K Plus Proche Voisins et Consistance

Exercice 1 Soit μ une probabilité à densité sur $\mathbb{R}^d$. Son support est défini par

$$\operatorname{supp}\mu := \{x \in \mathbb{R}^d,\ \forall \epsilon > 0,\ \mu(B(x,\epsilon)) > 0\}$$

où $B(x,\epsilon)$ désigne la boule fermée centrée en x de rayon ϵ. Soit $(X_i, Y_i)_{1\le i\le n}$ i.i.d. à valeurs dans $\mathbb{R}^d \times \{0,1\}$ avec $(X_i)_{1\le i\le n}$ suivant la loi μ. On définit pour tout $k < n$ et tout $x \in \mathbb{R}^d$ le point aléatoire $X_{(k)}(x)$ par

$$\#\{1 \le i \le n,\ d(x, X_i) \le d(x, X_{(k)}(x))\} = k$$

avec $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ la distance Euclidienne.

1. Montrer que pour tout $x \in \operatorname{supp}\mu$, la convergence presque sûre a lieu

   $$\forall k \in \mathbb{N},\ \lim_{n\to\infty} d(X_{(k)}(x), x) = 0.$$

   Déduire, pour une v.a. X de loi μ indépendante des observations, la convergence p.s.

   $$\lim_{n\to\infty} d(X_{(k)}(X), X) = 0.$$

2. Admettons le lemme technique suivant

   Lemma 1 (Stone)**.** *Soit une fonction $f \in \mathbb{L}^1(\mathbb{R}^d, \mu)$. Sous les hypothèses précédentes, pour tout $(k,n) \in \mathbb{N}^2$ vérifiant $k \le n$, il existe une constant γ_d qui ne dépend que de la dimension d telle que*

   $$\frac{1}{k}\sum_{i=1}^{k} \|f(X_{(i)}(X))\|_1 \le \gamma_d \|f(X)\|_1.$$

   Toujours sous nos hypothèses, montrer pour toute fonction $f \in \mathbb{L}^1(\mathbb{R}^d, \mu)$

   $$\lim_{n\to\infty} \frac{1}{k}\sum_{i=1}^{k} \|f(X) - f(X_{(i)}(X))\|_1 = 0.$$

3. La règle de k-PPV peut s'écrire

   $$g_n(x) = \begin{cases} 1, & \phi(x, Y_{(1)}(x), \cdots, Y_{(k)}(x)) > 0 \\ 0, & sinon \end{cases}$$

   avec $Y_{(i)}(x)$ l'étiquette associée à $X_{(i)}(x)$. Préciser la fonction ϕ.

4. Soit $(U_i)_{1\le i\le n}$ v.a. uniforme i.i.d. indépendantes de $(X_i)_{1\le i\le n}$. Justifier la représentation

$$Y_{(i)}(x) = 1_{\eta(X_{(i)}(x))>U_i}.$$

En les substituant par

$$Y'_{(i)}(x) = 1_{\eta(x)>U_i}$$

on définit un nouveau classifieur $g'_n(x)$. Montrer

$$\mathbb{P}(g_n(X) \neq g'_n(X)) \le \sum_{i=1}^{k} \|\eta(X) - \eta(X_{(k)}(X))\|_1.$$

5. Soit L_n l'erreur de Bayes du classifieur g_n. Déduire, pour un k impair fixé, la convergence

$$L_{knn} := \lim_{n\to\infty} \mathbb{E}L_n = L^* + \mathbb{E}\big[(1-2\min(\eta(X),1-\eta(X)))\mathbb{P}(B(k,\min(\eta(X),1-\eta(X))) > \frac{k}{2})|X)\big]$$

avec $B(k,\alpha)$ une v.a. binomiale de paramètre α.

6. Déduire la borne supérieure du risque L_{knn}

$$L_{knn} \le L^* + \frac{1}{\sqrt{ke}}.$$

7. Soit $(Z_i)_{i\ge1}$ i.i.d. vérifiant $\mathbb{P}(Z_i = 1) = 1 - \mathbb{P}(Z_i = -1) < 1/2$. Montrer

$$\forall m > 1,\ \mathbb{P}\Big(\sum_{i=1}^{2m+1} Z_i > 0\Big) \le \mathbb{P}\Big(\sum_{i=1}^{2m-1} Z_i > 0\Big).$$

8. Conclure avec l'inégalité de Cover-Hart

$$\forall m \ge 1,\ L^* \le L_{(2m+1)nn} \le L_{1nn} \le 2L^*.$$

Qu'en est-il si $L^* = 0$?

9. Supposons maintenant que k croît avec le nombre d'échantillon n et qu'il vérifie

$$\lim_{n\to\infty} k/n = 0 \ \text{ et } \ \lim_{n\to\infty} k = +\infty.$$

Montrer en utilisant le théorème de Stone que la méthode de k-PPV est universellement consistante.