Excès d'erreur, Consistance et Règle d'histogramme

Exercice 1 On se place dans le cadre de classification binaire où on définit pour un couple aléatoire $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ la probabilité conditionnelle $\eta(X) = \mathbb{P}(Y = 1|X)$. Soit $g^* : \mathbb{R}^d \mapsto \{0, 1\}$ le classifieur optimal qui minimise l'erreur de Bayes.

1. Pour tout classifieur g, on appelle L_g son erreur de Bayes

$$L_g = \mathbb{P}(g(X) \neq Y).$$

Montrer

$$L_g - L_{g^*} = 2\mathbb{E}\Big[\big|\eta(X) - \frac{1}{2}\big|1_{g(X)\neq g^*(X)}\Big].$$

2. (Plug-In Rule) Soit un classifieur g défini par

$$g(x) = 1_{\tilde{\eta}(x)>\frac{1}{2}} \tag{1}$$

avec $\tilde{\eta}(\cdot)$ un estimateur de $\eta(\cdot)$. Montrer et interpréter

$$L_g - L_{g^*} \leq 2\|\eta(X) - \tilde{\eta}(X)\|_1 \leq 2\|\eta(X) - \tilde{\eta}(X)\|_2.$$

3. Soit $\tilde{\eta}_1$ (resp. $\tilde{\eta}_0$) une approximation de η (resp. $1 - \eta$) qui ne respecte pas nécessairement la contrainte $\tilde{\eta}_1 + \tilde{\eta}_0 = 1$. On construit un nouveau classifieur

$$g(x) = \begin{cases} 0 & \text{si } \tilde{\eta}_1(x) \leq \tilde{\eta}_0(x) \\ 1 & \text{sinon.} \end{cases}$$

Montrer

$$L_g - L_{g^*} \leq \|1 - \eta(X) - \tilde{\eta}_0(X)\|_1 + \|\eta(X) - \tilde{\eta}_1(X)\|_1.$$

4. Montrer que la condition $L_{g^*} = 0$ conduit à une meilleure vitesse de convergence

$$\forall p \geq 1, \;\; L_g \leq 2^p \|\eta(X) - \tilde{\eta}(X)\|_p^p.$$

5. Soit une suite d'estimateurs η_n vérifiant

$$\lim_{n\to+\infty} \|\eta_n - \eta\|_2 = 0.$$

A tout η_n, on associe un classifieur g_n en utilisant la définition (1). Montrer

$$\lim_{n\to\infty} \frac{L_{g_n} - L_{g^*}}{\|\eta_n - \eta\|_2} = 0.$$

La régression est donc bien plus difficile que la classification.

Exercice 2 Reprenons le modèle probabiliste précédent. On note $\mu(\cdot)$ la probabilité induite par X sur $\mathbb{R}^d$. Soit une partition $\mathcal{P}_n = \{A_{n1}, A_{n2}, \cdots\}$ de l'espace $\mathbb{R}^d$ où A_{ni} représente une hypercube indicée par $i \in \mathbb{N}$ de côté h_n vérifiant

$$\lim_n h_n = 0 \quad \text{et} \quad \lim_n n h_n^d = +\infty.$$

L'objectif de cet exercice est d'établir la consistance forte de la règle d'histogramme.

1. Rappeler les deux notions de consistance d'une règle de classification.
2. Rappeler la règle d'histogramme.
3. Soit $A_n(x)$ l'hypercube à laquelle appartient le point $x \in \mathbb{R}^d$ et

$$\eta_n(x) = \frac{\sum_{i=1}^n Y_i 1_{X_i \in A_n(x)}}{n\mu(A_n(x))}.$$

   Montrer

$$\lim_{n\to\infty} \mathbb{E} \int |\eta(x) - \eta_n(x)| \mu(dx) = 0.$$

4. Montrer, à l'aide d'inégalité de McDiarmid,

$$\forall \epsilon > 0,\ \mathbb{P}\Big(\int |\eta(x) - \eta_n(x)| \mu(dx) - \mathbb{E} \int |\eta(x) - \eta_n(x)| \mu(dx) > \epsilon \Big) \leq e^{-n\epsilon^2/2}.$$

5. Établir la concentration

$$\forall \epsilon > 0,\ \exists N,\ \forall n > N,\ \mathbb{P}(L_{g_n} - L_{g^*} > \epsilon) \leq 2e^{-n\epsilon^2/32}.$$

6. Conclure.