Correction exercice 3 TD 1

1. Une règle de classification est une fonction $g : \mathcal{X} \to \{0, 1\}$.
2. On a :

$$\begin{aligned} R(g) &= \mathbb{P}[g(X) \neq Y] = \mathbb{E}[\mathbb{1}_{g(X)=1}\mathbb{1}_{Y=0}] + \mathbb{E}[\mathbb{1}_{g(X)=0}\mathbb{1}_{Y=1}] \\ &= \mathbb{E}[\mathbb{1}_{g(X)=1}(1-\eta(X))] + \mathbb{E}[(1-\mathbb{1}_{g(X)=1})\eta(X)] = \mathbb{E}[\eta(X)] + \mathbb{E}[\mathbb{1}_{g(X)=1}(1-2\eta(X))] \\ &= \mathbb{E}[\eta(X)] + \mathbb{E}[g(X)(1-2\eta(X))] \end{aligned}$$

   Puisque $\eta(X) = 3/4$, on en déduit

$$R(g) = \frac{3}{4} + \mathbb{E}\left[g(X)\left(1 - 2 \times \frac{3}{4}\right)\right] = \frac{3}{4} - \frac{1}{2}\mathbb{E}[g(X)].$$

3. $g^*(x) = \mathbb{1}_{\eta(x)>1/2}$, $E[g^*(X)] = \mathbb{P}[\eta(x) > 1/2] = 1$ et donc

$$R(g^*) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}.$$

   Le classifieur de Bayes vérifie la propriété suivante :

$$\forall g \in \mathcal{G}, R(g^*) \leq R(g).$$

4. Par définition, $\hat{g}(x) = \sum_{i=1}^n Z_i Y_i$. On remarque que $\hat{g}(x)$ dépend uniquement de l'échantillon d'apprentissage $\mathcal{D}_n$. Soit X une variable aléatoire indépendante de $\mathcal{D}_n$. La loi de $\hat{g}(X)$ sachant $\mathcal{D}_n$ est aléatoire en X et

$$\mathbb{E}[\hat{g}(X) \mid \mathcal{D}_n] = \sum_{i=1}^n Y_i \mathbb{E}[Z_i \mid \mathcal{D}_n].$$

   Or $\mathbb{E}[Z_i \mid \mathcal{D}_n]$ est la probabilité qu'un X_i soit le plus proche voisin d'un X indépendant de X_i et vaut $1/n$.

5. En reprenant le raisonnement de la question 2., on montre facilement qu'on a :

$$R(\hat{g}) := \mathbb{P}[\hat{g}(X) \neq Y \mid \mathcal{D}_n] = \frac{3}{4} - \frac{1}{2}\mathbb{E}[\hat{g}(X) \mid \mathcal{D}_n] = \frac{3}{4} - \frac{1}{2n}\sum_{i=1}^n Y_i.$$

   Puisque $\mathbb{P}[Y = 1 \mid X] = 3/4$, on a $\mathbb{E}[Y] = \mathbb{P}[Y = 1] = 3/4$ et donc

$$\mathbb{E}[R(\hat{g})] = \frac{3}{4} - \frac{1}{2} \times \frac{3}{4} = \frac{3}{8}.$$

   L'algorithme des 1 plus proches voisins n'est pas consistant !

6. On peut définir $\hat{g}_k(X) = \mathbb{1}\{\sum_{i=1}^n Y_i Z_i/k \geq 1/2\}$, avec $Z_i = \mathbb{1}\{X_i \text{ est un des plus proches voisins de } X\}$. Sachant X, $\sum_{i=1}^n Y_i Z_i$ est une somme de k v.a. indépendantes de Bernoulli de paramètres $\eta(X) = 3/4$. Puisque $\eta(X)$ ne dépend pas de X, $\sum_{i=1}^n Y_i Z_i$ est indépendant de X et suit la loi $\mathcal{B}(k, 3/4)$ (loi Binomiale). On a donc :

$$\mathbb{E}[\hat{g}_k(X)] = \mathbb{P}[\mathcal{B}(k, 3/4) \geq k/2].$$

   On rappelle alors que

$$F(t, m, p) := \mathbb{P}[\mathcal{B}(m, p) \leq t] = \sum_{j=0}^{\lfloor t \rfloor} \binom{m}{j} p^j (1-p)^j,$$

et on peut facilement montrer qu'on a

$$F(m-t,m,1-p) = 1 - F(t,m,p), \text{ pour } t \in \mathbb{R}_+ \setminus \mathbb{N}.$$

On en déduit

$$\mathbb{E}[\hat{g}_k(X)] = 1 - F(k/2,k,3/4) = F(k/2,k,1/4) = 1 - \mathbb{P}[\mathcal{B}(k,1/4) \geq k/2].$$

On définit $W_i \sim \mathcal{B}e(1/4)$ (loi de Bernoulli), $i = 1,\ldots,k$, i.i.d et on écrit

$$\begin{aligned}\mathbb{P}[\mathcal{B}(k,1/4) \geq k/2] &= \mathbb{P}\left[\sum_{i=1}^{k}\left(W_i - \frac{1}{4}\right) \geq \frac{k}{2} - k \times \frac{1}{4}\right] \\ &= \mathbb{P}\left[\sum_{i=1}^{k}\left(W_i - \frac{1}{4}\right) \geq \frac{k}{4}\right].\end{aligned}$$

On utilise alors l'inégalité de Hoeffding (voir cours). On a : $a_i = -1/4 \leq W_i - 1/4 \leq b_i = 3/4$, $\sum_i (b_i - a_i)^2 = k$ et donc

$$\mathbb{P}\left[\sum_{i=1}^{k}\left(W_i - \frac{1}{4}\right) \geq \frac{k}{4}\right] \leq \exp\left(-\frac{2k^2}{16k}\right) = \exp\left(-\frac{k}{8}\right).$$

Au final :

$$\mathbb{E}[\hat{g}_k(X)] \geq 1 - \exp\left(-\frac{k}{8}\right) \xrightarrow[k\to\infty]{} 1.$$

On a donc $\mathbb{E}[\hat{g}_k(X)]$ qui tend vers 1 quand k tend vers l'infini et $\mathbb{E}[R(\hat{g}_k)]$ qui tend vers

$$\frac{3}{4} - \frac{1}{2} \times 1 = \frac{1}{4}$$

quand k tend vers l'infini. On atteint donc le risque oracle, l'algorithme des k plus proches voisin est consistant quand k tend vers l'infini.