Exercice 1 Soit un espace probabilisé comportant un couple aléatoire $(X,Y) \in \mathbb{R}^d \times \{0,1\}$ où on définit $\eta(X) = \mathbb{P}(Y=1|X)$. Soit $g : \mathbb{R}^d \mapsto \{0,1\}$ un classifieur quelconque et g^* le classifieur optimal (de Bayes).

1. Soit L_g l'erreur de classification induite par le classifieur g. Montrer que la différence entre L_g et l'erreur optimale L_{g^*} est

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

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

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

où $\tilde{\eta}(\cdot)$ est un estimateur de $\eta(\cdot)$. Montrer

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

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

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

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 $L_{g^*} = 0$ permet de déduire une meilleure vitesse de convergence

$$L_g \leq 4\|\eta(X) - \tilde{\eta}(X)\|_2^2$$

5. Soit une suite d'estimateurs η_n vérifiant

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

et on construit une suite de classifiers g_n de la même manière qu'avant. Montrer

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

Ce résultat illustre que la régression est plus difficile que la classification.

Exercice 2 Dans cet exercice, on va montrer la consistance forte de la règle d'histogramme. Pour cela, on reprend le modèle probabiliste de l'exercice précédent et 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 $\forall i \in \mathbb{N}, A_{ni}$ représente une hypercube de côté h_n vérifiant $\lim_n h_n = 0$ et $\lim_n nh_n^d = +\infty$. On note $A_n(x)$ l'hypercube à laquelle appartient le point $x \in \mathbb{R}^d$.

1. Rappeler la différence entre les deux notions de consistances d'une règle de classification.
2. Rappeler la règle d'histogramme.
3. On note

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

   Montrer

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

4. A l'aide d'inégalité de McDiarmid, montrer

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

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.