

→ version apprentissage

# théorie de la décision

$D_n = \{(x_i, y_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} P$

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

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

$R_p(f) = \mathbb{E}_{(x,y) \sim P} [l(f(x), y)]$

↳ risque de sans de Vapnik

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

tâches	perte	alg
Classification	0-1	
regression ( $y \in \mathbb{R}$ )	quadratique	régl. linéaire
[estimation de : $\theta \in \mathbb{R}^d$ ? densité : $p(\theta)$ ? -log p(y) ?]		approche M.V. → régl. logistique etc.

## surapprentissage

Complexité

(besoin de régularisation)

gérer explicitement :  $\mathcal{F}$

implicitement → pénalisation / rég.

borne d'Occam

$C_L$  de Mallows

(risque empirique) erreur d'entraînement

# THEORIE

$R_p(A(n))$

$\mathbb{E} [R_p(A(n))]$

- risque fréquentiste  $R_p^F(A(n)) = \mathbb{E}_{D_n \sim P^n} [R_p(\hat{f}_n)]$
- sample complexity  $n$  t.a.  $R_p^F(A(n)) - R_p(f^*) \leq \epsilon$
- PAC  $P \{ R_p(A(n)) \leq \text{borne}(\dots) \} \geq 1 - \delta$
- sample complexity  $n$  t.a.  $R_p(A(n)) - R_p(f^*) \leq \epsilon$  avec prob  $\geq 1 - \delta$

outils :

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

Uniforme  $\sup_{P \in \mathcal{P}} [R_p^F(A, n) - R_p(f^*)]$

Consistence  $\lim_{n \rightarrow \infty} [R_p^F(A, n) - R_p(f^*)] \rightarrow 0$

Consistence uniforme universelle  $\mathcal{P} = \text{tous les distributions}$

- impossible quand  $\mathcal{X}$  infini (no free lunch)
- possible si  $\mathcal{X}$  est fini

$R_p^F \rightarrow$  perte quadratique  $R_p^F(A_n) - R_p^F(f^*) = \underbrace{E_X \left[ \left( E_{D_n}[\hat{f}_n(x)|X] - f^*(x) \right)^2 \right]}_{\text{biais}} + \underbrace{E_X E_{D_n} \left( \hat{f}_n(x) - E_{D_n}[\hat{f}_n(x)|X] \right)^2}_{\text{variance}}$

décomposition biais-variance

- James-Stein
- CL Mallows design  $X$  fixe } régression
- consistance de alg. partition [plug-in]
  - $h_n \rightarrow 0$  } biais  $\rightarrow 0$
  - $n h_n^d \rightarrow +\infty$  } variance  $\rightarrow 0$
  - $n \approx \frac{1}{h_n^d}$  } plan de dimension

borne d'Occam  $\left[ \forall f R_p(f) \leq \hat{R}_n(f) + \frac{1}{\sqrt{n}} \sqrt{\ln 2 K(f) + \ln \frac{1}{\delta}} \right]$  avec prob.  $\geq 1-\delta$

$K(f)$  complexité(f)  $\rightarrow$  à priori

### Approches

• moyennage local (régression)  $\hat{f}(x) = \sum_{i=1}^n W_i(x) Y_i$

- histogramme (partition)  $W_i(x) \propto \mathbb{I}\{X_i \in A(x)\}$   $\sum_i W_i(x) = 1$


- K p.p.v.  $\mathbb{I}\{X_i \in V_K(x)\}$

$\hookrightarrow$  plug-in  $\hat{f}(x) = \mathbb{I}\{\hat{\pi}(x) \geq \frac{1}{2}\}$   $f^*(x) = E[Y|X=x]$

$Y_i \in \{0,1\}$

$E \left[ R_{D_n}^{0,1}(\hat{f}) \right] - R^{0,1}(f^*) \leq 2 \sqrt{E[R(\hat{\pi})] - K/n}$

$\rightarrow$  pour perte quadratique



min risque empirique régularisée

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

l quadratique  $\rightarrow$  régression ridge

l 0-1  $\rightarrow$  problème min 0-1 est NP-diff

$\hookrightarrow$  même pente convexe de substitut

class.  $f: X \rightarrow \{-1, 1\}$

$\tilde{f}: X \rightarrow \mathbb{R}$  (plug-in):  $f(x) \triangleq \text{sgn}(\tilde{f}(x))$

exemple: pour SVM

$\tilde{\ell} \rightarrow [1 - y_i \tilde{f}(x_i)]_+$

perte log  $\rightarrow$  maximum de vraisemblance

$$f_w(x) \rightarrow \log \frac{P_w(y=1|x)}{P_w(y=-1|x)}$$

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

$\hookrightarrow$  régression logistique

Optimisation & Analyse Convexe

classe de fonctions "tractables"

- ensembles convexes  $\hookrightarrow$  épigraphe
- fonctions convexes
- convexité forte
- ineq. Jensen etc

Optimisation sous contraintes:

min  $f(x)$

t.q.  $h_i(x) = 0$

$g_j(x) \leq 0$

problème dual  $\sup_{\lambda, \mu} q(\lambda, \mu)$

$$q(\lambda, \mu) = \inf_x \left[ f(x) + \sum \lambda_i h_i(x) + \sum \mu_j g_j(x) \right]$$

$L(x, \lambda, \mu)$  Lagrangien

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

conditions de KKT pour  $(x^*, \lambda^*, \mu^*)$

$x^*$  optimise  $L(x, \lambda^*, \mu^*)$

$\lambda^*, \mu^*$  sont faisables

$\mu_j^* g_j(x^*) = 0 \forall j$  "complémentarité"

algor. thms

- descente de gradient
- ellipsoïde
- Newton (régression log  $\rightarrow$  IRLS)

# \* sélection de modèle / hyperparamètre ( $\lambda$ )

- validation croisée

- $C_L$  de Mallows



$$p(x|y) = \sum_k \pi_k N(x | \mu_k, \Sigma_k)$$

## \* générative vs. conditionnelle

$p(x, y)$   
 ↓  
 LDA  
 & DA

$p(y|x)$   
 ↓  
 régression logistique

$$p(y|x) = \frac{1}{1 + \exp(-\tilde{f}(x))}$$

## \* astuce du noyau

$$\langle \tilde{\varphi}(x), \tilde{\varphi}(x') \rangle = K(x, x')$$

- thm. du représentant

- dualité
- etc.

→ régression ridge design  $X$

$$\underbrace{X^T X}_{d \times d} \rightarrow \underbrace{X X^T}_{n \times n} = K$$

→ SVM

(RBF  $K(x, x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$ )

$$\tilde{\varphi}(x) = \exp(-\frac{\|x-\mu\|^2}{2\sigma^2})$$

$$\varphi(x) \in \mathbb{R} \rightarrow \mathbb{R}$$

$$\langle \varphi(x), \varphi(x') \rangle$$

$$\Rightarrow \int \exp(-\frac{\|x-\mu\|^2}{2\sigma^2} - \frac{\|x'-\mu\|^2}{2\sigma^2}) dx$$

