— Classification supervisée —

Problème : une inégalité exponentielle pour les k plus proches voisins

Cadre de Travail

• Soit (X, Y) un couple de variables aléatoires sur $\mathbb{R}^d \times \{0, 1\}$. On note μ la loi (à densité) de X et, pour tout $x \in \mathbb{R}^d$, $\eta(x) = \mathbb{P}(Y = 1|X = x)$. On note $g^\star$ la règle de classification définie par

$$g^\star(x) = 1 \quad \text{si et seulement si} \quad \eta(x) > \frac{1}{2}.$$

$R^\star$ désigne son risque défini par $R^\star = \mathbb{P}(g^\star(X) \neq Y)$.

• On se donne n répétitions $(X_i, Y_i)_{1\leq i\leq n}$ de même loi que (X, Y). Elles sont indépendantes et indépendantes de (X, Y). Il s'agit d'un jeu de données d'entrainement à partir duquel on cherche à apprendre une règle de prédiction de Y lorsqu'on observe X.

• Pour tout $x \in \mathbb{R}^d$, on note $(X_{(i)}(x), Y_{(i)}(x))_{1\leq i\leq n}$ la permutation des données telle que

$$\|X_{(1)}(x) - x\| \leq \|X_{(2)}(x) - x\| \leq \cdots \leq \|X_{(n)}(x) - x\|.$$

• Soient $v_{n,1} \geq \cdots \geq v_{n,n} \geq 0$ tels que $\sum_{i=1}^n v_{n,i} = 1$. La règle de classification des plus proches voisins est définie comme

$$g_n(x) = 1 \quad \text{si et seulement si} \quad \eta_n(x) > \frac{1}{2}, \quad \text{où} \quad \eta_n(x) = \sum_{i=1}^n v_{n,i} Y_{(i)}(x).$$

On note $R(g_n)$ son risque donné par $R(g_n) = \mathbb{P}(g_n(X) \neq Y|(X_i, Y_i)_{1\leq i\leq n})$.

• On suppose qu'il existe une suite $(k_n)_{n\geq 2}$ et $\alpha > 0$ tels que

1. $1 \leq k_n < n$;
2. $k_n \to \infty$ et $\frac{k_n}{n} \to 0$;
3. $v_{n,i} = 0$ si $i > k_n$;
4. $k_n v_{n,1} \leq \alpha$.

Objectif : Étudier le comportement de $R(g_n) - R^\star$ lorsque n tend vers l'infini.

Partie I

1. En une ligne, que dire de $R(g_n) - R^\star$?
2. Montrer que

$$\mathbb{P}(g_n(X) \neq Y|(X_i, Y_i)_{1\leq i\leq n}, X) = 1 - \left[\mathbb{1}_{\{g_n(X)=1\}}\eta(X) + \mathbb{1}_{\{g_n(X)=0\}}(1 - \eta(X))\right].$$

3. Montrer que

$$\mathbb{P}(g_n(X) \neq Y|(X_i, Y_i)_{1\leq i\leq n}, X) - \mathbb{P}(g^\star(X) \neq Y|X) \leq |2\eta(X) - 1|\mathbb{1}_{\{g_n(X)\neq g^\star(X)\}}.$$

4. En déduire que

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

Partie II

5. Justifier l'existence de $(w_{n,i}(x))_{1\leq i\leq n}$ tel que

$$\eta_n(x) = \sum_{i=1}^{n} w_{n,i}(x) Y_i.$$

6. En une ligne, justifier l'existence de $\rho_n(x)$ tel que

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

Désormais, on note

$$\eta_n^\star(x) = \sum_{i=1}^{n} w_{n,i}(x)\, Y_i\, \mathbb{1}_{B(x,\rho_n(x))}(X_i).$$

Partie III

7. On note $H_{k_n}(x) = \|X_{(k_n)}(x) - x\|$. Justifier que

$$\eta_n(x) = \sum_{i=1}^{n} w_{n,i}(x)\, Y_i\, \mathbb{1}_{B(x,H_{k_n}(x))}(X_i).$$

8. Montrer que $|\eta_n^\star(x) - \eta_n(x)| \leq Z_n(x)$, où

$$Z_n(x) = \frac{\alpha n}{k_n} \left|\mu_n\left[B(x, \rho_n(x))\right] - \mu\left[B(x, \rho_n(x))\right]\right|,$$

où μ_n désigne la loi empirique des X_i.

9. On rappelle que la variance de la loi binomiale de paramètre (m, p) est $mp(1-p)$. Montrer, via l'inégalité de Jensen, que

$$\mathbb{E}Z_n(x) \leq \frac{\alpha}{\sqrt{k_n}}.$$

Partie IV

On admet que $\mathbb{E}|\eta_n(X) - \eta(X)|$ tend vers 0 quand n tend vers l'infini sous les conditions de l'énoncé (estimateur des plus proches voisins de la fonction de régression).

Soit $\epsilon > 0$.

10. Montrer via l'inégalité triangulaire que, pour n assez grand, on a

$$\mathbb{E}|\eta_n^\star(X) - \eta(X)| < \frac{\epsilon}{20}.$$

11. À nouveau via l'inégalité triangulaire, montrer que

$$\begin{aligned}\mathbb{P}\left(\int_{\mathbb{R}^d} |\eta_n(x) - \eta(x)|\mu(dx) \geq \frac{\epsilon}{2}\right) \leq\ & \mathbb{P}\left(\left|\int_{\mathbb{R}^d} Z_n(x)\mu(dx) - \mathbb{E}\int_{\mathbb{R}^d} Z_n(x)\mu(dx)\right| \geq \frac{\epsilon}{5}\right) \\ & + \mathbb{P}\left(\left|\int_{\mathbb{R}^d} |\eta_n^\star(x) - \eta(x)|\mu(dx) - \mathbb{E}\int_{\mathbb{R}^d} |\eta_n^\star(x) - \eta(x)|\mu(dx)\right| \geq \frac{\epsilon}{5}\right).\end{aligned}$$

Partie V

On note $h_x(X_1, Y_1, \ldots, X_n, Y_n) = \eta_n^\star(x)$.

12. Justifier que

$$|h_x(x_1, y_1, \ldots, x_n, y_n) - h_x(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i', y_i', x_{i+1}, y_{i+1}, \ldots, x_n, y_n)| \leq \frac{\alpha}{k_n}.$$

13. Justifier que si

$$|h_x(x_1, y_1, \ldots, x_n, y_n) - h_x(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i', y_i', x_{i+1}, y_{i+1}, \ldots, x_n, y_n)| > 0,$$

alors $x_i \in B(x, \rho_n(x))$ ou $x_i' \in B(x, \rho_n(x))$.

14. Justifier que

$$x_i \in B(x, \rho_n(x)) \quad \text{si et seulement si} \quad \mu(B(x, \|x_i - x\|)) \leq \frac{k_n}{n}.$$

15. On admet qu'il existe une constante γ_d (qui ne dépend que de d) telle que

$$\mu\left(\left\{x \in \mathbb{R}^d \ : \ \mu(B(x, \|x_i - x\|)) \leq \frac{k_n}{n}\right\}\right) \leq \gamma_d \frac{k_n}{n}.$$

En déduire que

$$\int_{\mathbb{R}^d} |h_x(x_1, y_1, \ldots, x_n, y_n) - h_x(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i', y_i', x_{i+1}, y_{i+1}, \ldots, x_n, y_n)| \mu(dx) \leq \frac{2\alpha\gamma_d}{n}.$$

Partie VI

On admet l'inégalité de McDiarmid suivante. Soient $U_1, \ldots, U_n$ des variables aléatoires indépendantes à valeurs dans $A \subset \mathbb{R}$. Soit $g : A^n \to \mathbb{R}$ une fonction mesurable telle que

$$|g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i,$$

pour certaines constantes $c_i > 0$. Alors

$$\mathbb{P}\left(|g(U_1, \ldots, U_n) - \mathbb{E}g(U_1, \ldots, U_n)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

16. Appliquer l'inégalité de McDiarmid pour majorer

$$\mathbb{P}\left(\left|\int_{\mathbb{R}^d} |\eta_n^\star(x) - \eta(x)|\mu(dx) - \mathbb{E}\int_{\mathbb{R}^d} |\eta_n^\star(x) - \eta(x)|\mu(dx)\right| \geq \frac{\epsilon}{5}\right)$$

en fonction de n, ϵ, α et γ_d.

17. On admet que le terme

$$\mathbb{P}\left(\left|\int_{\mathbb{R}^d} Z_n(x)\mu(dx) - \mathbb{E}\int_{\mathbb{R}^d} Z_n(x)\mu(dx)\right| \geq \frac{\epsilon}{5}\right)$$

admet exactement le même majorant. Conclure en donnant un majorant de

$$\mathbb{P}(R(g_n) - R^\star \geq \epsilon).$$

Bonus : Où a-t-on utilisé l'hypothèse que X est à densité ?