Apprentissage statistique: TD5
The Nearest Neighbor Rule

Emile Contal
http://econtal.perso.math.cnrs.fr/teaching

8 février 2016

Exercice 1. Soient (X, Y, U) et $D_n = \big((X_1, Y_1, U_1), \ldots, (X_n, Y_n, U_n)\big)$ des variables aléatoires telles que : X est de mesure μ sur $\mathbb{R}^d$, $Y \in \{0, 1\}$ et $\mathbf{E}[Y \,|\, X] = \eta(X)$, U est uniformément distribué sur $[0, 1]$ et indépendant de X, les triplets (X_i, Y_i, U_i) sont indépendants et identiquement distribués à (X, Y, U). Soit $x \in \mathbb{R}^d$. On peut alors définir $Y'_i(x) = \mathbb{1}_{U_i \leq \eta(x)}$. On ordonne la séquence de quadruplets $\big((X_i, Y_i, Y'_i, U_i)\big)_{i \leq n}$ par valeur croissante de $\|X_i - x\|$, que l'on note alors $\big((X_{(i)}(x), Y_{(i)}(x), Y'_{(i)}(x), U_{(i)})\big)_{i \leq n}$.

On fixe $k < n$ impair. La règle de prédiction g_n des k-plus-proches-voisins est définie comme suit :

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^k Y_{(i)}(x) > \frac{k}{2} \\ 0 & \text{sinon}\,. \end{cases}$$

Ainsi que la règle de prédiction g_n^w des k-plus-proches-voisins pondérée par un vecteur $w = w_1, \ldots, w_k$:

$$g_n^w(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^k w_i \mathbb{1}_{Y_{(i)}(x)=1} > \sum_{i=1}^k w_i \mathbb{1}_{Y_{(i)}(x)=0} \\ 0 & \text{sinon}\,. \end{cases}$$

1. Soit g'_n la règle qui suivant $\eta(X)$ et Y' :

   $$g'_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^k Y'_{(i)}(x) > \frac{k}{2} \\ 0 & \text{sinon}\,. \end{cases}$$

   On définit $L_n = \Pr[g_n(X) \neq Y \,|\, D_n]$ et $L'_n = \Pr[g'_n(X) \neq Y \,|\, D_n]$. Rappeler comment montrer que $\mathbf{E}[L'_n] - \mathbf{E}[L_n] \to 0$.

2. Soit $L_{k\mathrm{NN}} = \lim_{n \to \infty} \mathbf{E}[L_n]$. On note $p = \eta(X)$. Montrer que :

   $$L_{k\mathrm{NN}} = \mathbf{E}\Big[p + (1 - 2p) \Pr\Big[\mathcal{B}\big(k, p\big) > \frac{k}{2} \,|\, X\Big]\Big]\,,$$

   et une formule similaire (moins propre) pour L_w l'erreur asymptotique de la règle g_n^w. Simplifier la formule précédente dans le cas où $p < 1/2$ et $\Pr\Big[\sum_{i=1}^k w_i(2Y'_i - 1)\Big] = 0$.

3. Soit N_l le nombre de vecteurs $z = (z_1, \dots, z_k)$ dans $\{-1, 1\}^k$ tel que $\sum \mathbb{1}_{z_i=1} = l$ et $\sum w_i z_i > 0$. Montrer que $N_l + N_{k-l} = \binom{k}{l}$.

4. En déduire que lorsque $p < 1/2$, avec $q = 1 - p$:

$$L_w = \mathbf{E}\Big[p + (1-2p)\Big(\Pr\Big[\mathcal{B}(k,p) > \frac{k}{2} \,|\, X\Big] + \sum_{l<\frac{k}{2}} N_l(p^l q^{k-l} - p^{k-l}q^l)\Big)\Big] .$$

5. Dans le cas où $p < 1/2$, en déduire que pour tout w, $L_w \geq L_{k\mathrm{NN}}$. Donner un résultat identique dans le cas où $p > 1/2$.

6. Utiliser un raisonnement similaire pour montrer que :

$$L^\star \leq \cdots \leq L_{(2k+1)\mathrm{NN}} \leq L_{(2k-1)\mathrm{NN}} \leq \cdots \leq L_{\mathrm{NN}} \leq 2L^\star .$$

On rappelle que $L_{\mathrm{NN}} \leq 2L^\star$.

Exercice 2.

1. On se place maintenant dans le cas où k est pair, et on définit la règle suivante des $2k$-plus-proches-voisins :

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^{2k} Y_{(i)}(x) > k \\ 0 & \text{si } \sum_{i=1}^{2k} Y_{(i)}(x) < k \\ Y_{(1)}(x) & \text{sinon} . \end{cases}$$

Montrer que l'erreur asymptotique $L_{2k\mathrm{NN}}$ de cette règle vérifie :

$$L_{(2k-1)\mathrm{NN}} = L_{2k\mathrm{NN}} .$$

Exercice 3. Dans cet exercice on cherche à déterminer des bornes sur $L_{k\mathrm{NN}} - L^\star$ d'après la formule donnée en question 1.2 qui peut se réécrire avec $\epsilon = \min\{p, q\}$:

$$L_{k\mathrm{NN}} = \mathbf{E}\Big[\epsilon + (1-2\epsilon)\Pr\Big[\mathcal{B}(k,\epsilon) > \frac{k}{2} \,|\, X\Big]\Big] .$$

1. Montrer que :

$$L_{k\mathrm{NN}} - L^\star \leq \frac{1}{\sqrt{ke}} .$$

Indice : l'inégalité de Hoeffding une fois de plus.

2. Montrer que :

$$L_{k\mathrm{NN}} - L^\star \leq \sqrt{\frac{2L_{\mathrm{NN}}}{k}} .$$

Indice dans l'ordre : Markov, Cauchy-Schwarz, $\mathbb{V}[\mathcal{B}(k,\epsilon)]$, Jensen.

3. Dire ce qu'on pourrait faire pour améliorer ces résultats.