Apprentissage: cours 6 (2ème partie)
Méthodes par minimisation du risque empirique

Sylvain Arlot

22 mars 2012

1 Définition, exemples

- Risque empirique de $f \in \mathcal{F}$:

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f;(X_i,Y_i))$$

- L'espérance du risque empirique est égale au risque : pour tout $f \in \mathcal{F}$ déterministe,

$$\mathbb{E}\left[\widehat{\mathcal{R}}_n(f)\right] = \mathbb{E}\left[\ell(f;(X_1,Y_1))\right] = \mathcal{R}(f) \quad .$$

- Intuition : on sur-apprend quand on minimise le risque empirique sur l'ensemble de tous les prédicteurs
- Définition formelle :
  Modèle : $S \subset \mathcal{F}$ (ensemble de prédicteurs).
  Minimiseur du risque empirique sur S :

$$\widehat{f}_S \in \arg\min_{f \in S} \left\{ \widehat{\mathcal{R}}_n(f) \right\}$$

- Exemples de modèles classiques : partitions (voir cours du 23/02), séparation linéaire dans $\mathbb{R}^d$, intervalles dans $[0,1]$.

- Difficultés algorithmiques potentielles (perte convexe ou pas), minimiseur approché du risque empirique.

2 Majoration générale du risque (Vapnik)

- Erreur d'approximation

$$\inf_{f \in S} \mathcal{R}(f) - \mathcal{R}(f^\star) = \mathcal{R}(S) - \mathcal{R}(f^\star)$$

et erreur d'estimation

$$\mathcal{R}\left(\widehat{f}_S\right) - \inf_{f\in S} \mathcal{R}\left(f\right)$$

Notation

$$f_S^\star \in \arg\min_{f\in S} \mathcal{R}\left(f\right)$$

(s'il existe), de telle sorte que

$$\mathcal{R}\left(S\right) = \mathcal{R}\left(f_S^\star\right)$$

Proposition 1 (Majoration de l'erreur d'estimation, Vapnik)**.** *Soit $\widehat{f}_S$ un algorithme d'apprentissage minimisant le risque empirique sur $S \subset \mathcal{F}$. Alors,*

$$\mathcal{R}\left(\widehat{f}_S\right) - \inf_{f\in S} \mathcal{R}\left(f\right) \leqslant 2\sup_{f\in S}\left|\mathcal{R}\left(f\right) - \widehat{\mathcal{R}}_n\left(f\right)\right| \ . \tag{1}$$

Démonstration. Soit $\delta > 0$ et $f_{S,\delta}^\star \in S$ tel que

$$\mathcal{R}\left(f_{S,\delta}^\star\right) \leqslant \inf_{f\in S} \mathcal{R}\left(f\right) + \delta \ .$$

Par définition de $f_{S,\delta}^\star$ et de $\widehat{f}_S$,

$$\begin{aligned}
\mathcal{R}\left(\widehat{f}_S\right) - \inf_{f\in S} \mathcal{R}\left(f\right) &\leqslant \mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}\left(f_{S,\delta}^\star\right) + \delta \\
&= \mathcal{R}\left(\widehat{f}_S\right) - \widehat{\mathcal{R}}_n\left(\widehat{f}_S\right) + \widehat{\mathcal{R}}_n\left(\widehat{f}_S\right) - \widehat{\mathcal{R}}_n\left(f_{S,\delta}^\star\right) + \widehat{\mathcal{R}}_n\left(f_{S,\delta}^\star\right) - \mathcal{R}\left(f_{S,\delta}^\star\right) + \delta \\
&\leqslant \mathcal{R}\left(\widehat{f}_S\right) - \widehat{\mathcal{R}}_n\left(\widehat{f}\right) + \widehat{\mathcal{R}}_n\left(f_{S,\delta}^\star\right) - \mathcal{R}\left(f_{S,\delta}^\star\right) + \delta \\
&\leqslant 2\sup_{f\in S}\left|\mathcal{R}\left(f\right) - \widehat{\mathcal{R}}_n\left(f\right)\right| + \delta
\end{aligned}$$

car $f_{S,\delta}^\star, \widehat{f}_S \in S$ p.s. On en déduit le résultat en faisant tendre δ vers zéro. □

Exercice 1. Soit $\rho > 0$ et $S \subset \mathcal{F}$. Soit $\widehat{f}_S$ un algorithme d'apprentissage tel que $\widehat{\mathcal{R}}_n\left(\widehat{f}_S\right) \leqslant \inf_{f\in S}\left\{\widehat{\mathcal{R}}_n\left(f\right)\right\} + \rho$. Montrer que

$$\mathcal{R}\left(\widehat{f}_S\right) - \inf_{f\in S} \mathcal{R}\left(f\right) \leqslant 2\sup_{f\in S}\left|\mathcal{R}\left(f\right) - \widehat{\mathcal{R}}_n\left(f\right)\right| + \rho \tag{2}$$

et

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right) - \inf_{f\in S} \mathcal{R}\left(f\right)\right] \leqslant \mathbb{E}\left[\sup_{f\in S}\left\{\mathcal{R}\left(f\right) - \widehat{\mathcal{R}}_n\left(f\right)\right\}\right] + \rho \ . \tag{3}$$

– Intuition sur le compromis approximation/estimation. Notions de sur-apprentissage et sous-apprentissage associées.

3 Cas d'un modèle fini avec une perte bornée

Théorème 1. *On suppose que la fonction de perte ℓ prend ses valeurs dans $[0,1]$, et l'on note S un sous-ensemble fini de $\mathcal{F}$. Alors,*

$$\mathbb{E}\left[\sup_{f\in S}\left\{\mathcal{R}\,(f)-\widehat{\mathcal{R}}_n\,(f)\right\}\right]\leqslant\frac{1}{\sqrt{2n}}\left[\sqrt{\ln(\mathrm{card}(S))}+\frac{\sqrt{\pi}}{2}\right]\ . \tag{4}$$

Note : en utilisant le Théorème 4 et la Proposition 3 de la boîte à outils de probabilités, on peut améliorer (4) en supprimant le terme $\frac{\sqrt{\pi}}{2}$.
Exemple : classification avec la perte 0–1. Régression avec la perte quadratique si les données sont bornées.

Corollaire 2. *On considère la perte $0-1$ $\ell\,(f(x),y)=\mathbb{1}_{f(x)\neq y}$ en classification binaire supervisée, et l'on suppose que $\mathcal{X}$ est fini. On définit l'algorithme de majorité suivant : pour tout $n\in\mathbb{N}$, $n\geqslant 1$, pour tout $x\in\mathcal{X}$,*

$$\widehat{f}^{\,\mathrm{maj}}(x;D_n):=\begin{cases}1 & \textit{si } \mathrm{card}\,\{\,i\ \textit{t.q. } X_i=x\ \textit{et } Y_i=1\,\}>\mathrm{card}\,\{\,i\ \textit{t.q. } X_i=x\ \textit{et } Y_i=0\,\}\\ 0 & \textit{sinon.}\end{cases}$$

Alors, pour tout $n\in\mathbb{N}$,

$$\sup_P\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}\left(\widehat{f}^{\,\mathrm{maj}}(D_n)\right)-\mathcal{R}\,(f^\star)\right]\leqslant\frac{\sqrt{\mathrm{card}(\mathcal{X})\ln(2)}+\frac{\sqrt{\pi}}{2}}{\sqrt{2n}}\ . \tag{5}$$

En particulier, $\widehat{f}^{\,\mathrm{maj}}$ est uniformément universellement consistante.

4 Classification binaire avec la perte 0–1

On suppose $\mathcal{Y}=\{0,1\}$ et que pour tout $y,y'\in\mathcal{Y}$, $\ell(y,y')=\mathbb{1}_{y\neq y'}$ (perte 0–1). Alors, toute fonction $f:\mathcal{X}\mapsto\mathcal{Y}$ est naturellement associée à l'ensemble

$$A(f)=\{\,x\in\mathcal{X}\ \text{t.q. }f(x)=1\,\}\subset\mathcal{X}\ .$$

Réciproquement, pour tout $A\subset\mathcal{X}$, on note $f_A=\mathbb{1}_A$ la fonction $\mathcal{X}\mapsto\mathcal{Y}$ associée à A. Cette correspondance se transfère au niveau des modèles : pour tout ensemble S de fonctions $\mathcal{X}\mapsto\mathcal{Y}$, on définit

$$\mathcal{A}(S)=\{\,A(f)\ \text{t.q. }f\in S\,\}\subset\mathfrak{P}\,(\mathcal{X})$$

et réciproquement, pour tout $\mathcal{A}\subset\mathfrak{P}\,(\mathcal{X})$, on note

$$S_{\mathcal{A}}=\{\,f_A\ \text{t.q. }A\in\mathcal{A}\,\}\ .$$

Le Théorème 1 peut s'étendre au cas d'un modèle S quelconque en remarquant que tout se passe comme si $S=S_{\mathcal{A}}$ était fini de cardinal

$$N_{\mathcal{A}}(X_1,\ldots,X_n):=\mathrm{card}\,\{\,A\cap\{\,X_1,\ldots,X_n\,\}\ \text{t.q. }A\in\mathcal{A}\,\}\leqslant 2^n\ .$$

Plus précisément, on a le résultat suivant (admis) :

$$\mathbb{E}\left[\sup_{f\in S}\left\{\mathcal{R}(f)-\widehat{\mathcal{R}}_n(f)\right\}\right]\leqslant\frac{2\sqrt{2}}{n}\mathbb{E}\left[\sqrt{\ln\left(N_{\mathcal{A}(S)}(X_1,\dots,X_n)\right)}\right]\ . \tag{6}$$

La quantité clé

$$H_S(X_1,\dots,X_n):=\ln\left(N_{\mathcal{A}(S)}(X_1,\dots,X_n)\right)$$

est appelée *entropie combinatoire* du modèle S.

La majoration (6) est notamment intéressante pour le cas des *classes de Vapnik-Chervonenkis*, c'est-à-dire, les modèles S tels que

$$V(S):=\sup\left\{k\geqslant 1\text{ t.q. }\sup_{x_1,\dots,x_k\in\mathcal{X}}N_{\mathcal{A}(S)}(x_1,\dots,x_k)=2^k\right\}<+\infty\ .$$

La quantité $V(S)$ est appelée *dimension de Vapnik-Chervonenkis* de S. Une propriété remarquable d'une classe de Vapnik-Chervonenkis S est qu'elle satisfait le lemme de Sauer (admis) :

$$\forall n\geqslant 1\,,\quad\forall x_1,\dots,x_n\in\mathcal{X}\,,\quad N_{\mathcal{A}(S)}(x_1,\dots,x_k)\leqslant\sum_{i=0}^{V(S)}\binom{n}{i} \tag{7}$$

et ce majorant est inférieur à $(en/V(S))^{V(S)}$ pour tout $n>2V(S)$.

Exercice 2. Déterminer si le modèle $S_{\mathcal{A}}$ est une classe de Vapnik-Chervonenkis lorsque $\mathcal{A}$ est :

1. $\mathcal{A}$ est fini.
2. $\mathcal{X}=\mathbb{R}$ et $\mathcal{A}$ est l'ensemble des demi-droites de la forme $]-\infty,a]$ avec $a\in\mathbb{R}$.
3. $\mathcal{X}=\mathbb{R}$ et $\mathcal{A}$ est l'ensemble des demi-droites.
4. $\mathcal{X}=\mathbb{R}$ et $\mathcal{A}$ est l'ensemble des intervalles de $\mathbb{R}$.
5. $\mathcal{X}=\mathbb{R}^2$ et $\mathcal{A}$ est l'ensemble des parties convexes de $\mathbb{R}^2$.
6. $\mathcal{X}=\mathbb{R}^d$ et $\mathcal{A}=\{\,]-\infty,a_1]\times\cdots\times]-\infty,a_d]$ t.q. $a_1,\dots,a_d\in\mathbb{R}\,\}$.
7. $\mathcal{X}=\mathbb{R}^d$ et $\mathcal{A}$ est l'ensemble des pavés de $\mathbb{R}^d$.
8. $\mathcal{X}=\mathbb{R}^d$ et $\mathcal{A}$ est l'ensemble des demi-espaces de $\mathbb{R}^d$ (commencer par $d=2$ ou 3).

Pour les cas où $S_{\mathcal{A}}$ est une classe de Vapnik-Chervonenkis, déterminer sa dimension de Vapnik-Chervonenkis ; si de plus $\mathcal{X}=\mathbb{R}$, minorer et majorer

$$\sup_{x_1,\dots,x_k\in\mathcal{X}}N_{\mathcal{A}}(x_1,\dots,x_k)$$

afin de tester la précision du Lemme de Sauer.

Indication/Bonus : on a le théorème général suivant : si $\mathcal{X}=\mathbb{R}^d$ et $\mathcal{A}=\{\,\{\,x\text{ t.q. }g(x)\geqslant 0\,\}\text{ t.q. }g\in\mathcal{G}\,\}$ avec $\mathcal{G}$ un espace vectoriel de dimension finie r de fonctions $\mathbb{R}^d\mapsto\mathbb{R}$, alors $S_{\mathcal{A}}$ est une classe de Vapnik-Chervonenkis de dimension $V(S_{\mathcal{A}})\leqslant r$.