théorie de décision → apprentissage

$D_n : \{(X_i, Y_i)\}_{i=1}^{n} \overset{iid}{\sim} P$

$l : \mathcal{A} \times (\mathcal{X} \times \mathcal{Y})$

$f : \mathcal{X} \longrightarrow \mathcal{Y}$

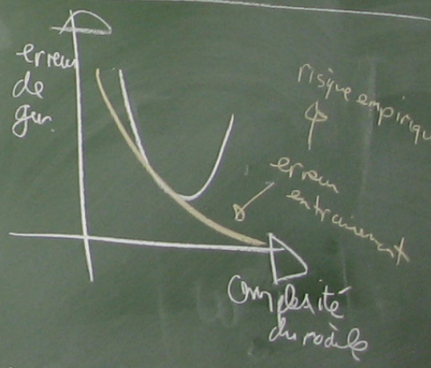
$R_P(f) = \mathbb{E}_{(X,Y)\sim P}\left[l(f(X), Y) \right]$

$\mathcal{A}(D_n) = \hat{f}_n$

tâches

tâches	pertes	alg
classification	0-1	
régression	quadratique	régression linéaire
estimation de densité	$-\log p(y)$	approche proba. MV régression logistique etc.

- surapprentissage ← complexité

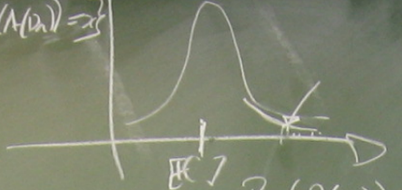


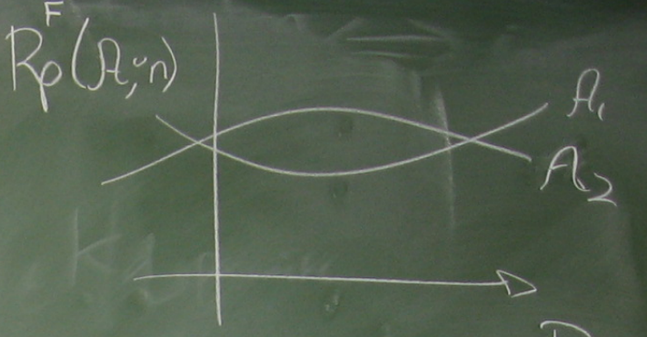
besoin de régularisation

- réduire :
  - implicite
  - explicite

→ borne d'Occam

→ C_L de Mallows





THÉORIE

- risque fréquentiste $\mathbb{E}_{D_n \sim P^{\otimes n}}\left[R_p(A(D_n))\right]$

Sample complexity : n t.q. $R_p^F(A,n) - R_p(f_p^*) \leq \varepsilon$

uniforme : $\sup_{P \in \mathcal{P}} \left[R_p^F(A;n) - R_p(f_p^*)\right]$

espérance de l'excès de risque

- PAC $P\{R_p(A(D_n)) \leq \text{borne}(\_\_)\} \geq 1-\delta$

sample complexity : n t.q. $R_p(A(D_n)) - R_p(f_p^*) \leq \varepsilon$ avec prob. $\geq 1-\delta$

Consistance

$\lim_{n\to\infty} [\ \ ] \to 0$

excès de risque

Consistance uniforme universelle → pas possible ∄ règle (no free lunch)

→ possible ∄ finie

Outils :

- loi des grands nombres
- thm. centrale limite
- } asymptotique
- inégalités de concentration — Chernoff, Hoeffding

R_p^F → perte quadratique

décomposition biais-variance

$$R_p^F(\hat{f}_n) - R_p(f_p^*) = \underbrace{\mathbb{E}_X\left[\left(\mathbb{E}_{D_n}[\hat{f}_n(x)] - f_p^*(x)\right)^2\right]}_{\text{biais}} + \underbrace{\mathbb{E}_{D_n,X}\left[\left(\hat{f}_n(x) - \mathbb{E}_{D_n}[\hat{f}_n(x)]\right)^2\right]}_{\text{variance}}$$

- James-Stein
- C_L Mallows X design fixe } régression
- Consistance alg. partition [plug-in]

$$\begin{cases} h_n \to 0 \\ n h_n^d \to +\infty \end{cases} \begin{matrix} \} \text{biais} \to 0 \\ \} \text{variance} \to 0 \end{matrix}$$

$n \gtrsim \frac{1}{h_n^d}$ [fléau de la dimension]

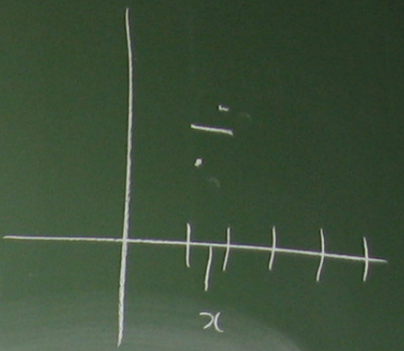
borne d'Occam

$$R_p(f) \leq \hat{R}_n(f) + \frac{1}{\sqrt{2n}}\underbrace{\sqrt{\ln 2|\mathcal{F}| + \ln\frac{1}{\delta}}}_{\text{complexité}(f)}$$

$\forall P$ avec prob. $\geq 1-\delta$

π : à priori

Approches



- moyennage local (régression)

$$\hat{f}(x) = \sum_{i=1}^{n} W_i(x)\, Y_i$$

  - histogramme (partition) $W_i(x) \propto \mathbb{1}\{X_i \in A(x)\}$ $\quad \sum_i W_i(x) = 1$
  - k p.p.v. $\quad \mathbb{1}\{X_i \in V_k(x)\}$

↳ **plug-in** $\quad \hat{f}(x) = \mathbb{1}\{\hat{\eta}(x) \geq \frac{1}{2}\}$ $\quad f^*(x) = \mathbb{E}[Y \mid X = x]$

$Y_i \in \{0,1\}$ ↳ pour perte quadratique

$$\mathbb{E}_{\mathcal{D}_n}\left[R^{01}(\hat{f})\right] - R^{01}(f^*) \leq 2\sqrt{\mathbb{E}\left[R^{reg}(\hat{\eta})\right] - R^{reg}(\eta^*)}$$

Min. risque empirique régularisé

$$\min_f \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f)$$

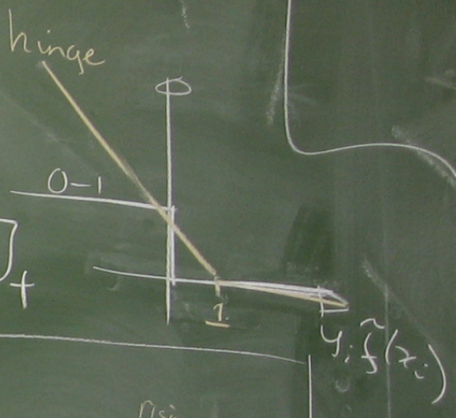
ℓ quadratique → régression ridge

$f : \mathcal{X} \to \mathcal{Y}$

→ problématique en classification NP-dur

$\tilde{\ell}(\tilde{f}(x_i), y_i)$

exemple : pour SVM $\tilde{\ell}$ → hinge $[1 - y_i \tilde{f}(x_i)]_+$



plug-in $\begin{cases} f : \mathcal{X} \to \{-1, 1\} \\ \tilde{f} : \mathcal{X} \to \mathbb{R}^q \end{cases}$

$f(x) \triangleq \text{sgn}\{\tilde{f}(x) \geq 0\}$

Maximum de vraisemblance → perte log

$f_w(x) \to P_w(y=1|x)$

$\ell(f_w(x), y) = -\log P_w(y|x)$

→ régression logistique

Surapprentissage

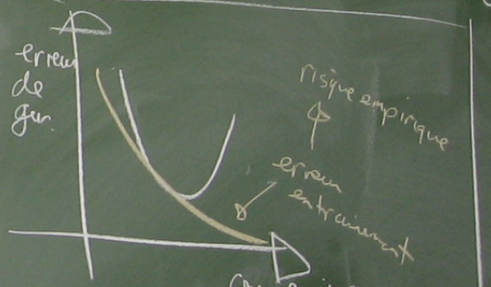
Complexité

Choix de régularisation

- réduire :
  - implicite
  - explicite

→ borne d'Occam

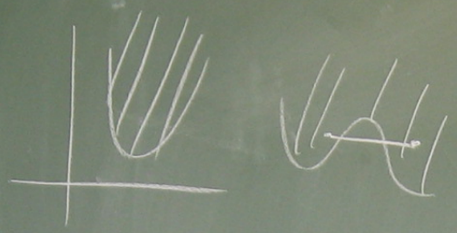
→ CL de Mallows



Optimisation & analyse convexe

classe de fct. "tractable"

- ensembles convexes
- fonctions convexes ⟸ épigraphe
- convexité forte
- Jensen



Algorithmes

- descente de gradient
- ellipsoïde
- Newton (régression logistique → IRLS)

Optimisation sous contrainte

$$\min_x f(x) \quad \text{t.q.} \quad h_i(x)=0,\ g_j(x)\le 0$$

problème dual

$$\sup_{\mu\ge 0,\lambda} \left[\inf_x \underbrace{f(x)+\lambda^T h(x)+\mu^T g(x)}_{L(x,\lambda,\mu)} \right]$$

fct. dual $q(\lambda,\mu)$

$$f(x) \ge f(x^*) \ge q(\lambda^*,\mu^*)$$

dualité forte ⟹ Slater

⟹ conditions KKT pour (x^*,λ^*,μ^*)

- x^* optimise $L(x,\lambda^*,\mu^*)$
- x^*,λ^*,μ^* faisable
- $\mu_j^* g_j(x^*)=0$ "comp. stricte"

* Sélection de modèle / hyper-paramètre (λ)
  - validation croisée
  - C_L Mallows

* générative vs. conditionnelle
  $p(x,y)$ — LDA, QDA
  $p(y|x)$ → régression logistique

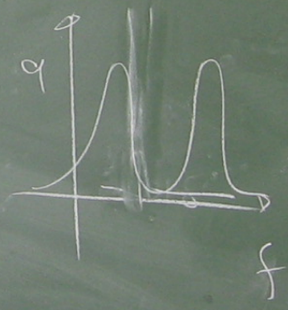
* astuce du noyau
  $\langle \varphi(x), \varphi(x') \rangle = k(x,x')$
  - régression ridge : $X^T X$ ($d \times d$) $\rightsquigarrow$ $XX^T = K$ ($n \times n$)
  - SVM
  - thm. du représentant
  - dualité
  - etc

$$\left(\text{RBF} \quad K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) = \sum_{p=0}^{\infty} \frac{\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)^p}{p!} \right)$$

optimalité

fréquentiste : $\hat{f}$

Bayésienne : $\hat{q}(f)$



$\max_f \hat{q}(f)$ → MAP

$\pi(f)$ $p(D_n|f)$ $\hat{q}(f) \propto p(D_n|f)\pi(f)$ $\max_f p(D_n|f)$

$y = f(x)$ $\int_f f(x)\hat{q}(f)\,df$ posterior mean estimate

$\tilde{X}$

$\rightsquigarrow \underbrace{XX^T}_{n \times n} = K$

Tâche

- classification
- régression

Connaissances à priori

données

$\mathcal{F} \rightsquigarrow \varphi(x)$

$K(x, x')$

ρ (MV)

explicite ou implicite

→ contraintes

→ régularisation / pénalisation

Modèle

algorithme

Validation croisée

principes d'analyse

$\max_f \hat{q}(f)$

$f \rightarrow$ MAP

$\max_f p(D_n | f) \rightarrow$ ML

max. Vrais.