

---

TRAVAUX DIRIGÉS N° 3 : Séparateurs linéaires

---

Stephan CLÉMENÇON <stephan.clemencon@telecom-paris.fr>  
Ekhine IRUOZKI <irurozki@telecom-paris.fr>

**EXERCICE 1.** On se place dans le cadre de la classification binaire : soient un descripteur aléatoire  $X$  à valeurs dans  $\mathbb{R}$  muni de sa tribu des Boréliens, et un label aléatoire  $Y$  valant 0 ou 1.

Soit  $\mathbb{G} := \{g : \mathbb{R} \rightarrow \{0, 1\}\}$  l'ensemble des classifieurs adaptés à ce contexte. L'erreur de classification est définie comme l'application  $L : g \in \mathbb{G} \mapsto \mathbb{P}(Y \neq g(X)) \in [0, 1]$  et on note  $L^* := \inf_{g \in \mathbb{G}} L(g)$ .

Dans cet exercice, on s'intéresse à la famille  $\mathcal{G}$  des classifieurs linéaires sur  $\mathbb{R}$  de la forme :

$$g_{(x_0, y_0)} : x \in \mathbb{R} \mapsto \begin{cases} y_0 & \text{si } x \leq x_0, \\ 1 - y_0 & \text{sinon,} \end{cases}$$

avec  $(x_0, y_0) \in \mathbb{R} \times \{0, 1\}$ . L'erreur de classification d'un tel  $g_{(x_0, y_0)}$  est notée plus simplement  $L(x_0, y_0)$  et on pose  $L_0 := \inf_{(x_0, y_0)} L(x_0, y_0)$ .

1) Exprimer l'erreur de classification d'un élément quelconque de  $\mathcal{G}$  en fonction des lois conditionnelles de  $X$  sachant  $Y$ . On utilisera les notations  $F_y(x) := \mathbb{P}\{X \leq x \mid Y = y\}$  pour  $(x, y) \in \mathbb{R} \times \{0, 1\}$  et  $p := \mathbb{P}(Y = 1)$ .

2) En considérant les points  $(x_0, y_0) = (-\infty, 0)$  et  $(x_0, y_0) = (-\infty, 1)$ , montrer que  $L_0 \leq \frac{1}{2}$ .

3) Montrer que  $L_0 = \frac{1}{2} - \sup_x \left| p F_1(x) - (1 - p) F_0(x) - p + \frac{1}{2} \right|$ . Simplifier l'expression quand  $p = \frac{1}{2}$ .

**Indication.** Pour tout  $(a, b) \in \mathbb{R}^2$  on peut écrire  $\min(a, b) = \frac{a + b - |a - b|}{2}$ .

4) Montrer que  $L_0 = \frac{1}{2}$  si et seulement si  $L^* = \frac{1}{2}$ .

5) Montrer l'inégalité de Chebychev-Cantelli : pour toute variable aléatoire réelle  $Z$  et tout  $t \geq 0$ ,

$$\mathbb{P}(Z - \mathbb{E}(Z) \geq t) \leq \frac{\mathbb{V}(Z)}{\mathbb{V}(Z) + t^2}.$$

6) On note respectivement  $m_y$  et  $\sigma_y^2$  l'espérance et la variance de la loi conditionnelle de  $X$  sachant  $Y = y$ , avec  $y \in \{0, 1\}$ . Montrer que :

$$L_0 \leq \left( 1 + \frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2} \right)^{-1}.$$

**Indication.** Utiliser l'inégalité démontrée à la question précédente.

7) Discuter de la performance du minimiseur empirique pris dans la classe  $\mathcal{G}$  et des limites des classifieurs linéaires.

---

**Solution:**

1) Soit  $(x, y) \in \mathbb{R} \times \{0, 1\}$ , alors ,

$$\begin{aligned}
 L(x, y) &= \mathbb{P}(Y \neq g_{(x, y)}(X)) = \mathbb{P}(Y = y, X > x) + \mathbb{P}(Y = 1 - y, X \leq x) \\
 &= \mathbb{P}(X > x \mid Y = y)\mathbb{P}(Y = y) + \mathbb{P}(X \leq x \mid Y = 1 - y)\mathbb{P}(Y = 1 - y) \\
 &= y(p(1 - F_1(x)) + (1 - p)F_0(x)) + (1 - y)((1 - p)(1 - F_0(x)) + pF_1(x)) \\
 &= (1 - y)(F_1(x)p + 1 - p - F_0(x) + F_0(x)p) + y(F_0(x) - F_0(x)p + p - F_1(x)p) \\
 &= (1 - y)(1 - \phi(x)) + y\phi(x).
 \end{aligned}$$

2) Soit  $y \in \{0, 1\}$ . D'après la question précédente, puisque  $F_0$  et  $F_1$  sont des fonctions de répartition, on a

$$\lim_{x \rightarrow -\infty} L(x, y) = \begin{cases} p, & \text{si } y = 1, \\ 1 - p, & \text{si } y = 0. \end{cases}$$

On en déduit que  $L_0 \leq p \wedge (1 - p) \leq \frac{1}{2}$ .

3) Posons  $\phi : x \in \mathbb{R} \mapsto (1 - p)F_0(x) - pF_1(x) \in [0, 1]$  et prenons  $(x, y) \in \mathbb{R} \times \{0, 1\}$ . D'après la première question, on a

$$L(x, y) = \begin{cases} \phi(x), & \text{si } y = 1, \\ (1 - \phi(x)), & \text{si } y = 0. \end{cases}$$

Ainsi,

$$\begin{aligned}
 L_0 &= \inf_{x \in \mathbb{R}} \min((1 - \phi(x)), \phi(x)) = \inf_{x \in \mathbb{R}} \frac{1 - \phi(x) + \phi(x) - |1 - \phi(x) - \phi(x)|}{2} \\
 &= \inf_{x \in \mathbb{R}} \left( \frac{1}{2} - \left| \frac{1}{2} - \phi(x) \right| \right) = \frac{1}{2} - \sup_{x \in \mathbb{R}} \left( \left| \frac{1}{2} - \phi(x) \right| \right) \\
 &= \frac{1}{2} - \sup_{x \in \mathbb{R}} \left( \left| \frac{1}{2} - (F_0(x) - F_0(x)p + p - F_1(x)) \right| \right) \\
 &= \frac{1}{2} - \sup_{x \in \mathbb{R}} \left| pF_1(x) - (1 - p)F_0(x) - p + \frac{1}{2} \right|
 \end{aligned}$$

Lorsque  $p = \frac{1}{2}$ , cela donne

$$L_0 = \frac{1}{2} \left( 1 - \sup_{x \in \mathbb{R}} |F_0(x) - F_1(x)| \right).$$

4) — Supposons que  $L^* = \frac{1}{2}$ . Par définition de  $L^*$ , on a  $L_0 \geq L^* = \frac{1}{2}$ .

Puisque  $L_0$  est aussi majorée par  $\frac{1}{2}$ , d'après la question 2, on a donc bien  $L_0 = \frac{1}{2}$ .

— Réciproquement, supposons que  $L_0 = \frac{1}{2}$ .

Puisque  $L_0 \leq p \wedge (1 - p) \leq \frac{1}{2}$  (cf. question 2), on a alors  $p = \frac{1}{2}$ .

En utilisant le dernier résultat de la question 3, on obtient ensuite  $\sup_{x \in \mathbb{R}} |F_0(x) - F_1(x)| = 0$ , ce qui signifie que  $F_0 = F_1$  (i.e.  $X \mid Y = 0$  a la même loi que  $X \mid Y = 1$ ).

On en déduit que pour tout  $g \in \mathbb{G}$ ,

$$\begin{aligned} L(g) &= \mathbb{P}(Y \neq g(X)) = \mathbb{P}(g(X) = 0 \mid Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1 \mid Y = 0)\mathbb{P}(Y = 0) \\ &= \frac{1}{2} (\mathbb{P}(g(X) = 0 \mid Y = 1) + \mathbb{P}(g(X) = 1 \mid Y = 1)) = \frac{1}{2}, \text{ (car } F_0 = F_1), \end{aligned}$$

d'où  $L^* = \frac{1}{2}$ .

5) Soit  $Z$  une variable aléatoire de carré intégrable et  $t \geq 0$ . Posons  $Z^* := Z - \mathbb{E}(Z)$ . On remarque que

$$0 \leq t = \mathbb{E}(t - Z^*) \leq \mathbb{E}((t - Z^*)\mathbb{1}_{\{Z^* < t\}}),$$

d'où

$$\begin{aligned} t^2 &\leq \mathbb{E}((t - Z^*)^2)\mathbb{E}(\mathbb{1}_{\{Z^* < t\}}) \quad (\text{Cauchy-Schwarz}) \\ &= \mathbb{E}(t^2 + (Z^*)^2 - 2tZ^*)\mathbb{P}(Z^* < t) \\ &= (\mathbb{V}(Z^*) + t^2)\mathbb{P}(Z^* < t) \\ &= (\mathbb{V}(Z) + t^2)\mathbb{P}(Z - \mathbb{E}(Z) < t). \end{aligned}$$

Par conséquent,

$$\begin{aligned} \mathbb{P}(Z - \mathbb{E}(Z) < t) &\geq \frac{t^2}{\mathbb{V}(Z) + t^2} \\ \iff 1 - \mathbb{P}(Z - \mathbb{E}(Z) \geq t) &\geq \frac{t^2}{\mathbb{V}(Z) + t^2} \\ \iff \mathbb{P}(Z - \mathbb{E}(Z) \geq t) &\leq 1 - \frac{t^2}{\mathbb{V}(Z) + t^2} = \frac{\mathbb{V}(Z)}{\mathbb{V}(Z) + t^2}. \end{aligned}$$

6) Commençons par remarquer que si  $m_0 = m_1$ , alors l'inégalité à démontrer est  $L_0 \leq 1$ , ce qui est toujours vrai.

Supposons maintenant que  $m_0 < m_1$ . On peut alors choisir deux réels  $\Delta_0, \Delta_1 > 0$  tels que  $m_1 - m_0 = \Delta_0 + \Delta_1$  et considérer le classifieur  $g_{(\Delta_0 + m_0, 0)}$  (cf. Figure 1).

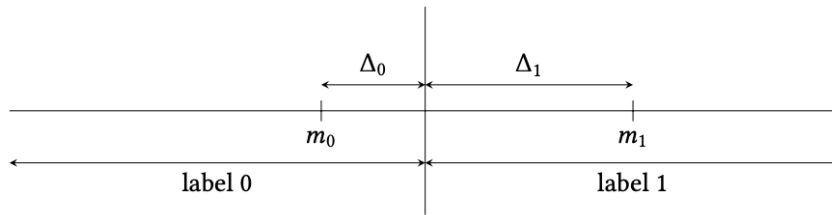


FIGURE 1 – Sketch of the classifier  $g(\Delta_0 + m_0, 0)$ .

D'après la question 1, en remarquant que  $\Delta_0 + m_0 = m_1 - \Delta_1$ , on obtient

$$\begin{aligned} L(\Delta_0 + m_0, 0) &= pF_1(\Delta_0 + m_0) + (1 - p)(1 - F_0(\Delta_0 + m_0)) \\ &= p\mathbb{P}(X \leq \Delta_0 + m_0 \mid Y = 1) + (1 - p)\mathbb{P}(X > \Delta_0 + m_0 \mid Y = 0) \\ &= p\mathbb{P}(X \leq m_1 - \Delta_1 \mid Y = 1) + (1 - p)\mathbb{P}(X > \Delta_0 + m_0 \mid Y = 0) \\ &= p\mathbb{P}(-X - (-m_1) \geq \Delta_1 \mid Y = 1) + (1 - p)\mathbb{P}(X - m_0 > \Delta_0 \mid Y = 0) \end{aligned}$$

$$\begin{aligned}
&= p \frac{\sigma_1^2}{\sigma_1^2 + \Delta_1^2} + (1-p) \frac{\sigma_0^2}{\sigma_0^2 + \Delta_0^2} \\
&= \frac{p}{1 + \frac{\Delta_1^2}{\sigma_1^2}} + \frac{1-p}{1 + \frac{\Delta_0^2}{\sigma_0^2}}.
\end{aligned}$$

En prenant  $\Delta_0 = \frac{m_1 - m_0}{\sigma_1 + \sigma_0} \sigma_0$  et  $\Delta_1 = \frac{\sigma_1}{\sigma_0} \Delta_0$ , on trouve

$$L(\Delta_0 + m_0, 0) \leq \left( 1 + \frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2} \right)^{-1}.$$

Puisque par définition  $L_0 \leq L(\Delta_0 + m_0, 0)$ , on retombe bien sur l'inégalité recherchée.

**Remarque.** Le cas  $m_1 < m_0$  peut être traité de la même manière en inversant tous les indices 0 et 1 des objets apparaissant dans la Figure 1.

**Interprétation.** Ce résultat nous dit que plus l'écart entre  $m_0$  et  $m_1$  est grand et plus  $\sigma_0^2$  et  $\sigma_1^2$  sont faibles, plus on sera à même de garantir un faible risque de classification sur  $G$ . Exprimé de manière plus grossière, mieux les classes sont séparées, plus on peut garantir une bonne classification en utilisant un simple séparateur linéaire (ce qui semble tout à fait logique). Une illustration est proposée à la Figure 2.

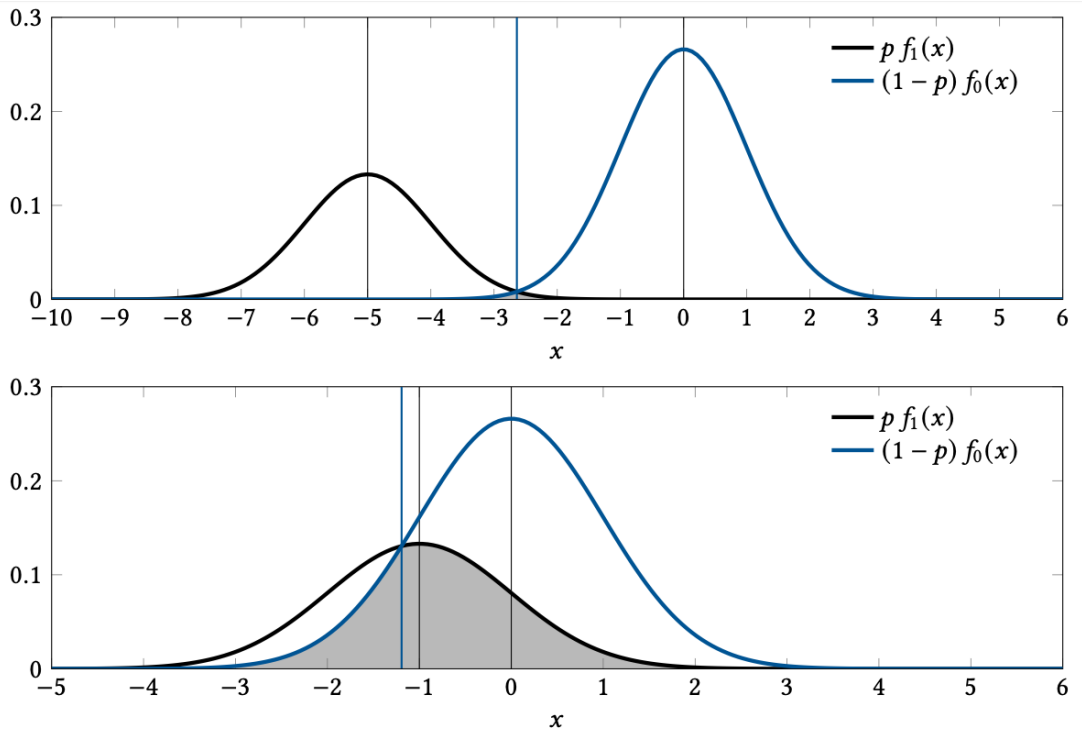


FIGURE 2 – Illustration du résultat de la question 6 :  $p = \frac{1}{3}$ ,  $X | Y = 0$  et  $X | Y = 1$  suivent toutes deux des gaussiennes réduites, d'espérances  $m_0 = 0$  et  $m_1 = -5$  (haut) ou  $m_1 = -1$  (bas), et de densités respectives  $f_0$  et  $f_1$ . Le classifieur optimal donne le label 0 à droite du séparateur (trait vertical bleu) et son risque  $L_0$  est indiqué par la zone grisée.

7) On s'intéresse maintenant au risque empirique

$$L_n : g \in \mathbb{G} \mapsto \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i \neq g(X_i)\},$$

minimisé sur  $\mathcal{G}$  par  $\hat{g}_n = g(\hat{x}_n, \hat{y}_n)$ . Par application directe des résultats vus en cours, nous avons d'une part

$$L(\hat{g}_n) \leq 2 \sup_{(x,y) \in \mathbb{R} \times \{0,1\}} |L(x,y) - L_n(x,y)| + L_0 \quad \text{p.s.}$$

puis d'autre part que pour tout  $\epsilon > 0$  et tout  $y \in \{0,1\}$ ,

$$\mathbb{P} \left( \sup_{x \in \mathbb{R}} |L(x,y) - L_n(x,y)| \geq \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

On déduit de ces deux inégalités que pour tout  $\epsilon > 0$ ,

$$\begin{aligned} \mathbb{P}(L(\hat{g}_n) - L_0 \geq \epsilon) &\leq \mathbb{P} \left( \sup_{(x,y) \in \mathbb{R} \times \{0,1\}} |L(x,y) - L_n(x,y)| \geq \frac{\epsilon}{2} \right) \\ &\leq \mathbb{P} \left( \sup_{x \in \mathbb{R}} |L(x,0) - L_n(x,0)| \geq \frac{\epsilon}{2} \right) + \mathbb{P} \left( \sup_{x \in \mathbb{R}} |L(x,1) - L_n(x,1)| \geq \frac{\epsilon}{2} \right) \\ &\leq 4 \exp \left( -n \frac{\epsilon^2}{2} \right). \end{aligned}$$

et ce quelle que soit la loi du vecteur  $(X, Y)$ . Évidemment, contrôler ce risque relatif n'a de sens que si le risque théorique  $L_0$  est lui-même petit. Les limites des séparateurs linéaires (ici quand  $X$  est à valeurs dans  $\mathbb{R}$ ) sont aisément illustrées par l'exemple donné en Figure 3. Dans ce type de configuration, aucun séparateur linéaire ne sera à même de faire mieux qu'une labellisation purement aléatoire.

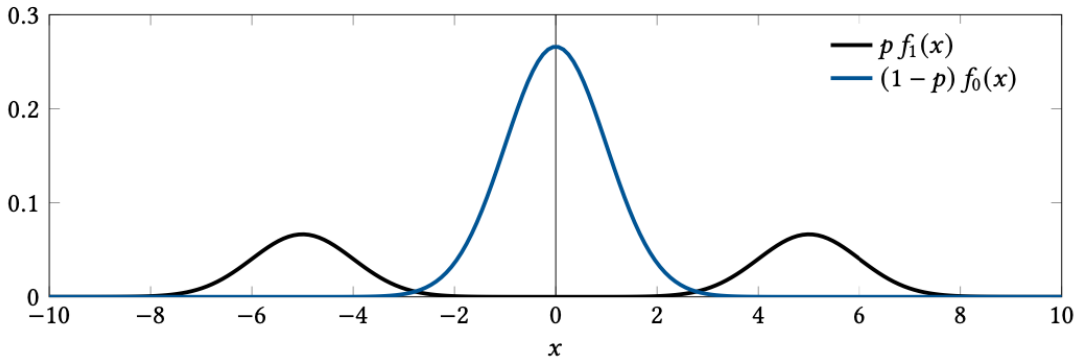


FIGURE 3 – Exemple où un séparateur linéaire n'est pas approprié :  $p = \frac{1}{3}$ ,  $X|Y = 0$  suit une loi Normale centrée réduite de densité  $f_0$ ,  $X|Y = 1$  suit un mélange équilibré de deux gaussiennes réduites, la première centrée en -5 et la seconde en 5, de densité  $f_1$ , de telle manière que  $m_0 = m_1 = 0$ .