Année 2017-2018 **M1 MINT**

Introduction au Machine learning : examen

Exercice 1

On a observé les données suivantes : les features sont dans $\mathbb{R}^2$ et les labels sont dans $\{\text{rouge}, \text{bleu}\}$.

1. Donner les valeurs de l'erreur empirique associée à la perte 0/1 des classifieurs construits par
   - l'algorithme des 1-plus proche voisins (1-NN)
   - l'algorithme des 3-plus proche voisins (3-NN).
2. Où metteriez vous le premier "split" d'un arbre de décision ? (vous pouvez le dessiner sur la figure)
3. A partir de quelle profondeur a-t-on un arbre d'erreur empirique nulle ?
4. Tracer l'hyperplan (ici la droite) associé à l'algorithme linear-SVM.

Exercice 2

On définit $r^\star$ comme le rectangle $[l, r] \times [b, t]$. On considèrent des données $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$ avec des couples features/label (X_i, Y_i) i.i.d., $X_i \in \mathbb{R}^2$ et $Y_i \in \{1 = \text{rouge}, -1 = \text{bleu}\}$ dont la loi vérifie

$$\begin{aligned}
&\mathbb{P}(Y_i = 1 | X_i \in r^\star) = 1 \\
&\mathbb{P}(Y_i = 1 | X_i \notin r^\star) = 0 \\
&\mathbb{P}(X_i \in r^\star) > \epsilon \text{ pour un } \epsilon > 0 \text{ fixé .}
\end{aligned}$$

On construit un classifieur en se restreignant à la classe des classifieurs indexés par des rectangles $\{c_r, r = [a, b] \times [c, d], a < b, c < d\}$ et définis par

$$\begin{cases} c_r(x) = 1 & \text{si } x \in r \\ c_r(x) = 0 & \text{si } x \notin r. \end{cases}$$

1. Quelle est l'erreur empirique (associée à la perte 0/1) du rectangle vert dessiné sur la figure 1 que l'on note ici $\hat{r}$?

2. Soit un classifieur c_r. On considère une nouvelle observation (X_+, Y_+). Dans quelle zone du plan doit être X_+ pour que cette observation soit mal classée par le classifieur c_r ?

3. On définit quatre rectangles $r_l^\star, r_t^\star, r_r^\star, r_b^\star$ (l pour "left", t pour "top", r pour "right" et b pour "bottom"), les rectangles $r_l^\star, r_t^\star$ ont été représentés sur la figure 2. Chacun de ces rectangles vérifie
$$\mathbb{P}(X_+ \in r_k^\star) = \epsilon/4 \;\; (k \in \{l, t, r, b\}).$$
Montrer que l'erreur de généralisation du classifieur $c_{r_{\text{minus}}}$ associé au rectangle (représenté sur la figure 3) $r_{\text{minus}}^\star = r^\star \setminus \cup_{k\in\{l,t,r,b\}} r_k^\star$ vérifie
$$R(c_{r_{\text{minus}}^\star}) \leq \sum_{k\in\{l,t,r,b\}} \mathbb{P}(X_+ \in r_k^\star) = \epsilon.$$

4. On cherche maintenant à borner l'erreur de généralisation de $c_{\hat{r}}$.
   (a) Montrer que si $r_{\text{minus}}^\star \subset \hat{r}$ alors $R(c_{r_{\text{minus}}^\star}) \geq R(c_{\hat{r}})$.
   (b) En déduire que
   $$\mathbb{P}\left(R(c_{\hat{r}}) > \epsilon\right) \leq \mathbb{P}\left(r_{\text{minus}}^\star \not\subset \hat{r}\right).$$
   (c) Montrer que
   $$\mathbb{P}\left(r_{\text{minus}}^\star \not\subset \hat{r}\right) \leq \sum_{k\in\{l,t,r,b\}} \mathbb{P}(\hat{r} \cap r_k^\star = \emptyset) \leq 4\Big(1 - \frac{\epsilon}{4}\Big)^n.$$
   (d) Que doit valoir n pour que le risque de $c_{\hat{r}}$ dépasse ϵ avec une probabilité inférieure à $\delta > 0$.

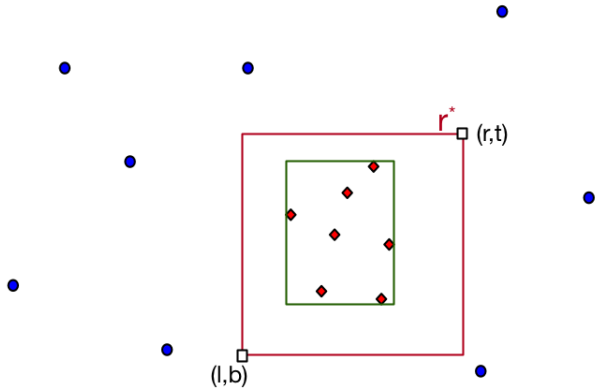

Figure 1
les données

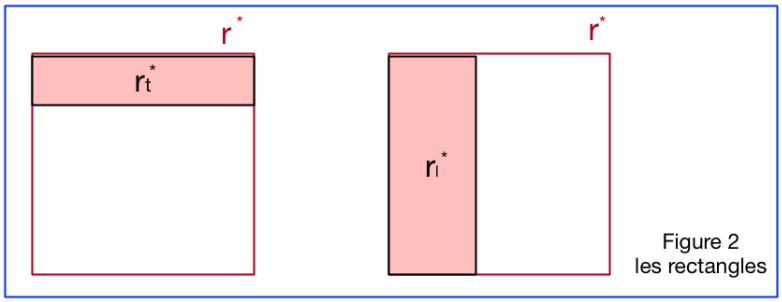

Figure 2
les rectangles

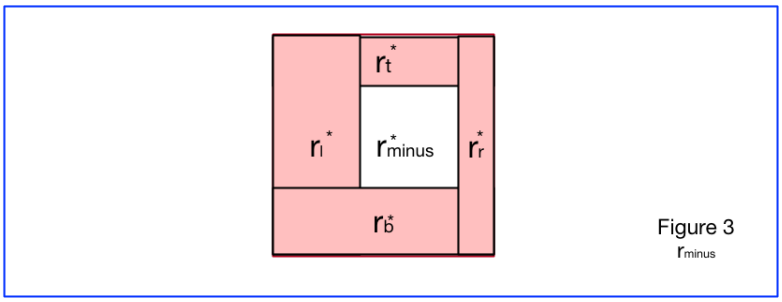

Figure 3
r_{minus}