Hilbert à noyau reproduisant.

Théorème de représentation $\hat{f}_{SVM}(x) = \sum_{i=1}^{n} v_i^* Y_i K(X_i, x)$.

Vitesses de convergence des SVM

On veut choisir α pour obtenir des vitesses de convergence du type :

$$\mathbb{E}R(\hat{f}_{SVM}, f^*) \leq Cn^{-\beta}.$$

On procède en deux étapes :

- Obtenir une inégalité oracle :

$$\mathbb{E}R(\hat{f}_{SVM}, f^*) \leq C \inf_{f \in \mathcal{H}_K} \left[R(f, f^*) + \alpha \|f\|_{\mathcal{H}}^2\right] + \delta(n).$$

- Contrôler l'erreur d'approximation :

$$a(\alpha) := \inf_{f \in \mathcal{H}_K} \left[R(f, f^*) + \alpha \|f\|_{\mathcal{H}}^2\right].$$

Vitesses de convergence non-adaptative

Soit π une probabilité sur $\mathbb{R}^d \times \{-1, 1\}$ telle que :

- π a un paramètre de marge $q \in [0, +\infty]$;
- $f^* \in \mathcal{B}_{s2\infty}(\mathbb{R}^d)$ pour $s > 0$.

On considère la minimisation SVM avec noyau Sobolev K_r, $r > d$.
Si on choisit α tel que

$$\alpha = n^{-\frac{r(r-s)(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}},$$

alors il existe $C > 0$ telle que :

$$\mathbb{E}R(\hat{f}_n, f^*) \leq Cn^{-\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}.$$

Choix de α : méthode adaptative

Principe de la méthode d'aggrégation :

- On sépare les observations $D_n = (D_{n_1}^1, D_{n_2}^2)$.
- On construit avec $D_{n_1}^1$ une famille de classifieurs SVM

$$\{\hat{f}_{\alpha_1}, \ldots, \hat{f}_{\alpha_M}\} \text{ où } \{\alpha_1, \ldots \alpha_M\} = \Lambda \text{ est une grille.}$$

- On calcule avec $D_{n_2}^2$ une suite de poids w_k, pour $k \in \{1 \ldots M\}$.
- On construit notre agrégat $\tilde{f}_n$ tel que

$$\tilde{f}_n = \sum_{k=1}^{M} w_k \hat{f}_{\alpha_k}.$$

Expérimentations

On a implémenté notre agrégat dans 2 cas :

- Cas Sobolev : $\tilde{f}_n$ issu de l'approche décrite précédemment. Noyau utilisé : $K_\sigma(x, y) = \exp(-\sigma\|x - y\|)$.
- Cas gaussien : $\tilde{f}_n$ issu des résultats de Steinwart et Scovel (2007). Noyau utilisé : $K_\sigma(x, y) = \exp\left(-\sigma\|x - y\|^2\right)$.

Données de classification

Dataset	d	n	p	realizations
Banana	2	400	4900	100
Titanic	3	150	2051	100
Thyroid	5	140	75	100
Diabetis	8	468	300	100
Breast-cancer	9	200	77	100
Flare-solar	9	666	400	100
Heart	13	170	100	100
Image	18	1300	1010	20
Waveform	21	400	4600	100

"Dataset"$=\{(D_n^1, T_p^1), (D_n^2, T_p^2), \ldots, (D_n^{100}, T_p^{100})\}$.

Résultats expérimentaux

Dataset	Laplace Aggregate	Gaussian Aggregate
Banana	11.31± 0.57	11.43±0.84
Titanic	22.77±1.13	22.57±0.79
Thyroid	5.45±2.68	6.31±2.97
Diabetis	28.34±2.27	27.80±2.06
Breast-cancer	32.74±5.16	32.13±4.77
Flare-solar	35.69±1.93	34.87±1.82
Heart	22.12±3.98	22.62±3.77
Image	3.95±0.74	5.66±0.74
Waveform	14.12±0.72	15.04±0.79

Résultats expérimentaux

Dataset	Laplace Aggregate	Gaussian Aggregate	Rästch et al. (2001)
Banana	11.31± 0.57	11.43±0.84	11.53±0.66
Titanic	22.77±1.13	22.57±0.79	22.42±1.02
Thyroid	5.45±2.68	6.31±2.97	4.80±2.19
Diabetis	28.34±2.27	27.80±2.06	23.53±1.76
Breast-cancer	32.74±5.16	32.13±4.77	26.04±4.74
Flare-solar	35.69±1.93	34.87±1.82	32.43±1.82
Heart	22.12±3.98	22.62±3.77	15.95±3.26
Image	3.95±0.74	5.66±0.74	2.96±0.6
Waveform	14.12±0.72	15.04±0.79	9.88±0.83