Université Paris Diderot | **Apprentissage statistique**
Master M2MO: 2017/2018 | **Joseph Salmon - Stéphan Clémençon**

Travaux dirigés n° 2 : Concentration, théorie de VC

Stéphan Clémençon <stephan.clemencon@telecom-paristech.fr>
Joseph Salmon <joseph.salmon@telecom-paristech.fr>

Exercice 1. On se place dans le modèle de classification binaire. On considère une classe finie $\mathcal{G}$ de classifieurs telle que les deux populations soient parfaitement séparables par un élément de $\mathcal{G}$. On a donc $\min_{g\in\mathcal{G}} L(g) = 0$, avec la notation $L(g) = \mathbb{P}(g(X) \neq Y)$.

On dispose d'un échantillon i.i.d. $\{(X_i, Y_i)\}_{i=1,\dots,n}$ suivant la même loi que (X, Y) et on note $\hat{g}_n$ le minimiseur de l'erreur empirique de classification.

— Montrer qu'alors $\min_{g\in\mathcal{G}} L_n(g) = 0$ où $L_n(g) := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbb{1}_{\{i\in[\![1,n]\!]:g(X_i)\neq Y_i\}}$.

— Montrer que, pour tout n et tout $\epsilon \in [0, 1]$, on a :

$$\mathbb{P}\{L(\hat{g}_n) > \epsilon\} \leq |\mathcal{G}|(1-\epsilon)^n$$

puis que pour tout $\epsilon > 0$

$$\mathbb{P}\{L(\hat{g}_n) > \epsilon\} \leq |\mathcal{G}|e^{-n\epsilon}$$

On pourra utiliser définir l'ensemble de classifieur $\mathcal{G}_{\text{bad}} = \{g \in \mathcal{G} : L(g) > \epsilon\}$ ainsi qu'une borne d'union.

— En déduire que, pour tout n :

$$\mathbb{E}(L(\hat{g}_n)) \leq \frac{\log(e|\mathcal{G}|)}{n}$$

On pourra utiliser la formule $\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X > t)dt$, vraie pour toute variable aléatoire positive.

Exercice 2. Soit X une variable aléatoire telle que $X \in [a, b]$ presque sûrement. On pose $\varphi(s) = \log \mathbb{E}(e^{sX})$.

1) Montrer que $\varphi'(s)$ et $\varphi''(s)$ sont les moments d'une variable aléatoire dont on précisera la loi.
2) Soit Z une variable aléatoire telle que $Z \in [a, b]$ presque sûrement, montrer que :

$$\mathrm{Var}(Z) \leq \frac{(b-a)^2}{4} \ .$$

3) Appliquer la formule de Taylor avec reste intégral et déduire de la question précédente l'inégalité de Hoeffding.

Exercice 3. Calculer la VC dimension des classes $\mathcal{A}$ d'ensembles suivantes :

a. $\mathcal{A} = \{\ ]-\infty, x_1] \times \ldots \times]-\infty, x_d]\ :\ (x_1, \ldots, x_d) \in \mathbb{R}^d\ \}$,
b. $\mathcal{A}$ est constituée des rectangles de $\mathbb{R}^d$.

Exercice 4. Donner une borne supérieure de la VC dimension de la classe des boules fermées dans $\mathbb{R}^d$:

$$\left\{ x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d \ : \ \sum_{i=1}^{d} |x_i - a_i|^2 \leq b \right\}$$

où $a_1, \ldots, a_d, b \in \mathbb{R}$.

Exercice 5. Soit $\mathcal{A}$ une classe d'ensembles de $\mathbb{R}^d$ de VC dimension $V < +\infty$ et de coefficient d'éclatement $s(\mathcal{A}, n)$, $\forall n \geq 1$.

1) Montrer que : $\forall n \geq 1$, $s(\mathcal{A}, n) \leq (n+1)^V$.
2) Montrer que : $\forall n \geq V$, $s(\mathcal{A}, n) \leq (ne/V)^V$.

Indication : on utilisera le lemme de Sauer : $\forall n \geq 1$, $s(\mathcal{A}, n) \leq \sum_{i=0}^{V} C_i^n$.