Apprentissage statistique: TD6
Consistency of Voting Classifiers

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

29 février 2016

Notations et définitions.

Soit $(X,Y),(X_1,Y_1),\dots,(X_n,Y_n)$ des paires iid de variables aléatoires sur $\mathbb{R}^d \times \{0,1\}$ telles que X soit de mesure μ et $\Pr[Y \mid X = x] = \eta(x)$. On note $D_n = \big\{(X_i,Y_i)\big\}_{i\leq n}$ l'ensemble d'apprentissage et g_n un classifieur (où la dépendance en D_n est implicite). On notera $L(g_n) = \Pr[g_n(X) \neq Y \mid D_n]$ l'erreur de g_n.

Une suite de classifieur $(g_n)_n$ est *consistante* pour une distribution de (X,Y) lorsque $\mathbf{E}[L(g_n)] \to L^\star$ où $L^\star$ est l'erreur du classifieur optimal. Une suite de classifieur est *universellement consistante* si elle est consistante pour toute distribution de (X,Y).

Soit $Z_m = \big(U_1,\dots,U_m\big)$ une suite de variables aléatoires iid et $x, U \mapsto g_n(x,U)$ un classifieur *randomisé*. On définit $\bar{g}_n(x,Z_m) = \mathbb{1}_{\frac{1}{m}\sum_{j=1}^m g_n(x,U_j) \geq \frac{1}{2}}$ le *classifieur par votes*.

Exercice 1. (Consistency is Preserved by Voting)

Soit (g_n) une séquence de classifieurs randomisés consistants pour une certaine distribution de (X,Y). Montrer que le classifier par vote $\bar{g}_n$ est également consistant.

Théorème 1 (Devroye, 1996)**.** *Soit $A_n = \big(A_n^1, A_n^2, \dots\big)$ une partition de $\mathbb{R}^d$ et $A_n(x)$ la cellule de A_n contenant x. Soit (g_n) une suite de classifieurs votants sur les cellules de la partition :*

$$g_n = 1 \text{ ssi } \sum_{i=1}^n \mathbb{1}_{Y_i=1}\mathbb{1}_{X_i\in A_n(x)} \geq \sum_{i=1}^n \mathbb{1}_{Y_i=0}\mathbb{1}_{X_i\in A_n(x)}\,.$$

On a que (g_n) est consistante lorsque :

(1) $\sup_{x,y\in A_n(X)} \|x-y\| \to 0$ *en probabilité,*

(2) $N_n(X) := \sum_{i=1}^n \mathbb{1}_{X_i\in A_n(X)} \to \infty$ *en probabilité.*

Exercice 2. (Hyper-cube Histograms)

Pour cet exercice, on considère la partition A_n de $\mathbb{R}^d$ définie par des hyper-cubes de côté h_n, et le classifieur g_n votant sur A_n.

1. Soit $M > 0$ et A_n^j tel que $\mu(A_n^j) > \frac{2M}{n}$. On note $\mu_n(A_n^j) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A_n^j}$. Montrer que :

$$\Pr\left[X \in A_n^j, N_n(X) \leq M\right] \leq \mu(A_n^j) \Pr\left[\mu_n(A_n^j) - \mathbf{E}\left[\mu_n(A_n^j)\right] \leq -\frac{1}{2}\mu(A_n^j)\right].$$

2. En utilisant l'inégalité de Chebyshev, en déduire que pour un tel A_n^j :

$$\Pr\left[X \in A_n^j, N_n(X) \leq M\right] \leq \frac{4}{n}.$$

3. En déduire que lorsque $nh_n^d \to \infty$, alors pour tout $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[N_n(X) \leq M\right] < \epsilon.$$

4. En déduire que lorsque $nh_n^d \to \infty$ et $h_n \to 0$, la règle g_n est universellement consistante.

Exercice 3. (Random Forests)

On suppose que le support de μ est $\mathcal{X} = [0,1]^d$. Un *arbre de décision randomisé* $g_n(x, U)$ est un classifieur qui construit une partition hiérarchique de $\mathcal{X}$ sous forme d'arbre binaire dont les noeuds correspondent à une partie rectangulaire de $\mathcal{X}$. La racine de l'arbre est associée à $\mathcal{X}$ en entier. A chaque noeud, on découpe le rectangle correspondant en deux sous-rectangles suivant une certaine coordonnée et un certain seuil. L'ensemble des cellules associées aux feuilles maintient en permanence une partition de $\mathcal{X}$. Un *arbre de décision purement randomisé* de taille k est construit en répétant k fois la procédure :

- choisir une feuille uniformément parmi les feuilles de l'arbre,
- choisir une coordonnée uniformément parmi les d variables,
- choisir un seuil uniformément dans la longueur du rectangle correspondante,
- découper la feuille suivant la variable et le seuil en deux sous feuilles.

Une *forêt aléatoire* est le classifieur par votes associé à une séquence d'arbres de décision $\left(g_n(x, U)\right)_n$.

Montrer qu'une forêt purement aléatoire est consistante lorsque $k \to \infty$ et $\frac{k}{n} \to 0$.

Théorème 2 (Stone, 1977)**.** *Soient $W_{ni}(x)$ des poids positifs dépendants de $X_1, \ldots, X_n$, tels que $\sum_{i=1}^n W_{ni}(x) = 1$. Soit (g_n) une suite de classifieurs pondérés :*

$$g_n = 1 \text{ ssi } \sum_{i=1}^n \mathbb{1}_{Y_i=1} W_{ni}(x) \geq \sum_{i=1}^n \mathbb{1}_{Y_i=0} W_{ni}(x)\,.$$

On a que (g_n) est universellement consistante lorsque :

(1) Il existe une constante c telle que pour toute fonction f mesurable positive d'espérance finie, $\mathbf{E}\big[\sum_{i=1}^n W_{ni}(X) f(X_i)\big] \leq c\mathbf{E}[f(X)]$,

(2) Pour tout $a > 0$, $\mathbf{E}\big[\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{\|X_i - X\| > a}\big] \to 0$,

(3) $\mathbf{E}\big[\max_{1\leq i\leq n} W_{ni}(X)\big] \to 0$.

Exercice 4. (Consistency by Randomization)

On se place pour simplifier dans le cas $d = 1$ et on considère que le support de X n'est pas atomique. Pour un x fixé, soient $X_{(1)}(x), \ldots, X_{(n)}(x)$ les points $X_1, \ldots, X_n$ triés selon leur distance à x. Soient $U = \big(U_1, \ldots, U_n\big)$ des variables aléatoires uniformes sur $[0, 1]$. On définit une règle du plus proche voisin randomisée comme suit :

$$g_n(x, U) = Y_{(i_x)} \text{ où } i_x = \operatorname*{argmin}_i \max(i, mU_i)\,,$$

où $m \leq n$ est un paramètre.

Soit $\widetilde{g}_n(x) = \mathbb{1}_{\mathbf{E}_U g_n(x,U) \geq 1/2}$ le classifieur par espérance, c'est à dire la limite du classifieur par votes de $g_n(\cdot, \cdot)$. Montrer que $\widetilde{g}_n$ est consistant lorsque $m \to \infty$ et $\frac{m}{n} \to 0$. Vous pourrez utiliser le fait que pour X indépendant des X_i, $\big\|X_{(k)}(X) - X\big\| \to 0$ avec probabilité 1 lorsque $\frac{k}{n} \to 0$.