Apprentissage statistique et grande dimension

TD 3 : classification supervisée

Exercice 1 (Classifieur bayésien)

Soit (X, Y) un couple de loi P avec $Y \in \{0, 1\}$. Soit g^* le classifieur de Bayes, et R^* son risque de classification. On note également $p = \mathbb{P}(Y = 1)$.

1. Montrer que $R^* \leq \min\{p, 1-p\}$.
2. Montrer que si X et Y sont indépendants, $R^* = \min\{p, 1-p\}$.
3. Fabriquer un exemple où $R^* = \min\{p, 1-p\}$ et où X et Y ne sont pas indépendants. Indication : considérer X, Y binaires dépendants.

Exercice 2 (Classifieur bayésien (suite))

On suppose que X admet une densité f par rapport à la mesure de Lebesgue de $\mathbb{R}^d$ et l'on rappelle que

$$R^* = \mathbb{E}\left[\min\{\eta(X), 1-\eta(X)\}\right],$$

où $\eta(x) = \mathbb{P}(Y = 1|X = x)$ est la fonction de régression. Soit f_0 (resp. f_1) la densité de X conditionnellement à $Y = 0$ (resp. $Y = 1$) c'est-à-dire que pour $j = 0, 1$, pour toute fonction borélienne φ, $\mathbb{E}[\varphi(X)|Y = j] = \int_{\mathbb{R}^d} \varphi(x) f_j(x) dx$. Soit $p = \mathbb{P}(Y = 1)$.

0. Montrer que
   (a)
   $$f(x) = p f_1(x) + (1-p) f_0(x).$$
   (b) En déduire une écriture de $\eta(x)$ en fonction de f_0, f_1 et p.
1. Écrire R^* en fonction de f_0, f_1 et p.
2. Dans le cas $p = 1/2$, en déduire que
   $$R^* = \frac{1}{2} - \frac{1}{4} \int_{\mathbb{R}^d} |f_0(x) - f_1(x)| dx.$$
   On pourra utiliser la formule $\min\{a, b\} = \frac{a+b}{2} - \frac{|a-b|}{2}$.

Exercice 3 (Un classifieur universellement consistant)

Soit (X, Y) un couple de loi $\mathbb{P}$ avec $X \in \{1, ..., p\}$ et $Y \in \{0, 1\}$. On note toujours g^* le classifieur de Bayes et R^* son risque de classification. On se propose d'étudier la règle de classification suivante, étant donné $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$,

$$\hat{g}(x) = \mathbf{1}_{\{\hat{\eta}(x) > \frac{1}{2}\}},$$

avec

$$\hat{\eta}(x) = \frac{1}{\text{card}\{i, X_i = x\}} \sum_{i, X_i = x} Y_i$$

et $\hat{\eta}(x) = 0$ si $\text{card}\{i, X_i = x\} = 0$.

1. Déterminer $\hat{\eta}$ et $\hat{g}$ sur les données suivantes $\mathcal{D}_7$:

i	1	2	3	4	5	6	7
X_i	3	4	4	3	3	4	5
Y_i	0	1	0	1	0	1	0

2. Montrer que $\hat{g}$ est universellement consistante.

Exercice 4 (Classification et maximum de vraisemblance)

On suppose que la loi P du couple (X, Y) est donnée de la manière suivante : $Y \sim \mathcal{B}(p)$ pour un paramètre $p \in]0, 1[$ puis $X \in \mathbb{R}^d$ est donnée par sa loi conditionnelle sachant Y :

$$X|Y \sim \mathcal{N}(V_Y, \Sigma)$$

où Σ est une matrice définie positive et V_0, V_1 sont deux vecteurs distincts de $\mathbb{R}^d$.

1. Donner la loi jointe de (X, Y) ainsi que les lois marginales.
2. Déterminer la fonction de régression
$$\eta(x) = E(Y|X = x).$$
3. En déduire la forme du classifieur de Bayes, g^*, et montrer que son risque de classification s'écrit
$$R^* = pP\left(Z > \delta/2 + \delta^{-1}\log(\frac{p}{1-p})\right) + (1-p)P\left(Z < -\delta/2 + \delta^{-1}\log(\frac{p}{1-p})\right)$$
où $\delta = \left\|\Sigma^{-1/2}(V_1 - V_0)\right\|$ et $Z \sim \mathcal{N}(0, 1)$.
4. En pratique, on dispose d'un n-échantillon (X_i, Y_i), $1 \leq i \leq n$, de la loi P. On suppose que l'on est dans un cas où l'on connaît p et Σ et on se propose d'estimer V_0 et V_1 par maximum de vraisemblance. Donner la forme des estimateurs $\hat{V}_0$ et $\hat{V}_1$ ainsi obtenus.
5. En déduire un estimateur de la fonction de régression, $\hat{\eta}$ et une règle de classification, $\hat{g}$.
6. Montrer que $R(\hat{g}) \to R^*$ en probabilité, quand $n \to \infty$.
7. Montrer que $\hat{g}$ n'est pas universellement consistante. Il suffira de fabriquer une autre loi P' telle que si (X_i, Y_i) et (X, Y) sont iid de loi P', alors $R(\hat{g})$ ne converge pas vers R^*.

Exercice 5 (Risque de classification pondéré)

Soit $\mathcal{D}_n = \{(X_i, Y_i), i = 1, ..., n\}$ un échantillon de copies du couple de variables aléatoires (X, Y) avec $X \in \mathbb{R}^d$ et $Y \in \{0, 1\}$. Soient $\omega_0, \omega_1 > 0$ fixé, le risque de classification pondéré R_ω d'un classifieur h est la quantité

$$R_\omega(h) = \omega_0 \mathbb{P}(Y = 0, h(X) = 1|\mathcal{D}_n) + \omega_1 \mathbb{P}(Y = 1, h(X) = 0|\mathcal{D}_n).$$

1. Soit $\eta(X) = \mathbb{E}[Y|X = x]$ la fonction de régression. Montrer que
$$\mathbb{E}[R_\omega(h)] \geq \mathbb{E}\left[\min\left\{\omega_0(1 - \eta(X)), \omega_1\eta(X)\right\}\right].$$
2. Soit
$$h^*_\omega(x) = \mathbf{1}_{\left\{\eta(x) > \frac{\omega_0}{\omega_0+\omega_1}\right\}},$$
calculer $R_\omega(h^*_\omega)$ et commenter.
3. Montrer que, pour tout classifieur h
$$\mathbb{E}[R_\omega(h)] - R_\omega(h^*_\omega) = \mathbb{E}\left[|\omega_1\eta(X) - \omega_0(1 - \eta(X)|\mathbf{1}_{\{h(X) \neq h^*_\omega(X)\}}\right].$$
4. Soit $\widehat{\eta}$ un estimateur de la fonction η et
$$h_\omega(x) = \mathbf{1}_{\left\{\widehat{\eta}(x) > \frac{\omega_0}{\omega_0+\omega_1}\right\}}.$$

(a) Montrer que $h_\omega(x) \neq h^*_\omega(x)$ implique $|\widehat{\eta}(x) - \eta(x)| \geq \left|\eta(x) - \frac{\omega_0}{\omega_0+\omega_1}\right|$.

(b) En déduire que

$$\mathbb{E}[R_\omega(h_\omega)] - R_\omega(h^*_\omega) \leq (\omega_0 + \omega_1)\mathbb{E}\left[|\widehat{\eta}(x) - \eta(x)|\right].$$

Exercice 6 (Vitesse de convergence sous hypothèse de régularité de η)

On suppose que les couples i.i.d. $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ sont à valeurs dans $[0,1]^d \times \{0,1\}$ pour $d \geq 3$. L'hypothèse sur la loi de probabilité P est que $\eta(\cdot)$ est L-Lipschitz, pour une certaine constante $L > 0$. On note μ la première marginale de P.

Soit $\hat{g}_k$ le classifieur des k-plus proches voisins : $\hat{g}_k(x) = \mathbf{1}_{\hat{\eta}(x) \geq 1/2}$ où $\hat{\eta}(x) = \frac{1}{k}\sum_{m=1}^k Y_{i_m(x)}$ et $m \mapsto i_m(x)$ est une permutation $\sigma(X_1, \ldots, X_n)$-mesurable telle que $\left\|X_{i_1(x)} - x\right\| \leq \cdots \leq \left\|X_{i_n(x)} - x\right\|$.

1. Pour $0 < \varepsilon < 1$, montrer qu'il existe une partition de $[0,1]^d$ formée de moins de $(\lfloor \frac{1}{\varepsilon} + 1 \rfloor^d)$ éléments de diamètre maximal au plus $\sqrt{d}\varepsilon$.
   On note $A_1, \ldots, A_{m_\varepsilon}$ les éléments de cette partition.
2. On rappelle que $X_{i_1(x)}$ est le plus proche voisin de x parmi $X_1, \ldots, X_n$. Montrer que
$$\forall x \in A_j, \quad P\left(\left\|X_{i_1(x)} - x\right\| > \varepsilon\sqrt{d}\right) \leq [1 - \mu(A_j)]^n.$$
3. En déduire que (X est maintenant aléatoire)
$$P\left(\left\|X_{i_1(X)} - X\right\| > \varepsilon\sqrt{d}\right) \leq m_\varepsilon \sup_{\alpha>0} \alpha\exp(-\alpha n) \leq \left(\frac{2}{\varepsilon}\right)^d \frac{1}{n}$$
4. En déduire que pour $c = \frac{4d^2}{d-2}$ on obtient
$$E(\|X_{i_1(X)} - X\|^2) \leq 2d\int_0^1 \min\left\{1, \left(\frac{2}{\varepsilon}\right)^d \frac{1}{n}\right\}\varepsilon d\varepsilon \leq \frac{c}{n^{2/d}}$$
5. Vérifier que
$$\begin{aligned} &E[(\hat{\eta}(x) - \eta(x))^2] \\ =&E\left[\left(\frac{1}{k}\sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)})\right)^2\right] + E\left[\left(\frac{1}{k}\sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)})\right)^2\right] \\ =&A(x) + B(x) \end{aligned}$$
6. Montrer que pour $m \neq m'$ (attention $i_m(x)$ et $i_{m'}(x)$ sont aléatoires) :
$$E\left[(Y_{i_m(x)} - \eta(X_{i_m(x)}))(Y_{i_{m'}(x)} - \eta(X_{i_{m'}(x)}))|X_1, \ldots, X_n\right] = 0$$
7. En déduire que $A(x) \leq \frac{1}{k}$
8. Soient $S_1, \ldots, S_k$ des sous ensembles disjoints de $\{X_1, \ldots, X_n\}$ tous de taille $\lfloor n/k \rfloor$ et $\hat{X}_j(x)$ le plus proche voisin de x dans S_j. Montrer que
$$B(x) \leq \frac{L^2}{k}\sum_{j=1}^k E\|\hat{X}_j(x) - x\|^2$$
9. En déduire que $E[B(x)] \leq cL^2\left(\frac{2k}{n}\right)^{2/d}$ et en déduire la vitesse de convergence des k-plus proches voisins.
10. Qu'obtient-on pour $d = 1$ et $d = 2$?