TD 1

1. Soit (X, Y) un couple de variable aléatoire tel que X suit une loi uniforme sur $[0, 1]$ et $Y \in \{0, 1\}$. On suppose que $\eta(X) = \mathbb{E}[Y|X]$ est définie par

$$\eta(X) = \frac{1}{5}\mathbb{1}_{\{X \leq 1/4\}} + \frac{2}{5}\mathbb{1}_{\{1/4 < X \leq 1/2\}} + \frac{3}{5}\mathbb{1}_{\{1/2 < X \leq 3/4\}} + \frac{4}{5}\mathbb{1}_{\{3/4 < X\}}.$$

1. Donner une expression simple de $g^*(X) = \mathbf{1}_{\{\eta(X) \geq 1/2\}}$.
2. Calculer $\mathbb{P}[g^*(X) \neq Y]$.

2. On considère μ_1 et μ_0 deux réels tel que $\mu_1 > \mu_0$. Soit (X, Y) un couple de variables aléatoires tel que Y suit une loi de Bernouilli de paramètre $p \in (0, 1)$ et

$$X|Y = 0 \sim \mathcal{N}(\mu_0, \sigma^2) \quad \text{et} \quad X|Y = 1 \sim \mathcal{N}(\mu_1, \sigma^2).$$

1. Déterminer la densité de X par rapport à la mesure de Lebesgue.
2. Déterminer une expression de $\eta(x) = \mathbb{P}[Y = 1|X = x]$.
3. Montrer que le classifieur de Bayes g^* vérifie

$$g^*(x) = \mathbb{1}_{\{ax \geq b\}},$$

avec

$$a = \frac{(\mu_1 - \mu_0)}{\sigma^2} \text{ et } b = \log\left(\frac{1-p}{p}\right) + \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}$$

4. On considère le cas où $p = 1/2$.
   (a) Justifier que $2R(g^*) = \mathbb{P}[g^*(X) = 1|Y = 0] + \mathbb{P}[g^*(X) = 0|Y = 1]$
   (b) En déduire que

$$R(g^*) = 1 - \Phi\left(\frac{\mu_1 - \mu_0}{2\sigma}\right),$$

où Φ désigne la fonction de répartition d'une $\mathcal{N}(0, 1)$.

3. Le but de l'exercice est de démontrer que l'algorithme des k plus proches voisins employé avec $k = 1$ n'est pas consistant. Pour cela on considère un vecteur aléatoire (X, Y) tel que $X \in \mathcal{X} = [0, 1]$ et $Y \in \{0, 1\}$. On considère également $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, n vecteurs indépendants et de même loi que (X, Y). On suppose que

$$\eta(x) = \mathbb{P}[Y = 1|X = x] = \frac{3}{4}, \quad \forall x \in \mathcal{X}.$$

1. Rappeler la définition d'une règle de classification.
2. Pour toute règle de classification g, montrer que

$$R(g) = \mathbb{P}[g(X) \neq Y] = \mathbb{E}[\eta(X)] + \mathbb{E}[g(X)(1 - 2\eta(X))].$$

   En déduire que $R(g) = \frac{3}{4} - \frac{1}{2}\mathbb{E}[g(X)]$, pour toute règle de classification g.
3. Déterminer le classifieur de Bayes et en donner son risque. Quelle propriété fondamentale vérifie le classifieur de Bayes ?
4. Soit pour $x \in \mathcal{X}$, $\hat{g}(x) = \hat{g}(x, D_n)$ le classifieur du 1 plus proche voisin.
   Pour $i = 1, \ldots, n$, On définit la variable aléatoire

$$Z_i = \mathbb{1}_{\{X_i \text{ est le plus proche voisin de x}\}}.$$

   Montrer que

$$\mathbb{E}[\hat{g}(x)] = \sum_{i=1}^{n} \mathbb{E}[Y_i Z_i],$$

où l'espérance est par rapport à l'échantillon.

5. Montrer que $X_1, \ldots, X_n$ et Y_i sont indépendantes. En déduire que Y_i et Z_i sont indépendantes.
6. En utilisant les questions précédentes, montrer que $\mathbb{E}[R(\hat{g})] = \frac{3}{8}$. Conclure.
7. On considère le cas général où $k \in \mathbb{N}^*$. Soit pour $x \in \mathcal{X}$, $\hat{g}_k(x) = \hat{g}(x, D_n)$ le classifieur du k plus proche voisin. Montrer que $\mathbb{E}[R(\hat{g}_k)] \to 1/4$ lorsque $k \to +\infty$ et $k/n \to 0$. Conclure.

4. On considère un vecteur aléatoire (X, Y) tel que $X \in \mathcal{X} = \{x_1, x_2\}$ et $Y \in \{0, 1\}$ où x_1 et x_2 sont deux réels fixés. On suppose que $\mathbb{P}[Y = 1] \in]0, 1[$.
On rappelle qu'une règle de classification est la donnée d'une application $g : \mathcal{X} \to \{0, 1\}$. On note $\mathcal{G}$ l'ensemble des règles de classification. Pour tout $g \in \mathcal{G}$ on définit la mesure de risque R_2 par

$$R_2(g) = \mathbb{P}[g(X) \neq 1 | Y = 1] + \mathbb{P}[g(X) \neq 0 | Y = 0].$$

On définit les quantités suivantes $\pi_1 = \mathbb{P}[X = x_1 | Y = 1]$, $\pi_0 = \mathbb{P}[X = x_1 | Y = 0]$ et $\Delta = |\pi_1 - \pi_0|$.
On rappelle que $\mathbb{1}_A = 1$ si A est réalisé et 0 sinon.

1. Déterminer le cardinal de $\mathcal{G}$.
2. Soit $g \in \mathcal{G}$. On définit pour $i = 0, 1$, $A_i = \mathbb{P}[g(X) \neq i | Y = i]$.
   (a) Montrer que pour $i = 0, 1$, $A_i = \pi_i \mathbb{1}_{g(x_1) \neq i} + (1 - \pi_i)\mathbb{1}_{g(x_2) \neq i}$.
   (b) En déduire que $R_2(g) = 1 + (\pi_1 - \pi_0)\mathbb{1}_{g(x_1) \neq 1} + (\pi_0 - \pi_1)\mathbb{1}_{g(x_2) \neq 1}$.
3. Soit $h \in \mathcal{G}$ la règle de classification définie par : $h(X) = \mathbb{1}_{\pi_1 > \pi_0}\mathbb{1}_{X = x_1} + \mathbb{1}_{\pi_1 \leq \pi_0}\mathbb{1}_{X = x_2}$.
   (a) Déduire du résultat obtenu en question 2)b) que pour tout $g \in \mathcal{G}$ on a : $R_2(g) - R_2(h) = \Delta\left(\mathbb{1}_{g(x_1) \neq h(x_1)} + \mathbb{1}_{g(x_2) \neq h(x_2)}\right)$.
   (b) En déduire que $R_2(h) = \min_{g \in \mathcal{G}} R_2(g)$.
   (c) Déduire de la question 2)b) que $\Delta = 1 - R_2(h)$.
4. Dans cette question la loi du vecteur (X, Y) est définie par $\mathbb{P}[X = x_1] = 1/2$, $\mathbb{P}[Y = 1 | X = x_1] = 1/10$ et $\mathbb{P}[Y = 1 | X = x_2] = 3/10$.
   (a) Déterminer la loi de Y.
   (b) Déterminer $h(x_1)$ et $h(x_2)$.
   (c) Rappeler la définition du classifieur de Bayes que l'on note h^*
   (d) Quelle propriété fondamentale vérifie h^*
   (e) Déterminer h^*.
   (f) Calculer $R_2(h^*) - R_2(h)$.