Travaux dirigés n° 1 : Classifieur de Bayes

Stephan Clémençon <stephan.clemencon@telecom-paris.fr>
Ekhine Irurozki <irurozki@telecom-paris.fr>

On se place dans le cadre de la classification binaire. Dans tout le TD, on considère un descripteur aléatoire X à valeurs dans un espace mesurable $\mathcal{X} \subset \mathbb{R}^d$ ($d \in \mathbb{N}^*$) et un label aléatoire Y valant 0 ou 1. La distribution jointe du vecteur (X, Y) est notée P, et la fonction de régression (probabilité a posteriori)

$$\eta : x \in \mathcal{X} \mapsto \mathbb{P}(Y = 1 \mid X = x) \in [0, 1].$$

Exercice 1. On considère $\mathcal{X} = [0, 1]$ et P telle que :

- la distribution conditionnelle de X sachant $Y = 0$ est $P_0 = \mathcal{U}([0, \theta])$ où $\theta \in]0, 1[$,
- la distribution conditionnelle de X sachant $Y = 1$ est $P_1 = \mathcal{U}([0, 1])$,
- $p = \mathbb{P}(Y = 1) \in]0, 1[$.

Pour $x \in \mathcal{X}$, donner $\eta(x)$ en fonction de p et θ.

Exercice 2. On considère $\mathcal{X} = \mathbb{R}_+$ et P telle que :

- la distribution de X est P_X sur $\mathcal{X}$,
- la fonction de régression vaut $\eta(x) = \dfrac{x}{x + \theta}$ pour tout $x \in \mathcal{X}$, où $\theta > 0$ est fixé.

On note h^* le classifieur de Bayes. Expliciter le classifieur de Bayes dans ce modèle. Montrer ensuite que son risque "0-1" (sa probabilité d'erreur) vaut

$$L(h^*) = \int_{\mathcal{X}} \min\Big(\eta(x), 1 - \eta(x)\Big)\, dP_X(x).$$

Calculer le risque de Bayes lorsque $P_X = \mathcal{U}([0, \alpha\theta])$ où $\alpha > 1$.

Exercice 3. Soient des poids $\omega(0), \omega(1) \geq 0$ tels que $\omega(0) + \omega(1) = 1$. On considère le risque de classification pondéré :

$$L_\omega(g) = \mathbb{E}\left(2\omega(Y) \cdot \mathbb{1}_{\{Y \neq g(X)\}}\right), \qquad g : \mathcal{X} \to \{0, 1\}.$$

Donner le classifieur de Bayes et le risque de Bayes pour ce critère. Quel est l'intérêt de considérer un tel critère ?

Exercice 4. On considère $X = (T, U, V)$ où T, U, V sont des variables aléatoires réelles i.i.d. de loi exponentielle standard. On pose $Y = \mathbb{1}_{\{T+U+V<\theta\}}$ où $\theta \in \mathbb{R}_+$ est fixé.

1) Calculer le classifieur de Bayes $g^*(T, U)$, lorsque V n'est pas observée. Calculer le risque de Bayes "0-1" associé à ce classifieur.

2) On suppose à présent que seule T est observée. Reprendre les calculs précédents et comparer les risques bayésiens obtenus lorsque $\theta = 9$.
3) Proposer un classifieur lorsque X n'a aucune composante qui soit observée. Calculer son erreur de classification.

Exercise 1, solution On remarque tout d'abord que les distributions P_0 et P_1 étant Uniformes, elles admettent toutes deux des densités, notées respectivement

$$f_0 : x \in \mathbb{R} \mapsto \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x) \quad \text{et} \quad f_1 : x \in \mathbb{R} \mapsto \mathbb{1}_{[0,1]}(x).$$

Par conséquent, d'après la formule des probabilités totales, la loi non conditionnelle de X admet elle aussi une densité :

$$f : x \in \mathbb{R} \mapsto f_0(x)(1-p) + f_1(x)p = \frac{1-p}{\theta}\mathbb{1}_{[0,\theta]}(x) + p\mathbb{1}_{[0,1]}(x).$$

Soit maintenant $x \in [0,1]$. D'après la formule de Bayes, on a

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{f_1(x)p}{f(x)} = \frac{p\mathbb{1}_{[0,1]}(x)}{p\mathbb{1}_{[0,1]}(x) + \frac{1-p}{\theta}\mathbb{1}_{[0,\theta]}(x)} = \begin{cases} 1 & \text{si } x \in (\theta, 1], \\ \dfrac{\theta p}{1-p+\theta p} & \text{si } x \in [0, \theta], \end{cases}$$

Lorsque $\theta = \frac{1}{2}$, on obtient

$$\eta(x) = \begin{cases} 1 & \text{si } x \in \left(\frac{1}{2}, 1\right], \\ \dfrac{p}{2-p} & \text{si } x \in \left[0, \frac{1}{2}\right], \end{cases}$$

comme illustré à la Figure 1.

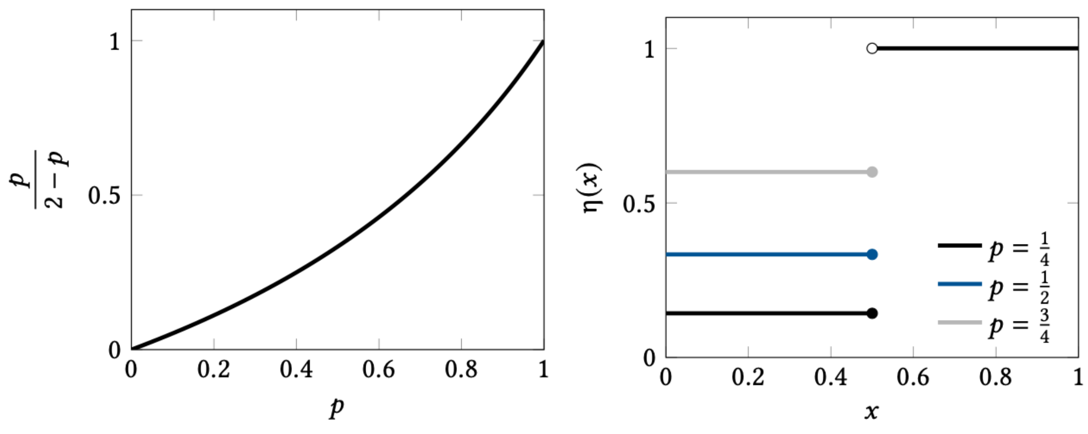

FIGURE 1 – Illustration pour $\theta = \frac{1}{2}$.

Exercise 2, solution

1) Le classifieur de Bayes pour des labels valant 0 ou 1 est la fonction

$$h^* : x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) > \frac{1}{2}\}}.$$

On peut remarquer que la fonction η étant à valeurs dans $[0,1]$, h^* peut être réécrit comme

$$h^* : x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) > 1-\eta(x)\}}.$$

Son risque 0-1 vaut donc :

$$\begin{aligned}
\mathrm{L}(h^*) &= \mathbb{E}[\mathbb{1}_{\{\mathrm{Y}\neq h^*(\mathrm{X})\}}] \\
&= \mathbb{E}[\mathbb{1}_{\{\mathrm{Y}=0\}}\mathbb{1}_{\{h^*(\mathrm{X})=1\}} + \mathbb{1}_{\{\mathrm{Y}=1\}}\mathbb{1}_{\{h^*(\mathrm{X})=0\}}] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\mathrm{Y}=0\}}\mathbb{1}_{\{h^*(\mathrm{X})=1\}} + \mathbb{1}_{\{\mathrm{Y}=1\}}\mathbb{1}_{\{h^*(\mathrm{X})=0\}} \mid \mathrm{X}\right]\right] \quad \text{(espérance totale)} \\
&= \mathbb{E}\left[(1-\eta(\mathrm{X}))\mathbb{1}_{\{h^*(\mathrm{X})=1\}} + \eta(\mathrm{X})\mathbb{1}_{\{h^*(\mathrm{X})=0\}}\right] \\
&= \mathbb{E}\left[(1-\eta(\mathrm{X}))\mathbb{1}_{\{\eta(\mathrm{X})>1-\eta(\mathrm{X})\}} + \eta(\mathrm{X})\mathbb{1}_{\{\eta(\mathrm{X})\leq 1-\eta(\mathrm{X})\}}\right] \\
&= \mathbb{E}[\min(\eta(\mathrm{X}), 1-\eta(\mathrm{X}))] \\
&= \int_{\mathcal{X}} \min(\eta(x), 1-\eta(x))\,\mathrm{dP}_{\mathrm{X}}(x).
\end{aligned}$$

2) (a) Soit $x \in \mathbb{R}^+$, alors (cf. Figure 2)

$$\frac{1}{\eta(x)} > 2 \quad \Leftrightarrow \quad \frac{x}{x+\theta} > \frac{1}{2} \quad \Leftrightarrow \quad x > \frac{x+\theta}{2} \quad \Leftrightarrow \quad x > \theta.$$

Ainsi, $h^*(x) = \mathbb{1}_{\{x>\theta\}}$ et d'après la question précédente

$$\begin{aligned}
\mathrm{L}(h^*) &= \int_0^{+\infty} \min(\eta(x), 1-\eta(x))\,\mathrm{P_X}(\mathrm{d}x) \\
&= \int_0^{+\infty} \min\left(\frac{x}{x+\theta}, \frac{\theta}{x+\theta}\right)\mathrm{P_X}(\mathrm{d}x) \\
&= \int_0^{\theta} \frac{x}{x+\theta}\mathrm{P_X}(\mathrm{d}x) + \int_\theta^{+\infty} \frac{\theta}{x+\theta}\mathrm{P_X}(\mathrm{d}x) \\
&= \int_0^{\theta} \left(1 - \frac{\theta}{x+\theta}\right)\mathrm{P_X}(\mathrm{d}x) + \int_\theta^{+\infty} \frac{\theta}{x+\theta}\mathrm{P_X}(\mathrm{d}x) \\
&= \mathbb{P}(\mathrm{X}\leq\theta) + \int_\theta^{+\infty} \frac{\theta}{x+\theta}\mathrm{P_X}(\mathrm{d}x) - \int_0^{\theta} \frac{\theta}{x+\theta}\mathrm{P_X}(\mathrm{d}x).
\end{aligned}$$

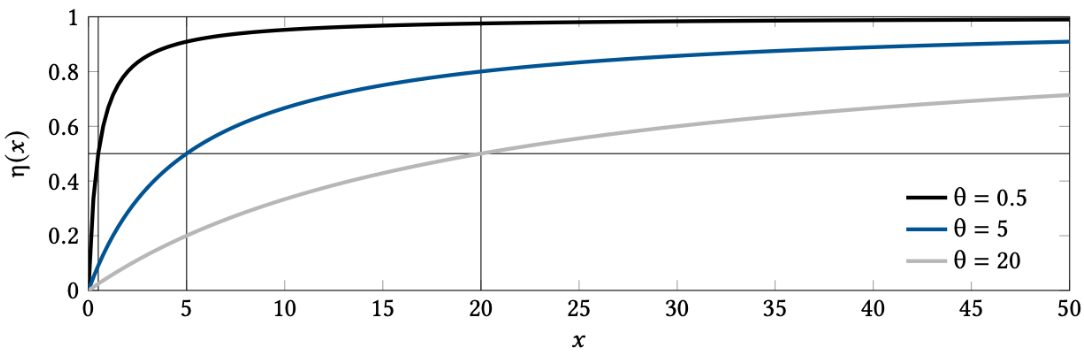

Figure 2 – Tracé de la fonction de régression pour différentes valeurs de θ.

(b) Lorsque $\mathrm{P_X}$ est une loi uniforme sur $[0, \alpha\theta]$ avec $\alpha > 1$, on obtient

$$\begin{aligned}
\mathrm{L}(h^*) &= \frac{\theta}{\alpha\theta} + \int_\theta^{+\infty} \frac{\theta}{x+\theta}\frac{1}{\alpha\theta}\mathbb{1}_{[0,\alpha\theta]}(x)\,\mathrm{d}x - \int_0^{\theta} \frac{\theta}{x+\theta}\frac{1}{\alpha\theta}\mathbb{1}_{[0,\alpha\theta]}(x)\,\mathrm{d}x \\
&= \frac{1}{\alpha}\left(1 + \int_\theta^{\alpha\theta} \frac{1}{x+\theta}\,\mathrm{d}x - \int_0^{\theta} \frac{1}{x+\theta}\,\mathrm{d}x\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\alpha}\left(1 + \int_{2\theta}^{(1+\alpha)\theta} \frac{1}{x}\,\mathrm{d}x - \int_{\theta}^{2\theta} \frac{1}{x}\,\mathrm{d}x\right) \\
&= \frac{1}{\alpha}(1 + \ln((\alpha+1)\theta) - \ln(2\theta) - \ln(2\theta) + \ln(\theta)) \\
&= \frac{1}{\alpha}\ln\left(e\frac{\alpha+1}{4}\right).
\end{aligned}
$$

Cette fonction de α est illustrée à la Figure 3.

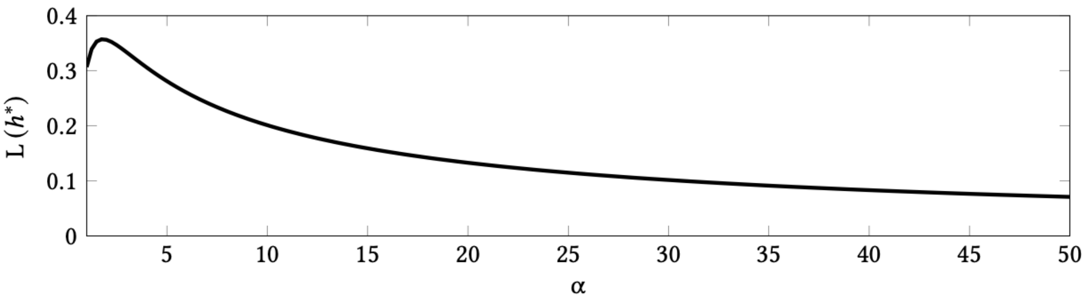

FIGURE 3 – Tracé de la fonction de régression pour différentes valeurs de θ.

Interprétation Posons $p := \mathbb{P}(Y = 1)$. On remarque que, sous les hypothèses de la question 2.b, on peut réécrire le risque de Bayes comme suit :

$$L(h^*) = \mathbb{P}(X \leq \theta | Y = 1)p + \mathbb{P}(X > \theta | Y = 0)(1 - p).$$

On peut expliciter p :

$$p := \mathbb{P}(Y = 1) = \int_X \eta(x) P_X(\mathrm{d}x) = \frac{1}{\alpha\theta}\int_0^{\alpha\theta} \frac{x}{x+\theta}\mathrm{d}x = 1 - \frac{1}{\alpha}\ln(\alpha+1).$$

Par ailleurs, en notant f la densité de X, d'après la formule de Bayes, X sachant Y = 1 possède une densité f_1 valant pour tout $x \in \mathbb{R}$:

$$f_1(x) = \frac{\eta(x) f(x)}{p} = \frac{x\mathbb{1}_{[0,\alpha\theta]}(x)}{(x+\theta)\theta(\alpha - \ln(\alpha+1))}.$$

et X sachant Y = 0 possède une densité f_0 valant pour tout $x \in \mathbb{R}$.

$$f_0(x) = \frac{(1-\eta(x)) f(x)}{1-p} = \frac{x\mathbb{1}_{[0,\alpha\theta]}(x)}{(x+\theta)(\ln(\alpha+1))}.$$

Cela nous permet de représenter graphiquement le risque de Bayes en fonction des lois du descripteur X dans chacune des classes déterminées par le label Y, comme à la Figure 4.
On retrouve bien dans l'aire grisée de la Figure 4 l'expression intégrale de la question 1 :

$$L(h^*) = \int_0^{+\infty} \min(\eta(x), 1-\eta(x)) f(x)\,\mathrm{d}x = \int_0^{+\infty} \min(p f_1(x), (1-p) f_0(x))\,\mathrm{d}x.$$

Remarque. Pour se convaincre que le classifieur de Bayes pour des labels valant 0 ou 1 est bien celui donné, on peut procéder comme suit. Dans le cours, il a été vu que pour des labels valant -1 ou 1, le classifieur de Bayes n'était autre que la fonction

$$g : x \in \mathcal{X} \mapsto 2\mathbb{1}_{\{\eta(x) > \frac{1}{2}\}} - 1,$$

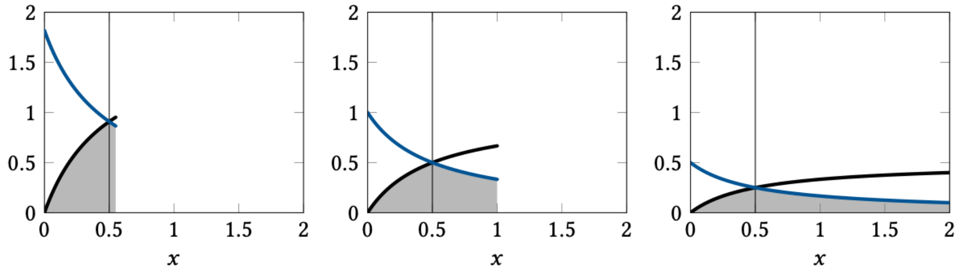

FIGURE 4 – Représentation graphique du risque de Bayes (aire grisée) selon pf_1 (courbe en noir) et $(1-p)f_0$ (courbe en bleu) pour $\theta = 0.5$ (trait vertical) et $\alpha = 1.1$ (gauche), $\alpha = 2$ (milieu), $\alpha = 4$ (droite), quand $P_X = \mathcal{U}([0, \alpha\theta])$.

qui minimise le risque 0-1 parmi tous les classifieurs $g : X \longrightarrow \{-1, 1\}$ possibles. On remarque que la variable aléatoire $Z := 2Y - 1$ prend ses valeurs dans {-1, 1}, et $\mathbb{P}(Z = 1) = \mathbb{P}(Y = 1)$. Alors

$$\mathbb{E}\{\mathbb{1}_{\{Y \neq h(X)\}}\} = \mathbb{E}\{\mathbb{1}_{\{Z \neq 2h(X)-1\}}\}$$

est bien minimal pour h^* tel que $2h^* - 1 = g^*$.

Exercise 3, solution Le classifieur de Bayes pour le risque de classification pondéré est par définition l'application $g^* : \mathcal{X} \longrightarrow \{0, 1\}$ telle que pour tout classifieur $g : \mathcal{X} \longrightarrow \{0, 1\}$,

$$R_\omega(g) := L_\omega(g) - L_\omega(g^*) \geq 0.$$

Soit $g : X \longrightarrow \{0, 1\}$ un classifieur quelconque. Alors

$$\begin{aligned}
L_\omega(g) &= \mathbb{E}\left[2\omega_Y \mathbb{1}_{\{Y \neq g(X)\}}\right] \\
&= 2\mathbb{E}\left[\omega_1 \mathbb{1}_{\{Y=1\}} \mathbb{1}_{\{g(X)=0\}}\right] + \omega_0 \mathbb{E}\left[\mathbb{1}_{\{Y=0\}} \mathbb{1}_{\{g(X)=1\}}\right] \\
&= 2\mathbb{E}\left[\omega_1 \eta(X) \mathbb{1}_{\{g(X)=0\}}\right] + \omega_0 \mathbb{E}\left[(1-\eta(X)) \mathbb{1}_{\{g(X)=1\}}\right] \quad \text{(espérance totale)} \\
&= 2\mathbb{E}\left[\omega_1 \eta(X)\right] + (\omega_0(1-\eta(X)) - \omega_1 \eta(X))\, \mathbb{1}_{\{g(X)=1\}} \\
&\quad (\mathit{since} \mathbb{1}_{\{g(X)=0\}} = 1 - \mathbb{1}_{\{g(X)=1\}} \text{ p.s.}) \\
&= 2\mathbb{E}\left[\omega_1 \eta(X)\right] + (\omega_0 - \eta(X)) \mathbb{1}_{\{g(X)=1\}}, \quad (\mathit{since} \omega_1 + \omega_0 = 1)
\end{aligned}$$

d'où

$$R_\omega(g) = 2\mathbb{E}\left[\omega_0 - \eta(X)\right] \left(\mathbb{1}_{\{g(X)=1\}} - \mathbb{1}_{\{g^*(X)=1\}}\right).$$

Or pour tout $x \in \mathcal{X}$ on a $(\omega_0 - \eta(x))(\mathbb{1}_{\{g(x)=1\}} - \mathbb{1}_{\{g^*(x)=1\}}) \geq 0$ ssi

$$\begin{aligned}
\omega_0 - \eta(x) &\geq 0, \\
\mathbb{1}_{\{g(x)=1\}} - \mathbb{1}_{\{g^*(x)=1\}} &\geq 0,
\end{aligned}$$

ou

$$\begin{aligned}
\omega_0 - \eta(x) &\leq 0, \\
\mathbb{1}_{\{g(x)=1\}} - \mathbb{1}_{\{g^*(x)=1\}} &\leq 0.
\end{aligned}$$

La fonction $g^* : x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) > \omega_0\}}$ satisfait cette condition sur tout $\mathcal{X}$, et en intégrant elle permet d'obtenir $R_\omega(g) \geq 0$ quel que soit g.

Son risque pondéré vaut alors

$$L_\omega(g^*) = 2\mathbb{E}\left[\omega_1 \eta(X) \mathbb{1}_{\{\eta(X) \leq \omega_0\}} + \omega_0(1-\eta(X)) \mathbb{1}_{\{\eta(X) > \omega_0\}}\right]$$

$$
\begin{aligned}
&= 2\mathbb{E}\left[\omega_1(\eta(X) + \omega_0 - \omega_0)\mathbb{1}_{\{\eta(X)\leq\omega_0\}} + \omega_0(1 - \omega_0 - \eta(X) + \omega_0)\mathbb{1}_{\{\eta(X)>\omega_0\}}\right] \\
&= 2\omega_0\omega_1 - 2\mathbb{E}\left[|\eta(X) - \omega_0| \left(\omega_1\mathbb{1}_{\{\eta(X)\leq\omega_0\}} + \omega_0\mathbb{1}_{\{\eta(X)>\omega_0\}}\right)\right].
\end{aligned}
$$

On retrouve bien le risque de Bayes classique $\frac{1}{2} - \mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right|\right]$ lorsque $\omega_0 = \omega_1 = \frac{1}{2}$.

En remarquant que pour tout $x \in \mathcal{X}$ on a $\eta(x) \leq \omega_0$ ssi $\omega_1\eta(x) \leq \omega_0(1 - \eta(x))$, on peut aussi écrire le risque pondéré de g^* comme

$$
L_\omega(g^*) = \mathbb{E}\left[\min\left(\omega_1\eta(X), \omega_0(1 - \eta(X))\right)\right].
$$

L'intérêt de ce critère est de se prémunir plus fortement contre un type d'erreur en particulier (faux positifs ou faux négatifs). Par exemple, pour un test de grossesse, il est plus important de garantir un faible taux de faux négatifs que de faux positifs, puisqu'en cas de test négatif, une femme n'ira pas consulter de médecin.

Conseils bibliographiques

Vous trouverez ci-dessous quelques points d'entrée utiles pour l'apprentissage automatique :

- Théorique et porté sur les aspects probabilistes : [?]
- Utilitaire et porté sur les aspects pratiques : [?]
- Livre récent porté essentiellement sur l'aspect optimisation : [?] (et du même auteur sur l'apprentissage en ligne [?])
- Méthodes Bayésiennes et modèles graphiques : [?]