Apprentissage statistique: TD4
The Nearest Neighbor Rule

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

6 février 2015

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 $\mathbf{E}[Y \mid X] = \eta(X)$,
- 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 $\big((X_i, Y_i, Y'_i(x), U_i)\big)_{i \leq n}$ 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 :

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

On définit également la règle g'_n (inconnue en pratique car dépend de η) :

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

1. On fixe $x \in \mathbb{R}^d$. Montrer que :

$$\Pr[g_n(x) \neq g'_n(x)] \leq \sum_{i=1}^k \mathbf{E}[\big|\eta(x) - \eta(X_{(i)}(x))\big|]\,.$$

2. Soit $L_n = \Pr[g_n(X) \neq Y \mid D_n]$ et de même pour L'_n. On rappelle que $\sum_{i=1}^k \mathbf{E}[\big|\eta(X) - \eta(X_{(i)}(X))\big|] \to_{n\to\infty} 0$. Montrer que $\mathbf{E}[L'_n] - \mathbf{E}[L_n] \to 0$.

3. Dans le cas du 1-plus-proche-voisin, montrer que $\mathbf{E}[L'_n] = \mathbf{E}[2\eta(X)(1-\eta(X))]$.

4. Soit f et g deux fonctions réelles respectivement croissante et décroissante. Montrer que $\mathbf{E}[f(X)g(X)] \leq \mathbf{E}[f(X)]\mathbf{E}[g(X)]$.

5. En déduire que l'erreur asymptotique du 1-plus-proche-voisin est au pire 2 fois l'erreur de Bayes.

Exercice 2.

1. On note $p = \min\{\eta(X), 1-\eta(X)\}$ et $q = 1-p$. Montrer que :

$$\lim_{n\to\infty} \mathbf{E}[L_n] = \mathbf{E}\Big[p + (1-2p)\Pr\big[\mathcal{B}\big(k, \eta(X)\big) > \frac{k}{2} \,|\, X\big]\Big] =: L_{k\mathrm{NN}}\,.$$

2. On considère maintenant la règle des plus proches voisins pondérée par un vecteur $w = w_1, \ldots, w_k$:

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

Montrer que l'erreur asymptotique L_w de cette règle peut s'écrire sous la forme :

$$L_w = \mathbf{E}\Big[p + (1-2p)\Pr\big[\sum_{i=1}^k w_i Z'_i > 0\big]\Big]\,,$$

où $Z'_i = 2Y'_i - 1$ avec Y'_i comme dans l'exercice précédent.

3. Soit N_l le nombre de vecteurs $z = (z_1, \ldots, z_k)$ 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 :

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

5. Dans le cas où $p = \eta(X)$, en déduire que pour tout w, $L_w \geq L_{k\mathrm{NN}}$. Donner un résultat identique dans le cas où $p = 1-\eta(X)$.

6. Utiliser un raisonnement similaire pour montrer que :

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