Classification Supervisée

1 Modélisation

$(X,Y) \in \mathcal{X} \times \{-1,+1\}$ de loi $\mathbb{P}$.

point, label, mesuré

- **Classifieur** (prédit Y) $h : \mathcal{X} \to \{-1,+1\}$ mesurable.

performance $L(h) = \mathbb{P}(h(X) \neq Y)$ proba de mauvaise classification.

- **Meilleur?** comme $|h(X)-Y| \in \{0,2\}$

$$L(h) = \mathbb{E}\left[|h(X)-Y|^2\right] \overset{\text{Pyth.}}{=} \mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y|X]-h(X))^2\right]$$

minimum pour $h_*(X) = \operatorname{sign}(\mathbb{E}[Y|X])$ Classifieur de Bayes.
si $\mathbb{P}$ connu $\to$ ok.

2 Classification à partir d'un échantillon

- $\mathbb{P}$ inconnu : on a seulement accès à $(X_1,Y_1), \dots, (X_n,Y_n)$ iid de loi $\mathbb{P}$.

on remplace $\mathbb{P}$ par $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$

$$\begin{cases} \widehat{L}_n(h) := \widehat{\mathbb{P}}_n(Y \neq h(X)) = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i \neq h(X_i)} \\ \widehat{h}_H = \underset{h \in H}{\operatorname{argmin}}\ \widehat{L}_n(h) \end{cases}$$

où $\underset{\text{dictionnaire}}{\underbrace{H}} \subset \{-1,+1\}^{\mathcal{X}}$

Discuter : H = tout le monde $\to$ on reproduit les données mais mauvaise classif. d'un nouveau point

Quel H? discuter : H petit : trop rigide. cf exemple. H gros : trop flexible et colle trop aux données

$$0 \leq L(\widehat{h}_H) - L(h_*) \leq \underbrace{\min_{h \in H} L(h) - L(h_*)}_{\text{erreur d'approximation}} + \underbrace{L(\widehat{h}_H) - \min_{h \in H} L(h)}_{\text{erreur stochastique (échantillon)}}$$

$\searrow$ si $H \nearrow$ $\qquad$ $\nearrow$ si $H \nearrow$

Complexité de H : coefficient de shattering (éclatement)

$$\mathcal{S}_m(H) = \max_{(x_1,\dots,x_m)\in\mathcal{X}^m} \operatorname{Card}\left\{(h(x_1),\dots,h(x_m)) : h\in H\right\}$$

Ce qui compte ce n'est pas la taille de H mais sa capacité à produire des classifications différentes

Théorème 1

Avec proba $\geq 1-e^{-t}$ on a :

a) performance : $L(\hat{h}_H) - \min_{h\in H} L(h) \leq 4\sqrt{\frac{2\log(2\mathcal{S}_m(H))}{m}} + \sqrt{\frac{2t}{m}}$

b) Intervalle de confiance : $L(\hat{h}_H) \in \left[\hat{L}_m(\hat{h}_H) - \delta_m(t), \hat{L}_m(\hat{h}_H) + \delta_m(t)\right]$

avec $\delta_m(t) = 2\sqrt{\frac{2\log(2\mathcal{S}_m(H))}{m}} + \sqrt{\frac{2t}{m}}$

Corollaire

Soit $H_1,\dots,H_M$ une famille de M dictionnaires. Sélectionnons $H_{\hat{m}}$ par

$$\hat{m} = \underset{m=1,\dots,M}{\operatorname{argmin}} \left\{\hat{L}_m(H_m) + \underbrace{\operatorname{pen}(H_m)}_{\text{expliquer}}\right\} \text{ avec } \operatorname{pen}(H) = 2\sqrt{\frac{2\log(2\mathcal{S}_m(H))}{m}} + \sqrt{\frac{\log(M)+t}{2m}}$$

alors on a :

$$L(\hat{h}_{H_{\hat{m}}}) \leq \min_{m=1,\dots,M}\left\{L(\hat{h}_{H_m}) + 2\operatorname{pen}(H_m)\right\}.$$

Discuter :

- 1 H : fluctuations en $\sqrt{\frac{\log \mathcal{S}_m(H)}{m}}$
- M H : idem + $\underbrace{\sqrt{\frac{\log M}{m}}}_{\text{prix à payer pour choisir parmi } M}$

Preuve du Corollaire :

Théorème 1 b) : $|L(\hat{h}_{H_m}) - \hat{L}_m(\hat{h}_{H_m})| \leq \text{pen}(H_m)$

simultanément pour tout $m \in \{1, \dots, M\}$ avec proba $\geq 1 - e^{-t}$

D'où

$$L(\hat{h}_{H_{\hat{m}}}) \leq \hat{L}_m(\hat{h}_{H_{\hat{m}}}) + \text{pen}(H_{\hat{m}}) \qquad (\text{Thm } 1\text{b})$$

(remarque I.C.)
$$\leq \inf_{m=1,\dots,M} \left\{ \hat{L}_m(\hat{h}_{H_m}) + \text{pen}(H_m) \right\} \qquad (\text{définition de } \hat{m})$$

$$\leq \inf_{m=1,\dots,M} \left\{ L(\hat{h}_{H_m}) + 2\,\text{pen}(H_m) \right\} \qquad (\text{Thm } 1\text{b})$$

□

Preuve du Théorème :

Lemme 1 : Soit $\hat{\Delta}_m(H) = \sup_{h \in H} |\hat{L}_m(h) - L(h)|$

On a :

(ok)

$$L(\hat{h}_H) - \min_{h \in H} L(h) \leq 2\hat{\Delta}_m(H) \quad \text{et} \quad |L(\hat{h}_H) - \hat{L}_m(\hat{h}_H)| \leq \hat{\Delta}_m(H)$$

preuve L1 : $\forall h \in H$ on a

$$L(\hat{h}_H) - L(h) = L(\hat{h}_H) - \hat{L}_m(\hat{h}_H) + \underbrace{\hat{L}_m(\hat{h}_H)}_{\leq \hat{L}_m(h) \text{ (par définition)}} - L(h)$$

$$\leq 2\hat{\Delta}_m(H)$$

□

Il suffit de borner $\hat{\Delta}_m(H)$

Lemme 2 : avec proba $\geq 1 - e^{-t}$

$$\hat{\Delta}_m(H) \leq \mathbb{E}\left[\hat{\Delta}_m(H)\right] + \sqrt{\frac{t}{2m}}$$

preuve L2 :

$$F((x_1,y_1),\dots,(x_m,y_m)) = \frac{1}{m}\sup_{h\in H}\left|\sum_{i=1}^m \mathbb{1}_{y_i\neq h(x_i)} - L(h)\right| \quad \text{vérifie}$$

$$\left|F((x_1,y_1),\dots,(x_i,y_i),\dots,(x_m,y_m)) - F((x_1,y_1),\dots,(x_i',y_i'),\dots,(x_m,y_m))\right| \leqslant \frac{1}{m} \quad \text{donc}$$

l'inégalité de concentration de Mc Diarmid assure que

$$\mathbb{P}\Big[\underbrace{F((X_1,Y_1),\dots,(X_m,Y_m))}_{\hat{\Delta}_m(H)} \geqslant \underbrace{\mathbb{E}\big[F((X_1,Y_1),\dots,(X_m,Y_m))\big]}_{\mathbb{E}[\hat{\Delta}_m(H)]} + s\Big] \leqslant e^{-2ms^2}$$

Pour conclure : $s = \sqrt{\frac{t}{2m}}$ $\square$

Pour finir de prouver le théorème 1 il reste à montrer que

Lemme 3 :

$$\boxed{\mathbb{E}[\hat{\Delta}_m(H)] \leqslant 2\sqrt{\frac{2\log(2\, S_H(m))}{m}}}$$

2 étapes : (i) symétrisation
(ii) Markov.

(i)

$$\mathbb{E}[\hat{\Delta}_m(H)] = \mathbb{E}\left[\sup_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \mathbb{1}_{Y_i\neq h(X_i)} - \tilde{\mathbb{E}}\left[\frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\tilde{Y}_i\neq h(\tilde{X}_i)}\right]\right|\right]$$

même loi que X_i, Y_i mais indépendant

$$\underset{\text{Jensen+Fatou}}{\leqslant} \mathbb{E}\tilde{\mathbb{E}}\left[\sup_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \underbrace{\left(\mathbb{1}_{Y_i\neq h(X_i)} - \mathbb{1}_{\tilde{Y}_i\neq h(\tilde{X}_i)}\right)}_{\overset{\text{loi}}{=}\ \sigma_i\left(\mathbb{1}_{Y_i\neq h(X_i)} - \mathbb{1}_{\tilde{Y}_i\neq h(\tilde{X}_i)}\right)}\right|\right]$$

iid $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$

$$\underset{\text{In. Triangulaire}}{\leqslant} 2\,\mathbb{E}\tilde{\mathbb{E}}\mathbb{E}_\sigma\left[\sup_{h\in H}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i \mathbb{1}_{Y_i\neq h(X_i)}\right|\right]$$

$$\leqslant 2\max_{y\in\{-1,1\}^m}\ \max_{x\in X^m}\ \mathbb{E}_\sigma\left[\sup_{h\in H}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i \mathbb{1}_{y_i\neq h(x_i)}\right|\right]$$

Soit $\mathcal{V}_H(x,y) = \left\{ \left(\mathbb{1}_{y_1 \neq h(x_1)}, \dots, \mathbb{1}_{y_m \neq h(x_m)} \right),\ h \in H \right\}$

on a donc $\mathbb{E}\left[\hat{\Delta}_m(H)\right] \leqslant \frac{2}{m} \max_y \max_x \mathbb{E}_\sigma\left[\sup_{v \in \mathcal{V}_H(x,y)} |\langle \sigma, v\rangle| \right]$

Comme $\operatorname{Card} \mathcal{V}_H(x,y) \leq \operatorname{Card} \mathcal{S}_m(H)$ il reste à montrer que

(ii) $\mathbb{E}_\sigma\left[\sup_{v \in \mathcal{V}} |\langle \sigma, v\rangle| \right] \leqslant \sqrt{2m \log(2 \operatorname{card} \mathcal{V})} \quad \forall \mathcal{V} \subset \{-1, 0, +1\}^m.$

$$\mathbb{E}_\sigma\left[\sup_{v \in \mathcal{V}} |\langle \sigma, v\rangle| \right] \underset{\mathcal{V}^\# = \mathcal{V} \cup -\mathcal{V}}{=} \mathbb{E}_\sigma\left[\sup_{v \in \mathcal{V}^\#} \langle \sigma, v\rangle \right]$$

$$\underset{\text{Jensen}}{\leqslant} \frac{1}{s} \log \mathbb{E}_\sigma\left[\underbrace{\sup_{v \in \mathcal{V}^\#}}_{\leqslant \sum_{v \in \mathcal{V}^\#}} e^{s\langle \sigma, v\rangle} \right] \quad \forall s > 0$$

(car termes $\geqslant 0$)

$$\underset{\text{iid}}{\leqslant} \frac{\log \operatorname{Card} \mathcal{V}^\#}{s} + \frac{1}{s} \log \prod_{i=1}^m \underbrace{\mathbb{E}\left[e^{s \sigma_i v_i} \right]}_{= \frac{1}{2}\left(e^{s v_i} + e^{-s v_i}\right) \leqslant e^{s^2 v_i^2/2}}$$

$$\leqslant \frac{\log \operatorname{Card} \mathcal{V}^\#}{s} + \frac{m s}{2}$$

prendre $s = \sqrt{\frac{2 \log(\operatorname{Card} \mathcal{V}^\#)}{m}}$ pour conclure.

$\square$

Dimension de Vapnik-Chervonenkis

discuter : - Calcul $\mathcal{S}_m(H)$ complexe / borne sup

$$d_H := \sup\{d \in \mathbb{N} : \mathcal{S}_d(H) = \underbrace{2^d}_{\text{maxi}}\} \in \mathbb{N} \cup \{+\infty\}$$

Lemme de Sauer :

Soit H de dimension de Vapnik $d_H < +\infty$. On a

$$\mathcal{S}_m(H) \underset{(i)}{\leq} \sum_{i=0}^{d_H} C_m^i \underset{(ii)}{\leq} (m+1)^{d_H}$$

Discuter : - $\mathcal{S}_m(H) \approx (m+1)^{d_H}$ dans IC + pen.

- ex : classifieurs affines : $d_H = d+1$

- preuve :

(i) - cas $d_H = 0$: $\mathcal{S}_m(H) = 1$ donc ok (Tout point ne peut être labellé que d'une seule manière)

On suppose $d_H \geq 1$.

- $n=1$: $\mathcal{S}_1(H) = 2 \leq C_1^0 + C_1^1$

- induction : vrai jusqu'à $m-1$

$$H(x_1, \dots, x_m) := \{(h(x_1), \dots, h(x_m)) : h \in H\}$$
$$= \tilde{H}(x_1, \dots, x_m) \text{ pour } \tilde{H} = \{h|_{\{x_1, \dots, x_m\}} : h \in H\}.$$

→ on peut remplacer H par $\tilde{H}$ pour majorer le card$(H(x_1, \dots, x_m))$

Soit $\tilde{H}' = \{h \in \tilde{H} : h(x_m) = 1 \text{ et } h - 2\mathbb{1}_{x_m} \in \tilde{H}\}$

$$\text{card}(H(x_1, \dots, x_m)) \leq \underbrace{\text{card}(\tilde{H}'(x_1, \dots, x_m))}_{1)} + \underbrace{\text{card}(\tilde{H} \setminus \tilde{H}'(x_1, \dots, x_m))}_{2)}$$

1) - $\mathrm{card}(\tilde{H}'(x_1,\dots,x_m)) = \mathrm{card}(\tilde{H}'(x_1,\dots,x_{m-1}))$ car $h(x_m)=1 \ \forall h \in \tilde{H}'$

- soit $1 \leq i_1 < \dots < i_d \leq m-1$:

Si $\mathrm{card}(\tilde{H}'(x_{i_1},\dots,x_{i_d})) = 2^d$ alors

$\mathrm{card}(\tilde{H}(x_{i_1},\dots,x_{i_d},x_m)) = 2^{d+1}$ par déf. de $\tilde{H}'$

$\Rightarrow d_{\tilde{H}'} \leq d_H - 1$

et par récurrence $\mathrm{card}(\tilde{H}'(x_1,\dots,x_m)) \leq \sum_{i=0}^{d_H-1} C_{m-1}^i$

2) - $\mathrm{card}(\tilde{H}\setminus\tilde{H}'(x_1,\dots,x_m)) = \mathrm{card}(\tilde{H}\setminus\tilde{H}'(x_1,\dots,x_{m-1}))$

car si $h,h' \in \tilde{H}\setminus\tilde{H}'$ vérifient $h(x_i)=h'(x_i)$ $\forall i=1,\dots,m$ alors $h(x_m)=h'(x_m)$

$\overset{(\text{rec})}{\Longrightarrow} \mathrm{card}(\tilde{H}'(x_1,\dots,x_m)) \leq \sum_{i=0}^{d_H} C_{m-1}^i$

donc $\mathrm{card}(H(x_1,\dots,x_m)) \leq \sum_{i=1}^{d_H} (C_{m-1}^i + C_{m-1}^{i-1}) + C_{m-1}^0 = \sum_{i=0}^{d_H} C_m^i$

(ii) $\sum_{i=0}^{d} C_m^i \leq \sum_{i=0}^{d} \frac{m^d}{i!} \leq \sum_{i=0}^{d} C_d^i m^i = (1+m)^d.$ $\square$

Remarque : On a mieux : $\sum_{i=0}^{d} C_m^i \leq \left(\frac{em}{d}\right)^d$ (exo)