# Apprentissage statistique

Chapitre 6 : Pertes générales, chaining

Lucie Le Briquer

6 mars 2018

## Table des matières

# 1 Pertes générales

- Les données sont $X \in \mathcal{X}$ (le plus souvent $\subseteq \mathbb{R}^d$)
- Les étiquettes $Y \in [-1, 1]$
- On dispose de la famille $\mathcal{F}$ de classifieurs $\subseteq \{x \in \mathcal{X} \longrightarrow [-1, 1]\}$
- Perte (générale) : $l(f(x), y) \in [0, 1]$

Le risque de $f \in \mathcal{F}$ est :

$$R(f) = \mathbb{E}[l(f(X), Y)]$$

**Exemple.** (classiques de pertes) $l_p(y, y') = |y - y'|^p$
À partir des données $(X_1, Y_1), \ldots, (X_n, Y_n)$ on définit le risque empirique d'un classifieur par :

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i)$$

Le minimiseur du risque empirique est $\hat{f}_n = \text{argmin } \hat{R}_n(f)$ quand il est défini (ce qui sera toujours le cas désormais).

**Lemme 1**

$$R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) \leqslant 2 \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|$$

et,

$$\mathbb{P}\left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geqslant \mathbb{E}\left[ \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right] + \varepsilon \right) \leqslant e^{-2n\varepsilon^2}$$

**Preuve.**
Ce n'est *pas* Hoeffding puisqu'on aurait :

$$\mathbb{P}\big( |\hat{R}_n(f) - R(f)| - \mathbb{E}[\hat{R}_n(f) - R(f)] \geqslant \varepsilon \big) \leqslant e^{-2n\varepsilon^2}$$

Et donc on aurait seulement $\leqslant |\mathcal{F}| e^{-2n\varepsilon^2}$.
Notons $Z_n = \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|$. On veut montrer que $\mathbb{P}(Z_n - E[Z_n] \geqslant \varepsilon) \leqslant e^{-2\varepsilon^2 n}$.

$$Z_n = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i) - \mathbb{E}[l(f(X), Y)] \right|$$

Si on change la valeur d'un seul $(X_i, Y_i)$ en maintenant les autres égaux, alors la valeur de $Z_n$ change d'au plus $\frac{1}{n}$. Inégalité de McDiarmid (ou des différences bornées) :

$$\mathbb{P}(Z_n - \mathbb{E}[Z_n] \geqslant \varepsilon) \leqslant e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n} \frac{1}{n^2}}} = e^{-2\varepsilon^2 n}$$

□

**Exemple.** (intéressant d'utilisation de McDarmid) Concentration de la variance empirique. Soient $X_1, \dots, X_n$ i.i.d. de variance $\sigma^2$.

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{2}{n(n-1)}\sum_{i<j}(X_i - X_j)^2$$

McDiarmid donne :

$$\mathbb{P}(\hat{\sigma}_n^2 - \sigma^2 \geqslant \varepsilon) \leqslant e^{-\frac{n\varepsilon^2}{2}}$$

Pour en revenir au learning, il suffit de contrôler $\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|$. En utilisant la symétrisation, on va simplifier son écriture.

**Lemme 2**

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|\right] \leqslant 2\sup_{\mathcal{D}_n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(f(x_i), y_i)\right|\right]$$

où les $\sigma_i = \pm 1$ avec probabilité $\frac{1}{2}$, i.i.d. et indépendantes des $(X_i, Y_i)$.

**Preuve.** (idées)
On introduit un "échantillon fantôme" $(X_i', Y_i')$ de même loi de $(X_i, Y_i)$ et indépendant. On remplace $R(f)$ par $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} l(f(X_i'), Y_i')\right]$. La fin de la preuve est presque identique au cas binaire. □

**Définition 1** (complexité de Rademacher)
La complexité de Rademacher pour la famille $\mathcal{F}$ au point $\mathcal{D}_n$ est :

$$\mathcal{R}[l\circ\mathcal{F}, \mathcal{D}_n] = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(f(x_i), y_i)\right|\right]$$

et on notera

$$\mathcal{R}_n[l\circ\mathcal{F}] = \sup_{\mathcal{D}_n}\mathcal{R}[l\circ\mathcal{F}, \mathcal{D}_n]$$

**Lemme 3**
Si $\mathcal{F}$ est finie, alors par Hoeffding :

$$\mathcal{R}_n[l\circ\mathcal{F}] \leqslant \sqrt{\frac{2\log(|\mathcal{F}|)}{n}}$$

## 1.1 Cas général, $\mathcal{F}$ infinie

On veut $\varepsilon-$approcher les fonctions croissantes à valeurs dans $[0,1]$. $\forall f$ croissante $\exists f_\varepsilon \in \mathcal{F}^\varepsilon$ $\forall i,\ |f(x_i) - f_\varepsilon(x_i)| \leqslant \varepsilon$. Il y a $\frac{1}{\varepsilon}$ sauts (de hauteur $\varepsilon$) à "placer" sur un des $n$ points. Donc au plus $n^{1/\varepsilon}$.

Par ailleurs si $l$ est $L-$lipschitz,

$$\mathcal{R}_n[l \circ \mathcal{F}] \leqslant \varepsilon + \mathcal{R}_n[l \circ \mathcal{F}^\varepsilon]$$

Donc dans le cas des fonctions croissantes on obtiendrait :

$$\mathcal{R}_n[l \circ \mathcal{F}] \leqslant \varepsilon + 2\sqrt{\frac{\log(n)}{\varepsilon n}} \leqslant 3\left(\frac{\log n}{n}\right)^{\frac{1}{3}}$$

Pour simplifier les notations, on va considérer la compexité de Rademacher :

$$\mathcal{R}_n[\mathcal{F}, \mathcal{D}_n] = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

Pour une classe $\mathcal{F}$ de fonctions $\mathcal{X} \longrightarrow [0,1]$.

**Définition 2** (covering number)

Étant donnés une classe $\mathcal{F} \subset \{\mathcal{X} \longrightarrow [0,1]\}$, une métrique $d$ sur $\{\mathcal{X} \longrightarrow [0,1]\}$ et $\varepsilon > 0$. Un $\varepsilon-$net de $(\mathcal{F}, d)$ est un ensemble $V$ tel que pour tout $f \in \mathcal{F}$, il existe $g \in V$ tel que $d(f,g) \leqslant \varepsilon$ (les éléments de $V$ n'appartiennent pas forcément à $\mathcal{F}$).

Les covering numbers de $(\mathcal{F}, d)$ sont :

$$\mathcal{N}(\mathcal{F}, d, \varepsilon) = \inf\{|V| \ ; \ V \text{ est un } \varepsilon - \text{net de } (\mathcal{F}, d)\}$$

En particulier pour les distances $d_p^x$ définies par :

$$d_p^x(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - g(x_i)|^p\right)^{1/p}$$

On notera

$$\sup_{x_1,\ldots,x_n} \mathcal{N}(\mathcal{F}, d_p^x, \varepsilon) = \mathcal{N}_n^p(\mathcal{F}, \varepsilon)$$

**Théorème 1** (discrétisation)

$$\mathcal{R}_n[\mathcal{F}, \mathcal{D}_n] \leqslant \inf_{\varepsilon>0}\varepsilon + \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^x, \varepsilon))}{n}}$$

et

$$\mathcal{R}_n[\mathcal{F}] \leqslant \inf_{\varepsilon>0}\varepsilon + \sqrt{\frac{2\log(2\mathcal{N}_n^1(\mathcal{F}, \varepsilon))}{n}}$$

**Exemples.**

- Si $\mathcal{F}$ est l'ensemble des fonctions croissantes de $\mathbb{R}$ dans $[0,1]$,

$$\mathcal{R}_n[\mathcal{F}] \leqslant 3\left(\frac{\log(n)}{n}\right)^{1/3}$$

- Si $\mathcal{F}$ est l'ensemble des fonctions linéaires $\theta^T x$ avec $\mathrm{c}\theta \in \mathcal{B}_p$ (sur $\mathcal{X} = \mathcal{B}_q$), alors :

$$\mathcal{N}(\mathcal{F}, \varepsilon) \leqslant \left(\frac{2}{\varepsilon}\right)^d \quad \text{et} \quad \mathcal{R}_n[\mathcal{F}] \leqslant 2\sqrt{\frac{d\log(16n/d)}{n}}$$

# 2 Chaining

**Théorème 2**

Si $f \in \mathcal{F}$ sont à valeurs dans $[0,1]$, alors, pour tout $x = (x_1, \ldots, x_n)$,

$$\mathcal{R}_n[\mathcal{F}, x] \leqslant 4 \inf_{\varepsilon > 0} \left\{ \varepsilon + 3 \int_\varepsilon^1 \sqrt{\frac{\log(2\mathcal{N}(\mathcal{F}, d_1^x, \varepsilon'))}{n}} d\varepsilon' \right\}$$

**Preuve.**

Fixons $x = (x_1, \ldots, x_n)$ de telle sorte que $f$ peut être identifiée par $(f(x_1), \ldots, f(x_n))$.

$$\mathcal{R}_n[\mathcal{F}, x] \leqslant \mathbb{E}_\sigma \sup_f |\sigma^T f| \qquad \text{où } \sigma = \left(\frac{\sigma_1}{n}, \ldots, \frac{\sigma_n}{n}\right)$$

On note $V_m$ un $2^{-m}-$net de $\mathcal{F}$ qui réalise $\mathcal{N}(\mathcal{F}, d_1^x, 2^{-m})$ et $f_m$ un élément de $V_m$ qui est à $2^{-m}$ de $f$. On écrit tout simplement :

$$f = (f_M - f_{M-1}) + (f_{M-1} - f_{M-2}) + \ldots + (f_1 - 0) + \underbrace{(f - f_M)}_{\leqslant 2^{-M}}$$

Donc

$$\begin{aligned}
\mathcal{R}_n[\mathcal{F}, x] &\leqslant \mathbb{E}_\sigma \sup_f \left|\sigma^T \left[(f - f_M) + (f_M - f_{M-1}) + \ldots + (f_1 - 0)\right]\right| \\
&\leqslant 2^{-M} + \sum_{m=1}^{M} \mathbb{E} \sup_{f_m, f_{m-1}} |\sigma^T (f_m - f_{m-1})|
\end{aligned}$$

Il reste à contrôler pour tout $m \leqslant M$ :

$$\mathbb{E} \sup_{f_m, f_{m-1}} |\sigma^T (f_m - f_{m-1})|$$

car par Hoeffding :

$$\mathbb{P}(\sigma^T g \geqslant \varepsilon) \leqslant e^{-\frac{2n^2\varepsilon^2}{\sum g_i^2}} = e^{-\frac{2n^2\varepsilon^2}{\|g\|^2}}$$

Donc,

$$\mathbb{P}\left(\sup_{g \in G} \sigma^T g \geqslant \varepsilon\right) \leqslant |G| e^{-\frac{2n^2\varepsilon^2}{\max \|g\|^2}}$$

Ainsi, ici :

$$\begin{aligned}
\mathbb{E} \sup_{f_m, f_{m-1}} |\sigma^T (f_m - f_{m-1})| &\leqslant \sup_f \|f_m - f_{m-1}\|_2 \frac{\sqrt{2\log(2|V_m||V_{m-1}|)}}{n} \\
&\leqslant \sup_f \|f_m - f_{m-1}\|_2 \frac{\sqrt{2\log(2|V_m|^2)}}{n}
\end{aligned}$$

mais,

$$\begin{aligned}\|f_m - f_{m-1}\|_2 &\leqslant \|f_m - f\|_2 + \|f - f_{m-1}\|_2 \\ &\leqslant \sqrt{n}(\|f_m - f\|_1 + \|f - f_{m-1}\|_1) \\ &\leqslant \sqrt{n}(2^{-m} + 2^{-m+1}) = 3\sqrt{n}2^{-m}\end{aligned}$$

D'où :

$$\mathcal{R}_n[\mathcal{F}, x] \leqslant 2^{-M} + 3\sum_{m=1}^{M} 2^{-m}\sqrt{\frac{2\log(2|V_m|^2)}{n}}$$

Avec $V_m = \mathcal{N}(\mathcal{F}, d_1^x, 2^{-m})$. En choisissant $M$ tel que $2^{-M-2} \leqslant \varepsilon 2^{-M-1}$, on obtient :

$$\mathcal{R}_n[\mathcal{F}, x] \leqslant 4\varepsilon + 12\int_{\varepsilon}^{1} \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^x, \varepsilon'))}{n}} d\varepsilon'$$

car $\sum_{m=1}^{M} 2^{-m}\sqrt{\dots} = 2\sum_{m=1}^{M}\left(2^{-m} - 2^{-(m+1)}\right)\sqrt{\dots}$ □

**Exemples.** Si $\mathcal{N}(\mathcal{F}, d^x, \varepsilon \approx n^{1/\varepsilon})$ alors :

$$\int_{\varepsilon}^{1} \sqrt{\frac{\log(2n^{1/\varepsilon'})}{n}} d\varepsilon' \leqslant \sqrt{2}\int_{\varepsilon}^{1} \sqrt{\frac{\log n}{n}\frac{1}{\varepsilon'}} d\varepsilon' \leqslant 2\sqrt{2}\sqrt{\frac{\log n}{n}} \quad \forall \varepsilon$$

Donc $\mathcal{R}_n[\mathcal{F}, x] \leqslant 36\sqrt{\frac{\log n}{n}}$.

# 3 Calculs de covering numbers

On veut *majorer* les coverings numbers.

*Technique 1.* On peut relier covering et packing numbers.

$$\mathcal{N}(\mathcal{F}, d, \varepsilon) = \text{ nombre de boules minimal pour recouvrir } \mathcal{F}$$

$$\mathcal{M}(\mathcal{F}, d, \varepsilon) = \text{ nombre maximal de points } f_i \text{ dans } \mathcal{F} \text{ tel que } |f_i - f_j| > \varepsilon$$

**Proposition 1**

$\forall \varepsilon, \forall \mathcal{F}, \ \forall d$ :

$$\mathcal{M}(\mathcal{F}, d, 2\varepsilon) \leqslant \mathcal{N}(\mathcal{F}, d, \varepsilon) \leqslant \mathcal{M}(\mathcal{F}, d, \varepsilon)$$

**Preuve.**

Pour la seconde inégalité, on remarque qu'un ensemble qui réalise $\mathcal{M}(\mathcal{F}, d, \varepsilon)$ est un $\varepsilon-$covering. Pour la première inégalité, considérons la contraposée. Supposons $\mathcal{M} \geqslant \mathcal{N} + 1$. Alors il existe une boule de taille $\varepsilon$ qui contient deux points du packing optimal. Donc ces deux points sont à une distance de plus de $2\varepsilon$. Contradiction. □

*Technique 2.* Utiliser le volume.

**Proposition 2**

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{Vol}(\mathcal{F})}{\mathrm{Vol}(\mathcal{B})} \leqslant \mathcal{N}(\mathcal{F}, d, \varepsilon) \leqslant \mathcal{M}(\mathcal{F}, d, \varepsilon) \leqslant \left(\frac{3}{\varepsilon}\right)^d \frac{\mathrm{Vol}(\mathcal{F})}{\mathrm{Vol}(\mathcal{B})}$$

si $\mathcal{F} \subset \mathbb{R}^d$ et est convexe, $\mathcal{B}$ est la boule unité.

**Preuve.**

$$\mathcal{F} \subset \bigcup_{f_i \in \mathcal{F}} \mathcal{B}(f_i, \varepsilon)$$

Donc,

$$\mathrm{Vol}(\mathcal{F}) \leqslant \mathrm{Vol}\left(\bigcup_{f_i \in \mathcal{F}} \mathcal{B}(f_i, \varepsilon)\right) \leqslant \sum_{f_i \in \mathcal{F}} \mathrm{Vol}(\mathcal{B}(f_i, \varepsilon)) = \mathcal{N}\mathrm{Vol}(\mathcal{B})\varepsilon^d$$

D'un autre côté,

$$\mathcal{M} \times \mathrm{Vol}(\mathcal{B}) \left(\frac{\varepsilon}{2}\right)^d \leqslant \mathrm{Vol}\left(\mathcal{F} + \frac{\varepsilon}{2}\mathcal{B}\right) \leqslant \mathrm{Vol}\left(\frac{3}{2}\mathcal{F}\right) = \left(\frac{3}{2}\right)^2 \mathrm{Vol}(\mathcal{F})$$

en supposant que $\mathrm{Vol}(\mathcal{F}) \geqslant \mathrm{Vol}(\varepsilon\mathcal{B})$. □

*Technique 2.* "Fat-shattering", généralise la VC-dimension.

**Définition 3**

Étant donnés $x_1, \ldots, x_n$, on dit que $\mathcal{F}$ éclate $x_1, \ldots, x_n$ au niveau $\alpha$, s'il existe des seuils $s_1, \ldots, s_n$ tels que $\forall E \subset \{1, \ldots, n\}$, il existe $f_E \in \mathcal{F}$ :

$$f_E(x_i) \geqslant s_i + \alpha \quad \text{si } i \in E$$

$$f_E(x_i) \leqslant s_i - \alpha \quad \text{si } i \notin E$$

$\mathrm{fat}_\alpha(\mathcal{F})$ est la plus petite cardinalité d'un ensemble pouvant être éclaté au niveau $\alpha$.

**Proposition 3**

$$\log(\mathcal{N}(\mathcal{F}, d_\infty, \alpha)) \leqslant \mathrm{fat}_{\frac{\alpha}{4}} \log\left(\frac{2en}{d\alpha}\right) \log\left(2n\left(1 + \frac{2}{\alpha}\right)^2\right) \approx \mathrm{fat}_{\frac{\alpha}{4}}(\mathcal{F}) \log^2\left(\frac{n}{d\alpha}\right)$$