Exercices classification binaire

Geneviève Robin

28 Septembre 2021

Exercice 1

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) $r^\star_{\text{minus}} = r^\star \setminus \cup_{k \in \{l,t,r,b\}} r_k^\star$ vérifie

$$R(c_{r^\star_{\text{minus}}}) \leq \sum_{k \in \{l,t,r,b\}} \mathbb{P}(X_+ \in r_k^\star) = \epsilon.$$

4. On cherche maintenant à borner l'erreur de généralisation de $c_{\hat{r}}$.
   (a) Montrer que si $r^\star_{\text{minus}} \subset \hat{r}$ alors $R(c_{r^\star_{\text{minus}}}) \geq R(c_{\hat{r}})$.
   (b) En déduire que

$$\mathbb{P}\left(R(c_{\hat{r}}) > \epsilon\right) \leq \mathbb{P}\left(r^\star_{\text{minus}} \not\subset \hat{r}\right).$$

(c) Montrer que

$$\mathbb{P}\left(r^{\star}_{\text{minus}} \not\subset \hat{r}\right) \leq \sum_{k \in \{l,t,r,b\}} \mathbb{P}(\hat{r} \cap r^{\star}_{k} = \emptyset) \leq 4\big(1 - \frac{\epsilon}{4}\big)^{n}.$$

(d) Que doit valoir n pour que le risque de $c_{\hat{r}}$ dépasse ϵ avec une probabilité inférieure à $\delta > 0$.

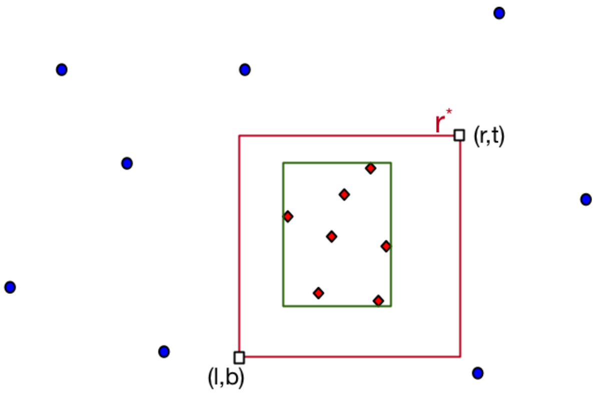


Figure 1
les données

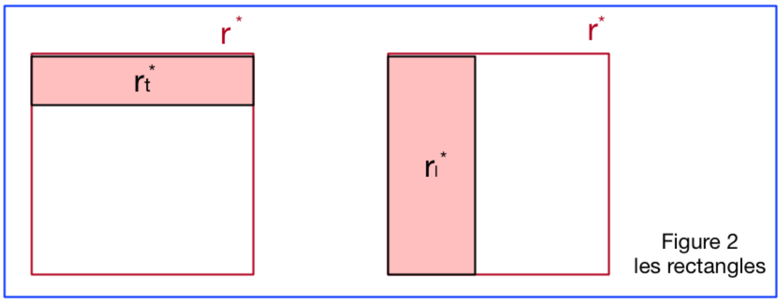


Figure 2
les rectangles

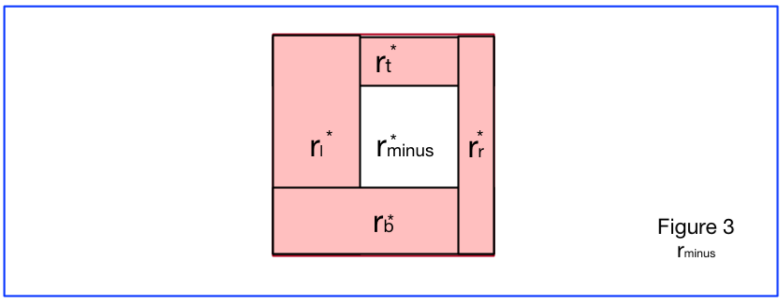


Figure 3
rminus

Exercice 2

On veut étudier l'algorithme des k plus proches voisins dans le cas où $k = 1$. On se place dans le cas où

- les covariables $X_i \in \mathcal{X} \subset [0,1]^d$
- les $Y_i \in \{0,1\}$.
- les couples (X_i, Y_i) sont i.i.d. de loi inconnue $\mathcal{L}$
- on note $\mathcal{D}_n = \big\{(X_i, Y_i), i = 1, \ldots, n\big\}$.

On considère de plus la perte 0/1 de sorte que le risque d'un classifieur est donné par

$$R(c) = \mathbb{E}_{\mathcal{L}}\big(\ell(Y_+, c(X_+))|\mathcal{D}_n\big) = \mathbb{E}_{\mathcal{L}}\big(\mathbb{1}_{Y_+ \neq c(X_+)}|\mathcal{D}_n\big)$$

où (X_+, Y_+) est indépendant des (X_i, Y_i) et de même loi.

1. On va étudier des classifieurs non-déterministes, c'est-à-dire qui dépendent des données. On note $\hat{c}$ un tel classifieur. Comme on va intéger par rapport aux lois de (X_+, Y_+) et $\mathcal{D}_n$ alternativement on va noter dans cette question

   - $\mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}$ quand on intègre par rapport à la loi de (X_+, Y_+)
   - $\mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}$ quand on intègre par rapport à la loi de $\mathcal{D}_n$

   avec ces notations

   $$R(\hat{c}) = \mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}(\ell(Y_+, \hat{c}(X_+)).$$

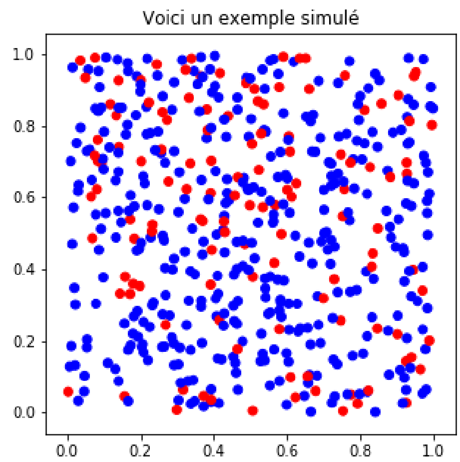
   Vérifier que

   $$\begin{aligned}\mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}\big(R(\hat{c})\big) &= \mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}\Big(\mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}\big(\ell(Y_+, \hat{c}(X_+))|\mathcal{D}_n\big)\Big) \\ &= \mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}\Big(\mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}\big(\ell(Y_+, \hat{c}(X_+))|(X_+, Y_+)\big)\Big)\end{aligned}$$

2. Dans toute la suite, chaque Y_i suit la loi de Bernoulli de paramètre $\pi^\star = 3/4$ pour toute valeur valeur de X_i, c'est-à-dire

   $$\mathbb{P}_{\mathcal{L}}\big(Y_i = 1|X_i = x\big) = \pi^\star = 3/4 \ \text{ pour tout } \quad x \in \mathcal{X}.$$

   Montrer que Y_i est indépendant de X_i. Proposez alors un classifieur déterministe (qui ne dépend pas de $\mathcal{D}_n$) c_0, c'est-à-dire une fonction de $\mathcal{X}$ dans $\{0,1\}$, qui vous paraît convenir au problème.



3. Montrer que le risque d'un classifeur déterministe c peut s'écrire successivement

$$R(c) = \mathbb{E}_{\mathcal{L}}\big(\mathbb{1}_{Y_+ \neq c(X_+)}\big) = \pi^\star + (1-2\pi^\star)\mathbb{E}_{\mathcal{L}}(c(X_+)).$$

   (a) Montrer que, pour tout classifieur déterministe c,

$$0 \leq \mathbb{E}_{\mathcal{L}}(c(X_+)) \leq 1.$$

   En déduire une borne inférieure de $R(c)$ et la valeur de $\mathbb{E}_{\mathcal{L}}(c(X_+))$ pour laquelle elle est atteinte.

   (b) En déduire que l'oracle $c^\star$ qui vérifie $c^\star = \text{argmin}_c R(c)$ vaut c_0.

   (c) Calculer le risque de l'oracle.

4. On considère le classifieur du plus proche voisin (1 plus proche voisin) noté $\hat{c}_1$. Pour chaque $x \in \mathcal{X}$, on introduit les v.a. $Z_i(x)$ $(i = 1, \dots, n)$ définies par

$$Z_i(x) = \begin{cases} 1 & \text{si } i \text{ est le plus proche voisin de } x \\ 0 & \text{sinon.} \end{cases}$$

   (a) Vérifier que $\sum_{i=1}^n Z_i(x) = 1$.

   (b) Montrer que $\hat{c}_1(x) = \sum_{i=1}^n Z_i(x) Y_i$.

   (c) Déduire de la question 1 que

$$\mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\big(\hat{c}_1(x)\big) = \sum_{i=1}^n \mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\big(Z_i(x) Y_i\big) = \pi^\star.$$

   (d) En déduire que

$$\mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\big(\ell(Y_+, \hat{c}_1(X_+)) | (X_+, Y_+)\big) = 3/8.$$

5. Déduire des questions 1 et 4 que $\mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\big(R(\hat{c}_1)\big) = 3/8$.

6. En déduire que le risque intégré de $\hat{c}_1$ ne tend pas vers celui de l'oracle $c^\star$.

7. Expliquer pourquoi.