Introduction à l'apprentissage statistique

Examen partiel

Durée : 1h30 - documents non autorisés

Rappels/Notations

- La fonction indicatrice $\mathbb{I}\{\Omega\}$ prend la valeur 1 si Ω est vrai, et 0 sinon.
- Soit X une variable aléatoire réelle et continue. La densité f de X est une fonction positive telle que $\mathbb{P}\{X \in [a,b]\} = \int_a^b f(t)\,dt$ pour tout $a < b$. La fonction de répartition de X est définie par $F(x) = \mathbb{P}\{X \leq x\}$.
- On note $\mathcal{U}([a,b])$ la loi uniforme sur l'intervalle $[a,b]$ définie par la fonction de densité $f(x) = \big(1/(b-a)\big)\ \mathbb{I}\{x \in [a,b]\}$.
- On note $\mathcal{L}(X \mid Y)$ la loi conditionnelle de X sachant Y.
- Dans le modèle de classification binaire, on note $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$ la probabilité a posteriori, et $f^*(x) = \mathbb{E}(Y \mid X = x)$ la fonction de régression de Y sur X.

Exercice 1 - On se place dans le cadre du modèle de classification où X est un vecteur aléatoire sur $\mathbb{R}^d$ et Y est une variable aléatoire à valeurs dans $\{-1,+1\}$.

1. Trouver la relation entre la probabilité a posteriori $\eta(x)$ et la fonction de régression $f^*(x)$.
2. Montrer que la fonction de régression f^* minimise le risque quadratique moyen :
$$\mathbb{E}\big((Y - f(X))^2\big)\ .$$
3. Déterminer en fonction de η la fonction f_1^* minimisant le critère :
$$A_1(f) = \mathbb{E}\big(\log(1 + e^{-Yf(X)})\big)$$
parmi les fonctions $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$.
4. Même question dans le cas du critère $A_2(f) = \mathbb{E}\big(\max\{0, 1 - Yf(X)\}\big)$.
5. Justifier l'intérêt d'utiliser de tels critères pour le problème considéré.

Exercice 2 - On considère un modèle de classification où le couple aléatoire (X,Y) est de loi P décrite par :

$$\mathcal{L}(X \mid Y = 0) = \mathcal{U}\big([-\theta,\theta]\big)\ ;\ \mathcal{L}(X \mid Y = 1) = \mathcal{U}\big([0,1]\big)\ ;\ p = \mathbb{P}\{Y = 1\}$$

où $p, \theta \in (0,1)$ sont fixés. Calculer la fonction de régression correspondante en fonction de p, θ. Donner l'application numérique pour $\theta = 1/3$. En déduire l'erreur minimale pour la loi considérée (appelée erreur de Bayes).

Exercice 3 - On se place dans le cadre du modèle de classification où X est une variable aléatoire de loi $\mathcal{U}\big([0,1]\big)$ et Y est une variable aléatoire à valeurs dans $\{-1,+1\}$. On suppose ici que $\eta(x) = 1/4$, pour tout $x \in [0,1/2]$ et $\eta(x) = 5/6$, pour tout $x \in]1/2,1]$.

1. Calculer $\mathbb{P}\{X \in [a,b], Y = +1\}$ pour $0 < a < b < 1$.
2. Soit la base d'apprentissage :

$$(X_1 = 0,4\ ;\ Y_1 = -1); (X_2 = 0,2\ ;\ Y_2 = -1); (X_3 = 0,5\ ;\ Y_3 = -1);$$
$$(X_4 = 0,2\ ;\ Y_4 = +1); (X_5 = 0,3\ ;\ Y_5 = +1); (X_6 = 0,8\ ;\ Y_6 = -1)\ .$$

   Calculer le classifieur $\widehat{g}_k$ des k-plus proches voisins, dans le cas où $k = 1$, puis $k = 3$.
3. Calculer, dans chaque cas, son erreur de classification au sens de $L(g) = \mathbb{P}\{Y \neq g(X)\}$ lorsque $g : [0,1] \to \{-1,+1\}$ est un classifieur.
4. Calculer l'erreur de Bayes $\min_g L(g)$ et commenter son écart avec l'erreur de classification $L(\widehat{g}_k)$ des classifieurs obtenus par la méthode des k-plus proches voisins. Quelle est la cause de ce résultat ?

Exercice 4 - On se place dans le modèle de classification binaire $(X,Y) \sim P$ où P est une loi sur $\mathbb{R} \times \{0,1\}$. On représente la loi conditionnelle de X sachant Y par les deux fonctions de répartition

$$F_y(x) = \mathbb{P}\{X \leq x \mid Y = y\}$$

pour $x \in \mathbb{R}$, $y \in \{0,1\}$. On considère la famille $\mathcal{G}_L$ des classifieurs linéaires sur $\mathbb{R}$ de la forme :

$$g(x) = g_{(x_0,y_0)}(x) = \begin{cases} y_0 & \text{si } x \leq x_0 \\ 1 - y_0 & \text{sinon}\ , \end{cases}$$

avec $x_0 \in \mathbb{R}$ et $y_0 \in \{0,1\}$.

1. Exprimer l'erreur de classification $L(g) = \mathbb{P}\{Y \neq g(X)\}$ pour les éléments de $\mathcal{G}_L$ en fonction des lois conditionnelles de X sachant Y. Pour un élément $g_{(x_0,y_0)}$, on note $L(x_0,y_0) = L(g_{(x_0,y_0)})$ et on pose $L_0 = \inf_{(x_0,y_0)} L(x_0,y_0)$.
2. En considérant les points $(x_0,y_0) = (-\infty,0)$ et $(x_0,y_0) = (-\infty,1)$, montrer que $L_0 \leq 1/2$.
3. On note l'erreur de Bayes $L^* = \inf_g L(g)$. Montrer que $L_0 = 1/2$ si et seulement si $L^* = 1/2$.
4. On rappelle que $\min(a,b) = (a + b - |a - b|)/2$. On suppose que $\mathbb{P}\{Y = 1\} = 1/2$. Trouver la relation entre L_0 et $\sup_x |F_1(x) - F_0(x)|$.
