Apprentissage statistique

Chapitre 5 : L'Empirical Risk Minimization (ERM)

Lucie Le Briquer

13 février 2018

Table des matières

1 Minimisation dans une classe finie

On se donne une famille de fonctions (classifieurs) $\mathcal{F} = \{f : \mathcal{X} \longrightarrow \{0,1\}\}$ (le plus souvent $\mathcal{X} \subset \mathbb{R}^d$) et le minimiseur du risque empirique dans $\mathcal{F}$ est le classifieur $\hat{f}$ défini par :

$$\hat{f}(\cdot) = \hat{f}(D_n, \cdot) = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \hat{R}_n(f)$$

où $D_n = \{(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \{0,1\}\}$ sont les données, et $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(x_i) \neq y_i}$ le nombre moyen d'erreur du classifieur.

Théorème 1

L'errreur "d'estimation" de l'ERM est contrôlée par :

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

$\int_{f \in \mathcal{F}} R(f)$ correspond à l'erreur d'approximation. Ainsi :

$$\mathbb{P}\left(R(\hat{f}) \geqslant \inf_{f \in \mathcal{F}} R(f) + \varepsilon\right) \leqslant 2|\mathcal{F}|e^{-\frac{n\varepsilon^2}{2}} \quad \text{et} \quad \mathbb{E}[R(\hat{f})] \leqslant \inf_{f \in \mathcal{F}} R(f) + \frac{3}{2}\sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}$$

Ici on a $R(f) = \mathbb{E}[\mathbb{1}_{f(X) \neq Y}] = \mathbb{P}(f(X) \neq Y)$.

Preuve.

1. Notons $f^* = \operatorname{argmin} R(f)$.

$$\begin{aligned} R(\hat{f}) - R(f^*) &= R(\hat{f}) - \hat{R}_n(\hat{f}) + \underbrace{\hat{R}_n(\hat{f}) - \hat{R}_n(f^*)}_{\leqslant 0} + R_n(f^*) - R(f^*) \\ &\leqslant 2 \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \end{aligned}$$

   Si l'argmin n'existe pas, on prend $R(f) + \varepsilon$ puis idem.

2. Pour (2) on veut contrôler :

$$\begin{aligned} \mathbb{P}\left(2 \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geqslant \varepsilon\right) &\leqslant \sum_{f \in \mathcal{F}} \mathbb{P}\left(|\hat{R}_n(f) - R(f)| \geqslant \frac{\varepsilon}{2}\right) \\ &\leqslant \sum_{f \in \mathcal{F}} 2e^{-2n\left(\frac{\varepsilon}{2}\right)^2} \\ &\leqslant 2|\mathcal{F}|e^{-\frac{n\varepsilon^2}{2}} \end{aligned}$$

3. On va utiliser le fait que si $X \geqslant 0$:

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}(x \geqslant \varepsilon) d\varepsilon$$

En particulier,

$$\begin{aligned}\mathbb{E}[R(\hat{f}) - \inf_{f\in F} R(f)] &\leqslant \int_{\varepsilon^*}^{+\infty} 2|\mathcal{F}|e^{-\frac{n\varepsilon^2}{2}}d\varepsilon + \varepsilon^* \\ &= \varepsilon + \int_{\varepsilon^*}^{+\infty} 2|\mathcal{F}|\frac{\varepsilon}{\varepsilon}e^{-\frac{n\varepsilon^2}{2}}d\varepsilon \\ &\leqslant \varepsilon + \int_{\varepsilon^*}^{+\infty} 2|\mathcal{F}|\frac{\varepsilon}{\varepsilon^*}e^{-\frac{n\varepsilon^2}{2}}d\varepsilon \\ &= \varepsilon^* + \frac{2|\mathcal{F}|}{\varepsilon^* n}e^{-n(\varepsilon^*)^2/2}\end{aligned}$$

où ε^* est tel que $2|\mathcal{F}|e^{-n(\varepsilon^*)/2} = 1$. Donc :

$$\mathbb{E}[R(\hat{f}) - \inf_{f\in F} R(f)] \leqslant \varepsilon^* + \frac{2}{\varepsilon^* n} = \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}} + \sqrt{\frac{2}{n\log(2|\mathcal{F}|)}}$$

□

Remarque. Soit $\mathcal{F}$ la famille des classifieurs affines de $\mathbb{R}^d$. $\mathcal{F} = \{f \mid f(x) = \mathbb{1}_{a^T x+b\geqslant 0}\}$.

$$\begin{array}{ccc} & \overset{0}{\times} & \\ 1\times & & \times 1 \\ & \underset{0}{\times} & \end{array}$$

Si on fixe un nombre de points n, on peut considérer uniquement les représentants dans la borne union de la preuve. De prime abord on peut penser qu'il y a 2^n, et on ne gagne rien, mais certaines combinaisons ne peuvent pas être classifiées (cf schéma) et on se ramène à n^p représentants.

2 La dimension de Vapnik-Chervonenkis

Définition 1 (coefficient de pulvérisation)

Le coefficient de pulvérisation de $\mathcal{F}$ est :

$$S(\mathcal{F}, n) = \max_{x_1,\ldots,x_n} |\{(f(x_1), \ldots, f(x_n)),\ f \in \mathcal{F}\}|$$

Il est vrai que $S(F, n) \leqslant 2^n$, mais il est bien souvent beaucoup plus petit (polynomial).

Définition 2 (VC-dimension)

La VC-dimension de $\mathcal{F}$ est le plus grand K tel que :

$$S(\mathcal{F}, k) = 2^k$$

Remarque. Dans l'exemple des classifieurs affines, pour 3 points on a 2^3 représentants, pour 4 points $14 \neq 2^4$. La VC-dimension de l'ensemble des classifieurs affines est donc 3.

Lemme 1 (Sauer-Shelah)

Soit d la VC-dimension de $\mathcal{F}$. Alors :

$$S(\mathcal{F}, n) \leqslant \begin{cases} 2^n & \text{si } n \leqslant d \\ \left(\frac{en}{d}\right)^d & \text{si } n > d \end{cases}$$

Preuve.

Récurrence sur $d+n$ ou d est la VC-dimension et n le nombre de points. Si $d = 0$ ok, $n = 0$ ok.

- On va montrer que $S(\mathcal{F}, n) \leqslant \sum_{k=0}^{d} \binom{n}{k}$.

  Soit d la VC-dimension de $\mathcal{F}$ et $\{x_1, \ldots, x_n\}$ n points qui réalise le maximum de la définition de $S(\mathcal{F}, n)$ et $A \subset \mathcal{F}$ une sous-famille de classifieurs de taille minimale qui permet d'obtenir tous les labellings. On va considérer $D' = \{x_2, \ldots, x_n\}$ et $A' \subset A$ le plus petit ensemble qui maximise le nombre de labellisations de D'.

  $$|A| = \mathcal{S}(\mathcal{F}, n) \qquad \text{et} \qquad |A| = |A'| + |A \backslash A'|$$

- La VC-dimension de A', d', est plus petite que que d. Par minimalité de A' on a :

  $$|A'| \leqslant \mathcal{S}(A', n-1) \leqslant \sum_{k=0}^{d'} \binom{n-1}{k} \leqslant \sum_{k=0}^{d} \binom{n-1}{k}$$

- Si $A \backslash A'$ pulvérise $E \subset D$, alors A pulvérise $E \cup \{x_1\}$. Sinon un élément de $A \backslash A'$ serait dans A' car pour tout $f' \in A \backslash A$, il existe $f' \in A'$ qui coïncide avec f sur D', mais par minimalité de A, f et f' diffèrent sur x_1. On doit alors avoir $|E| + 1 \leqslant d$ donc $VC - A \backslash A') \leqslant d - 1$. D'où :

  $$|A \backslash A'| \leqslant \sum_{k=0}^{d-1} \binom{n-1}{k}$$

Ainsi :

$$S(\mathcal{F}, n) \leqslant \sum_{k=0}^{d} \binom{n-1}{k} + \sum_{k=0}^{d-1} \binom{n-1}{k} = \sum_{k=0}^{d-1} \binom{n-1}{k+1} + \binom{n-1}{k} + 1 \leqslant \sum_{k=0}^{d} \binom{n}{k} \leqslant \left(\frac{en}{d}\right)^d$$

□

3 Inégalité de VC

Théorème 2

Pour tout famille $\mathcal{F}$ et $n \in \mathbb{N}$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| \geqslant \varepsilon\right) \leqslant 2S(\mathcal{F}, n)e^{-\frac{n\varepsilon^2}{32}}$$

Ainsi,

$$\mathbb{E}[R(\hat{f}_n)] \leqslant \inf_{f\in\mathcal{F}} R(f) + 4\sqrt{\frac{d\log\left(\frac{en}{d}\right) + \log(2)}{n}}$$

Preuve.

On suppose que $n\varepsilon^2 \geqslant 2$.

Technique 1 (symmétrisation) Soit $\{(X_1', Y_1'), \ldots, (X_n', Y_n')\} = D_n'$ un échantillon "fantôme" de même loi que D_n mais indépendant. Alors :

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f) \geqslant \varepsilon|\right) \leqslant 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - \hat{R}_n'(f)| \geqslant \frac{\varepsilon}{2}\right)$$

où $\hat{R}_n'(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(X_i')\neq Y_i'}$. Soit $\tilde{f}$ un classifieur de $\mathcal{F}$ (qui peut dépendre de D_n) tel que $|\hat{R}_n(\tilde{f}) - R_n(\tilde{f})| \geqslant \varepsilon$ si c'est possible (sinon $\tilde{f}$ vaut n'importe quoi), de sorte que

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| \geqslant \varepsilon\right) = \mathbb{P}(|\hat{R}_n(\tilde{f}) - R(\tilde{f})| \geqslant \varepsilon)$$

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - \hat{R}_n'(f)| \geqslant \frac{\varepsilon}{2}\right) &\geqslant \mathbb{P}\left(|\hat{R}_n(\tilde{f}) - \hat{R}_n'(\tilde{f})| \geqslant \frac{\varepsilon}{2}\right) \\
&\geqslant \mathbb{P}\left(|\hat{R}_n(\tilde{f}) - R(\tilde{f})| \geqslant \varepsilon \text{ et } |R(\tilde{f}) - \hat{R}_n'(\tilde{f})| \leqslant \frac{\varepsilon}{2}\right) \\
&= \mathbb{E}\left[\mathbb{1}_{|\hat{R}_n(\tilde{f}) - R(\tilde{f})|\geqslant\varepsilon} \times \mathbb{E}\left[\mathbb{1}_{|R(\tilde{f}) - \hat{R}_n'(\tilde{f})|\leqslant\frac{\varepsilon}{2}} | D_n\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{|\hat{R}_n(\tilde{f}) - R(\tilde{f})|\geqslant\varepsilon} \times \mathbb{P}\left(|\hat{R}_n'(\tilde{f}) - R(\tilde{f})| \leqslant \frac{\varepsilon}{2} | D_n\right)\right]
\end{aligned}$$

Une fois conditionné à D_n, $\tilde{f}$ est une fonction donnée. Donc par Hoeffding :

$$\mathbb{P}\left(|\hat{R}_n'(\tilde{f}) - R(\tilde{f})| \leqslant \frac{\varepsilon}{2} | D_n\right) \geqslant 1 - 2e^{-\frac{n\varepsilon^2}{2}} \geqslant \frac{1}{2}$$

Donc on veut contrôler $\mathbb{P}\left(|\hat{R}_n'(\tilde{f}) - R(\tilde{f})| \leqslant \frac{\varepsilon}{2} | D_n\right)$ mais :

$$\hat{R}_n(f) - \hat{R}_n'(f) = \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{1}_{f(X_i')\neq Y_i'}}_{\in\{-1,0,1\}}$$

Donc si on introduit des variables $\sigma_i = \pm 1$ avec probabilité $\frac{1}{2}$, la loi de $\hat{R}_n(f) - \hat{R}'_n(f)$ est la même que la loi de :

$$\frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{1}_{f(X'_i)\neq Y'_i} \right)$$

Donc

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\dots| \geqslant \frac{\varepsilon}{2} \right) = \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{1}_{f(X'_i)\neq Y'_i} \right) \right| \geqslant \frac{\varepsilon}{2} \right)$$

Mais,

$$\begin{aligned}
& \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X_i)\neq Y_i} - \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X'_i)\neq Y'_i} \right| \geqslant \frac{\varepsilon}{2} \right) \\
\leqslant\ & \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X_i)\neq Y_i} \right| \geqslant \frac{\varepsilon}{4} \text{ ou } \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X'_i)\neq Y'_i} \right| \geqslant \frac{\varepsilon}{4} \right) \\
\leqslant\ & 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X_i)\neq Y_i} \right| \geqslant \frac{\varepsilon}{4} \right)
\end{aligned}$$

Mais par définition de $S(\mathcal{F}, n)$, il n'existe qu'au plus $S(\mathcal{F}, n)$ valeurs distinctes des $\left(\mathbb{1}_{f(x_i)\neq y_i}\right)_{i=1\dots n}$ quand les x_i et y_i sont fixés.

$$\begin{aligned}
& \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X_i)\neq Y_i} \right| \geqslant \frac{\varepsilon}{4} \right) \\
=\ & \mathbb{E}\left[\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{f(X_i)\neq Y_i} \right| \geqslant \frac{\varepsilon}{4} \mid (X_i, Y_i) = (x_i, y_i) \right) \right] \\
\leqslant\ & \mathbb{E}\left[|S(\mathcal{F}, n)| 2e^{-\frac{2n}{4}\left(\frac{\varepsilon}{4}\right)^2} \right] \\
=\ & 2|S(\mathcal{F}, n)| e^{-\frac{n\varepsilon^2}{32}} \\
\leqslant\ & 2\left(\frac{en}{d}\right)^d e^{-\frac{n\varepsilon^2}{32}}
\end{aligned}$$

□