Classification par la méthode des k plus proches voisins

On considère le problème de classification binaire. On dispose d'un ensemble d'apprentissage $\mathscr{D}_n := \{(X_i, Y_i)\}_{i=1}^n$ de variables aléatoires i.i.d. définies sur un espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. La loi de X est absolument continue par rapport à la mesure de Lebesgue λ_{Leb} sur $\mathbb{R}^d$, de densité notée f. On suppose de plus que le problème de classification est équilibré : la séquence $(Y_i)_{i\geq 1}$ est à valeurs dans $\{0,1\}$ et sa loi marginale est une loi de Bernoulli de paramètre $1/2$.

Dans le problème, on notera $\mathbb{P}_X(\cdot)$ la loi de probabilité de la variable aléatoire X et $\mathbb{E}_X$ l'espérance associée. De la même façon, on notera $\mathbb{P}_n(\cdot)$ la loi de probabilité du n-échantillon d'apprentissage $((X_1,Y_1),\dots,(X_n,Y_n))$ et $\mathbb{E}_n$ l'espérance associée. Enfin, $\mathbb{P}$ désignera la loi de probabilité globale, *i.e.* $\mathbb{P}_X \otimes \mathbb{P}_n$ et $\mathbb{E}$ l'espérance associée.

On munit l'ensemble $\mathbb{R}^d$ de la distance euclidienne $d(x,y) = \|x-y\|$. L'algorithme des k plus proches voisins procède de la façon suivante.

- Nous choisissons un entier k impair (cet entier peut dépendre de n).
- Étant donnée une nouvelle observation x, nous considérons les k plus proches voisins au sens de d dans l'ensemble $\mathscr{D}_n$, noté $\mathscr{V}_{n,k}(x)$.
- Nous prédisons la classe majoritaire des labels dans le sous-ensemble de $\mathscr{D}_n$ formé par $\mathscr{V}_{n,k}(x)$.

Nous notons $\Phi_{n,k}(x)$ le résultat de la classification prédite par les k plus proches voisins au point x étant donné l'échantillon d'apprentissage $\mathscr{D}_n$.

Nous supposons que X est à valeurs dans un ensemble compact K de $\mathbb{R}^d$. Soit $\mathrm{B}(x,r)$ la boule euclidienne fermée centrée en x et de rayon r :

$$\mathrm{B}(x,r) = \{w \in K :\ \|x-w\| \leq r\} \ .$$

Le classifieur Bayésien est noté :

$$\Phi^\star(x) = \mathbb{1}_{\{\eta^\star(x) \geq 1/2\}} \ ,$$

où nous avons posé :

$$\eta^\star(x) = \mathbb{E}\left[Y \,|\, X = x\right] \ .$$

On rappelle que $\Phi^\star$ est le classifieur optimal pour la perte 0-1, notée L, et définie par :

$$\mathrm{L}(\Phi) = \mathbb{P}(\Phi(X) \neq Y) \ .$$

L'excès de risque $\mathscr{E}(\Phi)$ du classifieur Φ est ainsi donné par :

$$\mathscr{E}(\Phi) = \mathrm{L}(\Phi) - \mathrm{L}(\Phi^\star) \ .$$

1. Étant donné $x \in K$, on note $(X_{(i)}(x), Y_{(i)}(x))_{i=1}^n$ les observations de $\mathscr{D}_n$ ordonnées de la plus proche à la plus éloignée de x.

   (a) Démontrer qu'il n'y a pas d'ambiguïté sur la prédiction d'une nouvelle observation avec probabilité 1.

   (b) Écrire la valeur de la prédiction $\Phi_{n,k}(x)$ en termes d'indicatrice d'événements.

(c) Démontrer en particulier que $\Phi_{n,k}$ est une méthode « *plug-in* » utilisant une estimation de la fonction de régression $\eta^\star(x)$, estimation notée $\eta_{n,k}$, et donnée par :

$$\eta_{n,k}(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x)$$

Justifier le caractère non-paramétrique de l'algorithme des k plus proches voisins.

Nous avons vu dans le cours que le théorème "*no free-lunch*" impose d'effectuer des hypothèses structurelles sur le problème de classification pour que celui-ci soit raisonnablement résoluble.

Introduisons une hypothèse relative à la loi des observations X.

- $(\mathbf{H}_\mu)$: Un couple $(c_0, r_0) \in {\mathbb{R}_+^*}^2$ existe tel que le support de f est (c_0, r_0)-régulier :

$$\forall r \leq r_0 : \qquad \lambda_{\text{Leb}}(\text{supp}(f) \cap \text{B}(x,r)) \geq c_0 \lambda_{\text{Leb}}(\text{B}(x,r)) \ ,$$

et la densité f est minorée et majorée sur son support :

$$\exists \mu > 0 \quad \forall x \in \text{supp}(f) \qquad \mu \leq f(x) \leq \frac{1}{\mu} \ .$$

2. Démontrer que $(\mathbf{H}_\mu)$ est *équivalente* à l'hypothèse suivante : il existe $0 < \mu$ tel que $f(x) \geq \mu$ pour tout $x \in \text{supp}(f)$ et il existe $m > 0$ et $\delta_0 > 0$ tel que, pour tout $\delta \leq \delta_0$ et $x \in K$,

$$\mathbb{P}_X(\text{B}(x,\delta)) \geq m f(x) \delta^d \ . \tag{1}$$

Introduisons une hypothèse relative à la fonction de régression $\eta^\star$.

- $(\mathbf{H}_\mathbf{L})$: La fonction de régression $\eta^\star$ est L-Lipschitz :

$$\forall (x,y) \in K^2 \qquad |\eta^\star(x) - \eta^\star(y)| \leq L\|x-y\| \ .$$

Introduisons une hypothèse relative à la variation de fonction de régression autour de la zone critique de décision $\eta^\star = 1/2$.

- $(\mathbf{H}_\alpha)$: La fonction de régression η satisfait la condition de marge :

$$\exists C > 0 \quad \forall \varepsilon > 0 \qquad \mathbb{P}_X\left(|\eta^\star(X) - \frac{1}{2}| \leq \varepsilon\right) \leq C\varepsilon^\alpha \ .$$

3. Donner un cas où $(\mathbf{H}_\alpha)$ est vérifiée pour $\alpha = 1$ en dimension 1. Même question pour $\alpha = +\infty$ en dimension 1. Quelle situation semble la plus favorable à la classification ?
4. Sous les hypothèses $(\mathbf{H}_\mathbf{L})$ et $(\mathbf{H}_\mu)$, démontrer que $(\mathbf{H}_\alpha)$ ne peut être vérifiée que pour $\alpha \leq d$.
5. Supposons que la fonction $\eta^\star$ soit r fois continûment différentiable Démontrer que, s'il existe x_0 dans l'intérieur de K tel que $\eta^\star(x_0) = 1/2$ et

$$\frac{\partial^{r_1} \dots \partial^{r_d}}{\partial x^{(1)} \dots \partial x^{(d)}} \eta^\star(x_0) = 0 \ , \quad \text{pour tout } (r_1, \dots, r_d) \in \mathbb{N}^d, \, r_1 + \cdots + r_d \leq r \ ,$$

alors la condition de marge ne peut être vérifiée que pour $\alpha \leq \frac{d}{r+1}$.

6. Sous l'hypothèse $(\mathbf{H}_\alpha)$, démontrer que l'excès de risque de classification satisfait pour tout $\varepsilon > 0$.
$$\mathscr{E}(\Phi_{n,k}) \leq 2C\varepsilon^{1+\alpha} + \mathbb{E}\left[|2\eta^\star(X) - 1| \mathbb{1}_{\{\Phi_{n,k}(X) \neq \Phi^\star(X)\}} \mathbb{1}_{\{|\eta^\star(X) - 1/2| > \varepsilon\}}\right] .$$
On utilisera l'expression de l'excès de risque donnée en PC9.
7. Donner la loi conditionnelle : $\mathscr{L}((Y_{(i)}(x))_{1 \leq i \leq n} | (X_1, \dots, X_n))$.
En notant $\mathbb{P}_{n,X_1^n}$ la loi des labels $(Y_1, \dots, Y_n)$ conditionnés aux positions $(X_1, \dots, X_n)$ et $\mathbb{E}_{n,X_1^n}$ son espérance associée, démontrer que pour tout $x \in K$:
$$\mathbb{P}_{n,X_1^n}\left(\left|\eta_{n,k}(x) - \mathbb{E}_{n,X_1^n}[\eta_{n,k}(x)]\right| > s\right) \leq 2\mathrm{e}^{-2ks^2} .$$

Pour $x \in K$, nous définissons
$$\Delta_n(x) = \left| \frac{1}{k} \sum_{i=1}^k \eta^\star(X_{(i)}(x)) - \eta^\star(x) \right| .$$

8. Démontrer que pour tout $\varepsilon > 0$ et $x \in K$:
$$\mathbb{1}_{\{|\eta^\star(x) - \frac{1}{2}| > \varepsilon\}} \mathbb{E}_{n,X_1^n}[\mathbb{1}_{\{\Phi_{n,k}(x) \neq \Phi^\star(x)\}}] \leq 2\mathbb{1}_{\{|\eta^\star(x) - \frac{1}{2}| > \varepsilon\}} \mathrm{e}^{-2k\lfloor \varepsilon - \Delta_n(x) \rfloor_+^2}, \tag{2}$$
où $\lfloor t \rfloor_+$ désigne la partie positive de t.
9. Pour tout $x \in K$ et $t \geq 0$, démontrer que, sous l'hypothèse $(\mathbf{H_L})$, on a :
$$\Delta_n(x) \leq Lt + \mathbb{1}_{\{|X_{(k)}(x) - x| \geq t\}} . \tag{3}$$

Définissons pour $t > 0$,
$$m_t(x) = \mathbb{P}_X(\mathrm{B}(x,t)) = \int_{\mathrm{B}(x,t)} f(w) \lambda_{\mathrm{Leb}}(\mathrm{d}w) . \tag{4}$$

10. Montrer que pour tout $t > 0$ et $k \in \mathbb{N}$,
$$\mathbb{P}_n\left(|X_{(k)}(x) - x| > t\right) \leq \mathbb{P}_n\left(\frac{1}{n} \sum_{i=1}^n \left(\mathbb{1}_{\mathrm{B}(x,t)}(X_i) - m_t(x)\right) < \frac{k}{n} - m_t(x)\right) . \tag{5}$$

On considère une suite d'entiers $(k_n)_{n \geq 0}$ telle que
$$\lim_{n \longrightarrow +\infty} k_n = +\infty \qquad \text{et} \qquad \lim_{n \longrightarrow +\infty} \frac{k_n}{n} = 0 .$$

Considérons l'intervalle
$$I_n = \left[\left(\frac{2}{m\mu} \frac{k_n}{n} \right)^{1/d}, \delta_0 \right] ,$$
où δ_0 est donné dans la question (2). Notons que cet intervalle n'est pas vide pour n assez grand car $\lim_{n \to \infty} k_n/n = 0$.

11. Démontrer que, sous l'hypothèse $(\mathbf{H}_\mu)$, pour tout $t \in I_n$ et $x \in \mathrm{supp}(f)$,
$$\mathbb{P}\left(|X_{(k_n)}(x) - x| > t\right) \leq \mathbb{P}\left(\frac{1}{n} \left| \sum_{i=1}^n \left(\mathbb{1}_{\mathrm{B}(x,t)}(X_i) - m_t(x)\right) \right| > \frac{m_t(x)}{2}\right) .$$

12. En déduire que, sous l'hypothèse $(\mathbf{H}_\mu)$, pour tout $t \in I_n$ et $x \in \mathrm{supp}(f)$,

$$\mathbb{P}\left(|X_{(k_n)}(x) - x| \geq t\right) \leq 2\mathrm{e}^{-3k_n/14} . \tag{6}$$

13. Sous les hypothèses $(\mathbf{H}_\alpha)$, $(\mathbf{H}_\mu)$ et $(\mathbf{H}_L)$, démontrer que pour n assez grand, pour $t_n = \left(\frac{2}{m\mu}\frac{k_n}{n}\right)^{1/d}$, une constante $C > 0$ existe telle que :

$$\mathscr{E}(\Phi_{n,k}) \leq C\left(\varepsilon_n^{1+\alpha} + e^{-k_n\varepsilon_n^2/2} + e^{-3k_n/14}\right) + \mathbb{1}_{\{\varepsilon_n < 2Lt_n\}} .$$

14. Conclure que le choix optimal de la méthode des k plus proches voisins est obtenue en choisissant :

$$k_n \sim c\log(n)^{\frac{d}{d+2}} n^{\frac{2}{d+2}} .$$

Donner alors une borne de l'excès de risque ainsi obtenue.

15. Simulations.
    (a) Dans un contexte Gaussien d'analyse linéaire discriminante avec covariance I_d, restreinte au compact $[-1,1]^d$, et en choisissant comme centres des deux distributions gaussiennes sous-jacentes $(-1/2,\ldots,-1/2)$ et $(1/2,\ldots,1/2)$, déterminer le classifieur Bayésien.
    (b) Illustrer l'efficacité de l'algorithme des k plus proches voisins.
    (c) Illustrer la dégradation de la vitesse lorsque la dimension augmente.
    (d) Montrer numériquement que l'algorithme de LDA est plus efficace dans ce contexte.