Correction du devoir : classification supervisée

Romain Azaïs

19/12/2023

Partie I

1 $R(g_n) - R^* \geq 0$

2 $P(g_n(X) \neq Y \mid (X_i, Y_i), X)$

$$= 1 - P(g_n(X) \neq Y,\ Y = 1 \mid (X_i, Y_i), X) - P(g_n(X) \neq Y,\ Y = 0 \mid (X_i, Y_i), X)$$

$$= 1 - \eta(X)\, \mathbb{1}_{\{g_n(X) = 1\}} - (1 - \eta(X))\, \mathbb{1}_{\{g_n(X) = 0\}}$$

3 La même égalité est vraie pour g^*. Donc

$$P(g_n(X) \neq Y \mid (X_i, Y_i), X) - P(g^*(X) \neq Y \mid X)$$

$$= -\eta(X) \underbrace{\left[\mathbb{1}_{\{g_n(X) = 1\}} - \mathbb{1}_{\{g^*(X) = 1\}} \right]}_{|\ | = \begin{cases} 1 & \text{si } g_n(X) \neq g^*(X) \\ 0 & \text{sinon} \end{cases}} - (1 - \eta(X)) \underbrace{\left[\mathbb{1}_{\{g_n(X) = 0\}} - \mathbb{1}_{\{g^*(X) = 0\}} \right]}_{|\ | = \begin{cases} 1 & \text{si } g_n(X) \neq g^*(X) \\ 0 & \text{sinon} \end{cases}}$$

$$\leq |2\eta(X) - 1|\, \mathbb{1}_{\{g_n(X) \neq g^*(X)\}}$$

4 On passe à l'espérance sur X :

$$R(g_n) - R^* \leq E\left[|2\eta(X) - 1|\, \mathbb{1}_{\{g_n(X) \neq g^*(X)\}} \,\middle|\, (X_i, Y_i) \right]$$

Or, si $\mathbb{1}_{\{g_n(X) \neq g^*(X)\}} = 1$, $\left|\eta(X) - \frac{1}{2}\right| \leq |\eta(X) - \eta_n(X)|$

Donc $R(g_n) - R^* \leq 2 \int_{\mathbb{R}^d} |\eta_n(x) - \eta(x)|\, \mu(dx)$

Partie II

5 $\omega_{n,i}(x) = v_{n,j}$ où j est l'indice tel que $Y_i = Y_{(j)}$

6 μ est à densité donc $\mu(B(x,\cdot)) = \frac{k_n}{n}$ admet une unique solution notée $\rho_n(x)$

Partie III

7 $\omega_{n,i}(x) = 0$ pour $i > k_n$

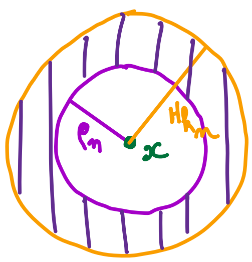
Or, par déf. de H_{k_n}, $\mathbb{1}_{B(x, H_{k_n})}(X_i) = 0$ ssi $i > k_n$

Ajouter cette indicatrice ne change donc pas la somme.

8 $\eta_n^{\infty}(x) - \eta_n(x) = \sum_{i=1}^{n} \omega_{n,i}(x)\, Y_i \left[\mathbb{1}_{B(x,\rho_n)}(X_i) - \mathbb{1}_{B(x,H_{k_n})}(X_i)\right]$

Or $\omega_{n,i}(x) \leq \frac{\alpha}{k_n}$ et $Y_i \leq 1$ donc

$$\left|\eta_n^{\infty}(x) - \eta_n(x)\right| \leq \frac{\alpha}{k_n} \underbrace{\sum_{i=1}^{n} \left|\mathbb{1}_{B(x,\rho_n)}(X_i) - \mathbb{1}_{B(x,H_{k_n})}(X_i)\right|}_{= \left| n\mu_n(B(x,\rho_n)) - k_n \right|}$$



$= \left| n\mu_n(B(x,\rho_n)) - k_n \right|$ par définition de la loi empirique et de H_{k_n}

Or, par 6, $\frac{k_n}{n} = \mu(B(x,\rho_n))$. Donc

$$\left|\eta_n^{\infty}(x) - \eta_n(x)\right| \leq \frac{\alpha n}{k_n} \left| \mu_n(B(x,\rho_n)) - \mu(B(x,\rho_n)) \right|$$

9 $\mu_n(B[x,\rho_n]) \times n \sim \text{Binom.}(\mu(B(x,\rho_n)), n)$

donc $\mathbb{E}[Z_n(x)^2] = \frac{\sigma^2}{k_n^2}\, n\, \underbrace{\mu(B(x,\rho_n))}_{\frac{k_n}{n}} \underbrace{[1-\mu(B(x,\rho_n))]}_{\leq 1}$

$$\mathbb{E}[Z_n(x)]^2 \leq \mathbb{E}[Z_n(x)^2] \leq \frac{\sigma^2}{k_n} \quad \text{donc} \quad \mathbb{E}[Z_n(x)] \leq \frac{\sigma}{\sqrt{k_n}}$$

Partie IV

10 $$|\eta_n^*(X) - \eta(X)| \leq |\eta_n^*(X) - \eta_n(X)| + |\eta_n(X) - \eta(X)|$$

Or $\mathbb{E}|\eta_n(X) - \eta(X)| \xrightarrow[n\to\infty]{} 0$ (par l'énoncé)

et $\mathbb{E}|\eta_n^*(X) - \eta_n(X)| \leq \mathbb{E}[Z_n(x)] \leq \frac{\sigma}{\sqrt{k_n}} \xrightarrow[n\to\infty]{} 0$

Donc, pour n assez grand, $\mathbb{E}|\eta_n^*(X) - \eta(X)| \leq \frac{\varepsilon}{20}$

11 $$|\eta_n(x) - \eta(x)| \leq \underbrace{|\eta_n(x) - \eta_n^*(x)|}_{(A)} + \underbrace{|\eta_n^*(x) - \eta(x)|}_{(B)}$$

(A) $$\int |\eta_n(x) - \eta_n^*(x)|\,\mu(dx) \leq \int Z_n(x)\,\mu(dx)$$

$$\leq \left| \int Z_n(x)\,\mu(dx) - \mathbb{E}\int Z_n(x)\mu(dx) \right|$$

$$+ \mathbb{E}\int Z_n(x)\mu(dx)$$

Or $\mathbb{E}\int Z_n(x)\mu(dx) \leq \frac{\sigma}{\sqrt{k_n}} \leq \frac{\varepsilon}{20}$ pour n assez grand

(B) $\int \left|\eta_n^{*}(x)-\eta(x)\right|\mu(dx)$

$$\leq \left|\int \left|\eta_n^{*}(x)-\eta(x)\right|\mu(dx) - \mathbb{E}\int \left|\eta_n^{*}(x)-\eta(x)\right|\mu(dx)\right|$$

$$+ \underbrace{\mathbb{E}\int \left|\eta_n^{*}(x)-\eta(x)\right|\mu(dx)}$$

$= \mathbb{E}\left|\eta_n^{*}(X)-\eta(X)\right| \leq \frac{\varepsilon}{20}$ pour n assez grand.

Bilan : $\int \left|\eta_n(x)-\eta(x)\right|\mu(dx)$

$$\leq \left|\int Z_n(x)\mu(dx) - \mathbb{E}\int Z_n(x)\mu(dx)\right|$$

$$+ \left|\int \left|\eta_n^{*}(x)-\eta(x)\right|\mu(dx) - \mathbb{E}\int \left|\eta_n^{*}(x)-\eta(x)\right|\mu(dx)\right| + \frac{\varepsilon}{10}$$

On conclut avec $\frac{\varepsilon}{2} - \frac{\varepsilon}{10} = \frac{2\varepsilon}{5}$.

Partie V

12 $\quad |\dots| \leq \underbrace{\left| y_i \mathbb{1}_{B(x,\rho_n)}(x_i) - y_i' \mathbb{1}_{B(x,\rho_n)}(x_i')\right|}_{\leq 1} \underbrace{\omega_{n,i}(x)}_{\leq \frac{\alpha}{k_n}}$

13 Si $|\ldots| > 0$

alors $\left| y_i \mathbb{1}_{B(x,\rho_n)}(x_i) - y'_i \mathbb{1}_{B(x,\rho_n)}(x'_i) \right| > 0$

Donc l'une des deux indicatrices (au moins) est non-nulle,
ie $x_i \in B(x,\rho_n)$ ou $x'_i \in B(x,\rho_n)$

14 Si $x_i \in B(x,\rho_n)$, $\|x_i - x\| \leq \rho_n$

donc $\mu(B(x,\|x_i - x\|)) \leq \mu(B(x,\rho_n)) = \frac{k_n}{n}$

Réciproquement, comme $B(x,\|x_i - x\|)$ et $B(x,\rho_n)$ partagent le même centre, et comme $\mu(B(x,\rho_n)) = \frac{k_n}{n}$,
$\mu(B(x,\|x_i - x\|)) \leq \frac{k_n}{n} \Rightarrow B(x,\|x_i - x\|) \subset B(x,\rho_n)$

et donc $\|x_i - x\| \leq \rho_n$ ie $x_i \in B(x,\rho_n)$

15 $\int |\ldots| \mu(dx) \leq \frac{\gamma}{k_n} \int \left[\mathbb{1}_{B(x,\rho_n(x))}(x_i) + \mathbb{1}_{B(x,\rho_n(x))}(x'_i) \right] \mu(dx)$

par **12** et **13**. Alors, par **14**,

$$\int |\ldots| \mu(dx) \leq \frac{\gamma}{k_n} \left[\int_{\{x : \mu(B(x,\|x_i - x\|)) \leq \frac{k_n}{n}\}} \mu(dx) + \int_{\{x : \mu(B(x,\|x_i - x\|)) \leq \frac{k_n}{n}\}} \mu(dx) \right]$$

Par le rés. admis de 15,

$$\int |\dots| \mu(dx) \leq \frac{\alpha}{k_n} \times 2 \gamma_d \frac{k_n}{n} = \frac{2 \alpha \gamma_d}{n}$$

Partie VI

16 On peut appliquer l'in. de McDiarmid par les rés. de la partie V. On a alors

$$\mathbb{P}\left(|\dots| \geq \frac{\varepsilon}{5} \right) \leq 2 \exp\left(- \frac{2 \left(\frac{\varepsilon}{5}\right)^2}{n \frac{4 \alpha^2 \gamma_d^2}{n^2}} \right)$$

$$\leq 2 \exp\left(- \frac{1}{50} \times \frac{n \varepsilon^2}{\alpha^2 \gamma_d^2} \right)$$

17 $\mathbb{P}(R(g_n) - R^* \geq \varepsilon) \leq \mathbb{P}\left(\int_{\mathbb{R}^d} |\eta_n(x) - \eta(x)| \mu(dx) \geq \frac{\varepsilon}{2} \right)$

par 4. Alors, par 11, 16 et l'hypothèse de 17,

$$\mathbb{P}(R(g_n) - R^* \geq \varepsilon) \leq 4 \exp\left(- \frac{n \varepsilon^2}{50 \alpha^2 \gamma_d^2} \right)$$

Bonus : seulement dans la question 6

Fin

Sujet fortement inspiré de Lectures on the Nearest Neighbor Method. Biau and Devroye (2015).