École Normale Supérieure
FIMFA — Statistique 2018–2019

Devoir 3 pour le 8 janvier 2018

1. classification

Les définitions utilisées ici sont tirées du chapitre 9 des notes de cours.

(a) vc-dimension.

i. Soit $\mathcal{X} = \{0,1\}^n$ (le n-cube discret). On définit sur $\mathcal{X}$ l'opération $\odot : x \odot y = \sum_{i=1}^n x_i y_i \mod 2$. Les fonctions parités (ou fonctions de Walsh sont définies par $\chi_w(x) = (-1)^{w \odot x}$). Quelle est la dimension de Vapnik-Chervonenkis de l'ensemble des parties définies par $\{\chi_w : w \in \{0,1\}^n\}$?

ii. Soit $\mathcal{X} = \mathbb{R}^d$, soit $\mathcal{C}$ la classe des parallélépipèdes rectangles dont les cotés sont parallèles aux axes. Quelle est la dimension de Vapnik-Chervonenkis de $\mathcal{C}$?

iii. ☕☕ Soit $\mathcal{X} = \mathbb{R}^d$, soit $\mathcal{C}$ la classe des boules euclidiennes dans $\mathcal{X}$. Quelle est la dimension de Vapnik-Chervonenkis de $\mathcal{C}$?

iv. Montrer que les polyèdres convexes ont une dimension de Vapnik-Chervonenkis infinie dans $\mathbb{R}^2$.

Avec les notations des notes de cours, on suppose que P-p.s. $|\eta(x)| = 1$, ce qui signifie que (P-p.s.) l'étiquette Y est déterminée par X et que pour le classifieur bayésien f^*, $L(f^*) = 0$ et $L_n(f^*) = 0$ (p.s.). On suppose que la classe $\mathcal{F}$ de classifieurs sur laquelle on effectue la minimisation du risque empirique est finie et contient le classifieur bayésien. Note $\widehat{f}_n$ un minimisant du risque empirique pris dans $\mathcal{F}$.

(b) Montrer que

$$\mathbb{E}[L(\widehat{f}_n)] \leq \frac{\log|\mathcal{F}| + 1}{n}$$

2. estimation de densité

On s'intéresse à une méthode simple et utile : l'estimation de densité par histogrammes. On s'intéresse aux densités à support sur $[0,1]$. Dans la suite $X_1, \ldots, X_n \sim_{\text{i.i.d.}} f$ où f est la densité à estimer. On suppose que

$$\|f\|_2^2 = \int_0^1 f(x)^2 \mathrm{d}x < \infty$$

On se donne une suite croissante de points $z_0 = 0 < z_1 < \ldots < z_D = 1$ ($1 < D \in \mathbb{N}$). On définit l'estimateur $\widehat{f}_n$ par

$$\widehat{f}_n(x) = \frac{F_n(z_j) - F_n(z_{j-1})}{z_j - z_{j-1}} \qquad \text{si } x \in (z_{j-1}, z_j]$$

(a) Pour $j = 1, \ldots, D$, on note $p_j = \int_{z_{j-1}}^{z_j} f(x)\mathrm{d}x$. Calculer $\mathbb{E}\widehat{f}_n(x)$, et $\mathrm{var}(\widehat{f}_n(x))$ lorsque $x \in (z_{j-1}, z_j]$

(b) Calculer le mise :

$$\mathbb{E}\left[\int_0^1 (f(x) - \widehat{f}_n)^2 \mathrm{d}x\right]$$

(c) Vérifier que si $\|f\|_\infty \leq M < \infty$,

$$\mathbb{E}\left[\int_0^1 (f(x) - \widehat{f}_n)^2 \mathrm{d}x\right] \leq \frac{M(D-1)}{n} + \|f - \mathrm{Proj}\, f\|^2$$

où Proj f est la projection orthogonale de f sur le sous espace vectoriel des fonctions constantes sur les intervalles $(z_{j-1}, z_j]$.

(d) On suppose à partir de maintenant, que $z_j - z_{j-1} = 1/D$ (intervalles réguliers), montrer que

$$\mathbb{E}\left[\int_0^1 (f(x) - \widehat{f}_n)^2 \mathrm{d}x\right] \leq \frac{D-1}{n} + \|f - \mathrm{Proj}\, f\|^2$$

Est-ce une amélioration ?

(e) Supposer f continue sur $[0,1]$. On laisse D_n dépendre de n, avec $D_n \to \infty$ mais $D_n/n \to 0$. Montrer que

$$\mathbb{E}\left[\int_0^1 (f(x) - \widehat{f}_n)^2 \mathrm{d}x\right] \to 0 \quad \text{quand } n \to 0\,.$$

(f) Supposer en plus f, L-Lipschitzienne, montrer que :

$$\mathbb{E}\left[\int_0^1 (f(x) - \widehat{f}_n)^2 \mathrm{d}x\right] \leq \frac{D_n - 1}{n} + \frac{L^2}{D_n^2}\,.$$

Comment choisir D_n (si on sait que la densité à estimer est L-Lipschitzienne) ?

3. TESTS NON-PARAMÉTRIQUES

Soit F, la fonction de répartition d'une loi absolument continue sur $\mathbb{R}$. On définit la statistique T_n par

$$T_n(x_1, \ldots, x_n) = \sqrt{n} \sup_{x<y} |F_n(y) - F_n(x) - (F(y) - F(x))|\,.$$

On veut construire un test d'adéquation à F avec une procédure de décision de la forme

$$\begin{cases} \text{rejet de } X_1, \ldots, X_n \sim_{\text{i.i.d.}} F & \text{si } T_n \geq \tau \\ \text{acceptation de } X_1, \ldots, X_n \sim_{\text{i.i.d.}} F & \text{si } T_n < \tau \end{cases}$$

(a) Montrer que si $X_1, \ldots, X_n \sim_{\text{i.i.d}} F$, la loi de T_n est libre de F.

(b) Montrer que si $X_1, \ldots, X_n \not\sim_{\text{i.i.d}} F$, T_n converge en probabilité vers $+\infty$.

(c) Montrer que si $X_1, \ldots, X_n \sim_{\text{i.i.d}} F$, $\mathbb{E}[T_n] \leq \frac{\kappa}{\sqrt{n}}$ pour une constante $\kappa > 0$.

BAYÉSIEN

On se place dans un cadre d'inférence bayésienne. Les paramètres θ à valeur dans Θ sont tirés selon une loi a priori Π. de densité π par rapport à une mesure σ-finie ν. Les lois P_θ sont de densité $p(\cdot \mid \theta)$ par rapport à une dominante μ. On résume cette situation par le diagramme :

$$\begin{array}{cc} \theta \sim \Pi & \mathrm{d}\Pi = \pi \mathrm{d}\nu \\ X_1, \ldots, X_n \mid \theta \sim P_\theta^{\otimes n} & \mathrm{d}P_\theta = p(\mid \theta)\mathrm{d}\mu\,. \end{array}$$

4. (a) Caractériser (par exemple en donnant leurs densités) les lois a posteriori de $\theta \mid X_1$ et $\theta \mid X_1, X_2$.

(b) Montrer que $\Pi(\cdot \mid X_1, X_2)$, la loi a posteriori sachant X_1, X_2 coincide avec la loi a posteriori $\Pi'(\cdot \mid X_2)$ si l'a priori $\Pi'()$ est la loi a posteriori $\Pi(\cdot \mid X_1)$.

(c) Étendre à n observations : montrer que $\Pi(\cdot \mid X_1, X_2, \ldots, X_n)$ coïncide avec la loi a posteriori $\Pi'(\cdot \mid X_n)$ si l'a priori $\Pi'()$ est la loi a posteriori $\Pi(\cdot \mid X_1, \ldots, X_{n_1})$.

(d) L'ordre du conditionnement $\theta \mid X_1, \ldots, X_n$ est il important ?

(e) On note P_X^n la loi de $X_1, \ldots, X_n$ définie par

$$P_X^n\{A\} = \int_\Theta P_\theta^{\otimes n}(A)\mathrm{d}\Pi(\theta)$$

pour $A \in \mathcal{B}(\mathbb{R}^n)$. Montrer que la loi P_X^n est invariante par permutation.

On précise maintenant $\Theta = \mathbb{R}^2$, $\theta = (\theta_1, \theta_2)^T$, et

$$X \mid \theta \sim \mathcal{N}\left(\frac{\theta_1 + \theta_2}{2}, 1\right) \qquad \theta \sim \Pi$$

avec Π de densité $2h(\theta_1 + \theta_2)h(\theta_1 - \theta_2)$, $h \geq 0$, et $\int h(u)\mathrm{d}u = 1$. On note

$$z_1 = \frac{\theta_1 + \theta_2}{2} \qquad z_2 = \frac{\theta_1 - \theta_2}{2}\,.$$

(a) Loi de $z = (z_1, z_2)^T$, loi de z_2.

(b) Densité a posteriori de $\theta \mid X$. Densité a posteriori de $z \mid X$

(c) Densité de $z_2 \mid X$.