Apprentissage statistique

TD 1 : introduction à l'apprentissage statistique

Les exercices 1 et 3 sont fortement inspirés d'exercices du polycopié de Arnak Dalalyan, Apprentissage Statistique 3ème année ENSAE.

Exercice 1 Classification multi-label

Soit $\mathcal{Y} = \{a_1, \dots, a_K\}$ et $\mathcal{X} = \mathbb{R}^d$. Nous supposons que la loi P de (X, Y) admet une densité f par rapport à la mesure $\lambda_d \otimes \nu$ où λ_d est la mesure de Lebesgue sur $\mathbb{R}^d$ et ν la mesure de comptage sur $\mathcal{Y}$.

1. Rappeler que X a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ donnée par
$$f_X(x) = \sum_{k=1}^{K} f(x, a_k).$$
2. Montrer que le classifieur de Bayes est
$$g_P^*(x) \in \arg\max_{a \in \mathcal{Y}} f(x, a).$$
3. Dans le cas de la classification binaire où $K = 2$ et $a_1 = 0$, $a_2 = 1$, retrouver le classifieur de Bayes.

Answer of exercise 1

1. On rappelle que, comme f est la densité de (X, Y) par rapport à la mesure $\lambda_d \otimes \nu$,
$$f_X(x) = \int_{\mathcal{Y}} f(x, y) d\nu(y),$$
ce qui donne bien le résultat voulu.
2. D'après le cours
$$g_P^*(x) \in \arg\max_{a \in \mathcal{Y}} \mathbb{P}_P(Y = a | X = x).$$
Nous avons, pour tout $a \in \mathcal{Y}$, et $x \in \mathbb{R}^d$ tel que $f_X(x) > 0$,
$$\mathbb{P}_P(Y = a | X = x) = \frac{f(x, a)}{f_X(x)}$$
ce qui donne bien le résultat.
3. En classification binaire cela donne
$$g_P^*(x) = \mathbf{1}_{f(x,1) > f(x,0)}$$
ou
$$g_P^*(x) = \mathbf{1}_{f(x,1) \geq f(x,0)}$$
les deux définitions donnant des classifieurs de Bayes qui coïncident sauf aux points x tels que $f(x, 0) = f(x, 1)$.
calculons la fonction de régression. Nous avons
$$\eta_P^*(x) = \mathbb{E}_P[Y | X = x] = \mathbb{P}_P(Y = 1 | X = x) = \frac{f(x, 1)}{f_X(x)} = \frac{f(x, 1)}{f(x, 0) + f(x, 1)}.$$
Donc
$$f(x, 1) > f(x, 0) \iff \eta_P^*(x) > 1/2,$$
ce qui donne bien le résultat voulu.

Exercice 2 Regression linéaire

Soient $\mathcal{Y} = \mathbb{R}$ et $\mathcal{X} = \mathbb{R}^d$. On suppose qu'il existe $\beta \in \mathbb{R}^d$ tel que

$$Y = \beta^t X + \varepsilon,$$

où ε est une variable réelle, centrée, de variance σ^2 finie et indépendante de X.

1. Calculer la fonction de régression η_P^* et en déduire la fonction oracle g_P^*.
2. Quelle est la valeur du risque de Bayes R_P^* ?

Exercice 3

On considère le problème de classification binaire avec $Y \sim \mathcal{B}(p)$ et

$$\begin{aligned} X|Y=0 &\sim \mathcal{U}([0,1/2]), \\ X|Y=1 &\sim \mathcal{U}([0,1]). \end{aligned}$$

1. Déterminer la fonction de répartition de X et sa densité f_X par rapport à la mesure de Lebesgue.
2. Pour tout $x \in [0,1]$, calculer $\mathbb{E}_P[Y\mathbf{1}_{\{X\leq x\}}]$.
3. Montrer que, pour tout $x \in [0,1]$,

$$\mathbb{E}_P[Y\mathbf{1}_{\{X\leq x\}}] = \int_0^x \eta_P^*(u) f_X(u) du,$$

où $\eta_P^*(x) = \mathbb{E}_P[Y|X=x]$ est la fonction de régression.

4. Déterminer la loi conditionnelle de Y sachant $X = x$ ainsi que la forme du classifieur de Bayes.

Answer of exercise 3

1. X est à support dans $[0,1]$ donc sa fonction de répartition F_X vérifie $F_X(x) = 0$ si $x < 0$ et $F_X(x) = 1$ si $x > 1$. Soit $x \in [0,1]$

$$F_X(x) = \mathbb{P}_P(X \leq x) = \mathbb{P}_P(X \leq x|Y=0)\mathbb{P}_P(Y=0) + \mathbb{P}_P(X \leq x|Y=1)\mathbb{P}_P(Y=1).$$

Nous avons,

$$\mathbb{P}_P(X \leq x|Y=1) = \int_0^x du = x$$

et

— Si $x \leq 1/2$, alors

$$\mathbb{P}_P(X \leq x|Y=0) = \int_0^x 2du = 2x.$$

donc

$$F_X(x) = 2x(1-p) + xp = (2-p)x.$$

— Si $x > 1/2$, alors

$$\mathbb{P}_P(X \leq x|Y=0) = 1$$

donc

$$F_X(x) = 1 - p + xp.$$

Nous avons ensuite, pour $x \in]0, 1/2[$,

$$f_X(x) = F_X'(x) = 2 - p,$$

et pour $x \in]1/2, 1[$

$$f_X(x) = F_X'(x) = p.$$

La densité de X est donc p.p.

$$f_X(x) = (2-p)\mathbf{1}_{]0,1/2[}(x) + p\mathbf{1}_{[1/2,1[}(x).$$

2.

$$\mathbb{E}_P[Y\mathbf{1}_{\{X\leq x\}}] = \mathbb{P}_P(Y=1, X \leq x) = \mathbb{P}_P(X \leq x|Y=1)\mathbb{P}_P(Y=1) = xp.$$

3.

$$\mathbb{E}_P[Y\mathbf{1}_{\{X\leq x\}}] = \mathbb{E}_P[\mathbb{E}[Y\mathbf{1}_{\{X\leq x\}}|X]] = \mathbb{E}_P[\mathbf{1}_{\{X\leq x\}}\mathbb{E}[Y|X]] = \mathbb{E}_P[\mathbf{1}_{\{X\leq x\}}\eta_P^*(X)] = \int_0^1 \mathbf{1}_{\{u\leq x\}}\eta_P^*(u)f_X(u)du.$$

(on peut aussi utiliser la formule vue en cours).

4. Comme Y est à valeurs dans $\{0,1\}$, la loi de $Y|X$ est une loi de Bernoulli de paramètre

$$\eta_P^*(x) = \mathbb{P}(Y=1|X=x).$$

En dérivant la formule de la question 3., en utilisant le résultat de la question 2. et en remarquant que $f_X(0)=0$, on obtient, pour tout $x \in]0,1[$

$$\eta_P^*(x)f_X(x) = p.$$

D'où par 1.,

$$\eta_P^*(x) = \frac{p}{2-p}, \text{ pour } x \in]0,1/2[$$

et

$$\eta_P^*(x) = 1, \text{ pour } x \in [1/2,1].$$

Puisqu'il s'agit d'un problème de classification binaire, le classifieur de Bayes s'écrit

$$g_P^*(x) = \mathbf{1}_{\{\eta_P^*(x)>1/2\}}.$$

Donc

si $x < 1/2$

$$g_P^*(x) = \mathbf{1}_{\{p/(2-p)>1/2\}} = \mathbf{1}_{\{p>3/2\}};$$

Si $x \geq 1/2$

$$g_P^*(x) = \mathbf{1}_{\{1>1/2\}} = 1.$$

Exercice 4 Classifieur bayésien

Soit (X,Y) un couple de loi P avec $Y \in \{0,1\}$. Soit g_P^* le classifieur de Bayes, et R_P^* son risque de classification. On note également $p = \mathbb{P}(Y=1)$.

1. Montrer que $R_P^* \leq \min\{p, 1-p\}$.
2. Montrer que si X et Y sont indépendants, $R_P^* = \min\{p, 1-p\}$.
3. Fabriquer un exemple où $R_P^* = \min\{p, 1-p\}$ et où X et Y ne sont pas indépendants. Indication : considérer X, Y binaires dépendants.

Answer of exercise 4

Rappel notation : $\mathcal{X} \subset \mathbb{R}^d$ espace des covariables X ; $i = 1, \ldots, n$ indice des observations ; X_i vecteur des covariables et Y_i variable réponse du $i$ème individu ; $\mathcal{D}_n = \{(X_1,Y_1), \ldots, (X_n,Y_n)\}$. l'espace d'échantillonage.

0. Voir page 223 de BiauDevroye.pdf déjà vu en cours

Un classifieur binaire est une fonction mesurable $h : \mathcal{X} \times \mathcal{D}_n \mapsto \{0,1\}$. $h(.,\mathcal{D}_n)$ détermine la règle de classification issue des données $\mathcal{D}_n$.
La fonction de régression pour un problème de classification est $\eta(x) = E(Y|X=x) = P(Y=1|X=x)$.
Le risque de classification pour un classifieur h conditionnel est $R_{\mathcal{D}_n}(h) = P(h(X) \neq Y|\mathcal{D}_n)$ tandis que son risque de classification (inconditionnel) est $R(h) = E(P(h(X) \neq Y|\mathcal{D}_n)) = P(h(X) \neq Y)$. Dans ce cas binaire pour tout classifieur h le risque est

$$\begin{aligned} R(h) &= E(1_{h(X)\neq Y}) = E(1_{h(X)=0,Y=1} + 1_{h(X)=1,Y=0}) \\ &= E(E(1_{h(X)=0,Y=1}|X)) + E(E(1_{h(X)=1,Y=0}|X))E(1_{h(X)=0}E(1_{Y=1}|X)) + E(1_{h(X)=1}E(1_{Y=0}|X)) \\ &= E\left(\eta(X)1_{h(X)=0} + (1-\eta(X))1_{h(X)=1}\right) \end{aligned}$$

1. Ici on se place en l'absence de données puisqu'on étudie le classifieur $g^\star$ de Bayes. Donc son risque vaut

$$R^\star = E(P(g^\star(X) \neq Y|X)) = E\left(P(Y=1|X)1_{g^\star(X)=0} + P(Y=0|X)1_{g^\star(X)=1}\right)$$

Le classifieur de Bayes est le classifeur issue de la vraie loi de la variable réponse, c'est à dire

$$g^\star(x) = \begin{cases} 1 & \text{si } P(Y=1|X=x) > P(Y=0|X=x) \\ 0 & \text{sinon} \end{cases}$$

Or $P(Y=1|X=x) > P(Y=0|X=x) \Leftrightarrow 2P(Y=1|X=x) > 1 \Leftrightarrow P(Y=1|X=x) > 1/2$. Donc

$$g^\star(x) = \begin{cases} 1 & \text{si } P(Y=1|X=x) > 1/2 \\ 0 & \text{sinon} \end{cases}$$

Réécriture avant majoration - Méthode 1 :
Déterminons son risque

$$\begin{aligned} R(g^\star) &= E\left(P(Y=1|X)1_{g^\star(X)=0} + (1-P(Y=1|X))1_{g^\star(X)=1}\right) \\ &= E\left(P(Y=1|X)1_{P(Y=1|X)\leq 1/2} + (1-P(Y=1|X))1_{P(Y=1|X)>1/2}\right) \\ &= E\left(\min(P(Y=1|X), 1-P(Y=1|X))\right) \end{aligned}$$

Réécriture avant majoration - Méthode 2 :
Déterminons son risque

$$R(g^\star) = E(P(g^\star(X) \neq Y|X)) = P(g^\star(X) \neq Y) = 1 - P(g^\star(X) = Y)$$

puisque $g^\star$ est indépendant des données. De plus $\{g^\star(X) = Y|X\} = \{g^\star(X) = 0 \cap Y = 0|X\} \cup \{g^\star(X) = 1 \cap Y = 1|X\}$ Donc

$$\begin{aligned} P(g^\star(X) = Y|X) &= P(g^\star(X) = 0 \cap Y = 0|X) + P(\{g^\star(X) = 1 \cap Y = 1|X) \\ &= 1_{g^\star(X)=0}P(Y=0|X) + 1_{g^\star(X)=1}P(Y=1|X) \end{aligned}$$

Or la variable aléatoire $P(g^\star(X) = Y|X)$ est une variable avec deux valeurs possibles de probabilités.
— Si $g^\star(X) = 0$ alors

$$P(Y=0|X) > 1/2 > P(Y=1|X) \Rightarrow P(g^\star(X) = Y|X) = \max(P(Y=0|X), P(Y=1|X)).$$

— Si $g^\star(X) = 1$ alors

$$P(Y=1|X) > 1/2 > P(Y=0|X) \Rightarrow P(g^\star(X) = Y|X) = \max(P(Y=0|X), P(Y=1|X)).$$

Ainsi $P(g^\star(X) = Y|X) = \max(P(Y=0|X), P(Y=1|X)) \Rightarrow P(g^\star(X) = Y) = E\left(\max(P(Y=0|X), P(Y=1|X))\right)$.
Donc

$$R(g^\star) = 1 - E\left(\max(P(Y=0|X), P(Y=1|X))\right) = E\left(\min(P(Y=0|X), P(Y=1|X))\right)$$

Majoration :
Comme

$$E\left(\min(P(Y=1|X), 1-P(Y=1|X))\right) = \int_{\mathbb{R}} \min(P(Y=1|X=x), P(Y=0|X=x))dF_X(x)$$

$$R(g^\star) \begin{array}{l} \leq \int_{\mathbb{R}} P(Y=1|X=x)dF_X(x) = P(Y=1) \\ \leq \int_{\mathbb{R}} P(Y=0|X=x)dF_X(x) = P(Y=0) \end{array} \leq \min(P(Y=1), P(Y=0))$$

On obtient la majoration du risque

$$R(g^\star) \leq \min(P(Y=1), 1-P(Y=1)) = \min(p, 1-p).$$

On a aussi

$$R(g^\star) = E\left(\min(P(Y=1|X), 1-P(Y=1|X))\right) = E(\min(\eta^\star(X), 1-\eta^\star(X)))$$

avec $\eta^\star(X) = P(Y=1|X)$.

2. Si indépendance alors $P(Y = .|X) = P(Y = .)$ et donc $g^\star(x) = 1_{p>1/2}$ est déterministe et

$$R(g^\star) = P(g^\star(X)) = 0)P(Y = 1) + P(g^\star(X)) = 1)P(Y = 0) = 1_{p\leq 1/2}p + 1_{p>1/2}(1-p) = \min(p, 1-p)$$

3. Il nous faut choisir $p \notin \{0, 1/2, 1\}$. En effet
si $p = 0$ alors $Y = 0$ p.s. donc Y est indépendant de X ;
si $p = 1$ alors $Y = 1$ p.s. donc Y est indépendant de X ;
si $p = 1/2$ alors d'après le cours $\eta(x) = 1/2$ et $R^\star = 1/2$ donc Y est indépendant de X.
Cherchons un cas dépendance
Fixons une covariable binaire $X \in \{0, 1\}$. Autrement dit on cherche les probabilités élémentaires suivantes :

	conditionnelle		
	$X = 0$	$X = 1$	incond.
$Y = 0$	$P(Y = 0\|X = 0) = p_{0\|0}$	$P(Y = 0\|X = 1) = p_{0\|1}$	$1 - p$
$Y = 1$	$P(Y = 1\|X = 0) = p_{1\|0}$	$P(Y = 1\|X = 1) = p_{1\|1}$	p
	$p_{0\|0} + p_{1\|0} = 1$	$p_{0\|1} + p_{1\|1} = 1$	

La dépendance entraine $p_{0|0} \neq p_{0|1}$ ou $p_{0|0} \neq p$. Dans ce cadre, le classifieur de Bayes se simplifie

$$g^\star(x) = \begin{cases} 1 & \text{si } p_{1|x} > 1/2 \\ 0 & \text{sinon} \end{cases}$$

Deux exemples simples ci-dessous

	$X = 0$	$X = 1$
$Y = 0$	$3/8 = p_{0\|0}$	$7/8 = p_{0\|1}$
$Y = 1$	$5/8 = p_{1\|0}$	$1/8 = p_{1\|1}$
$g^\star(X)$	1	0
$\eta^\star(X)$	5/8	1/8

,

	$X = 0$	$X = 1$
$Y = 0$	$4/8 = p_{0\|0}$	$3/8 = p_{0\|1}$
$Y = 1$	$4/8 = p_{1\|0}$	$7/8 = p_{1\|1}$
$g^\star(X)$	0	1
$\eta^\star(X)$	4/8	7/8

Dans le premier cas, on a

		$P(X = 1)$	
	général	1/2	1/4
$R^\star$	$3/8P(X = 0) + 1/8P(X = 1)$	1/4	5/16
p	$5/8P(X = 0) + 1/8P(X = 1)$	3/8	1/2
$1 - p$		5/8	1/2
$\min(p, 1 - p)$		3/8	1/2

Dans le second cas, on a

		$P(X = 1)$	
	général	1/2	1/4
$R^\star$	$1/2P(X = 0) + 1/8P(X = 1)$	5/16	13/32
p	$4/8P(X = 0) + 7/8P(X = 1)$	11/16	19/32
$1 - p$		5/16	3/32
$\min(p, 1 - p)$		5/16	3/32

Déterminons la conditionnelle $P(Y = .|X = .)$ dans le cas général à l'aide de deux équations par la formule des probabilités totales

$$P(X = 0)p_{1|0} + P(X = 1)p_{1|1} = p$$

$$P(X = 0)p_{0|0} + P(X = 1)p_{0|1} = 1 - p$$

Dans ce cas binaire on a $R^\star = E(\min(p_{1|X}, 1 - p_{1|X}))$. On impose

$$\min(p, 1-p) = \min(p_{1|0}, 1-p_{1|0})P(X = 0) + \min(p_{1|1}, 1-p_{1|1})P(X = 1).$$

Prenons $p = P(Y = 1) = 1/4$ on a donc

$$P(X = 0)p_{1|0} + P(X = 1)p_{1|1} = 1/4$$

$$P(X=0)p_{0|0} + P(X=1)p_{0|1} = 3/4$$

De plus, on veut une loi conditionnelle tel que $R^\star = \min(p, 1-p) = 1/4$.

$$1/4 = \min(p_{1|0}, 1-p_{1|0})P(X=0) + \min(p_{1|1}, 1-p_{1|1})P(X=1).$$

On obtient un système de 3 équations à 3 inconnues $p_{1|0}, p_{1|1}, P(X=0)$:

$$\begin{cases} P(X=0)p_{1|0} + P(X=1)p_{1|1} = 1/4 \\ P(X=0)p_{0|0} + P(X=1)p_{0|1} = 3/4 \\ \min(p_{1|0}, 1-p_{1|0})P(X=0) + \min(p_{1|1}, 1-p_{1|1})P(X=1) = 1/4 \end{cases}$$

Supposons $P(X=1) = 1/2 = P(X=0)$

$$\begin{cases} p_{1|0} + p_{1|1} = 1/2 \\ p_{0|0} + p_{0|1} = 3/2 \\ \min(p_{1|0}, 1-p_{1|0}) + \min(p_{1|1}, 1-p_{1|1}) = 1/2 \end{cases}$$

Si on choisit $p_{1|1} < 1/2 < 1-p_{1|1}$ alors on obtient

$$\begin{cases} p_{1|0} = 1/2 - p_{1|1} \\ \min(p_{1|0}, 1-p_{1|0}) = 1/2 - p_{1|1} \end{cases} \Rightarrow p_{1|0} = \min(p_{1|0}, 1-p_{1|0}) \Rightarrow p_{1|0} < 1/2.$$

Par exemple $p_{1|1} = 1/8$ donc $p_{1|0} = 3/8$ ainsi

		conditionnelle		inconditionnelle
Proba.		$X=0$	$X=1$	
	$Y=0$	$5/8 = p_{0\|0}$	$7/8 = p_{0\|1}$	$1-p = \frac{1}{2}\frac{5}{8} + \frac{1}{2}\frac{7}{8} = 3/4$
	$Y=1$	$3/8 = p_{1\|0}$	$1/8 = p_{1\|1}$	$p = \frac{1}{2}\frac{3}{8} + \frac{1}{2}\frac{1}{8} = 1/4$
Risque	$g^\star(X)$	0	0	
	$\eta^\star(X)$	$3/8$	$1/8$	
	$R^\star$	$3/8 \times 1/2 + 1/8 \times 1/2$		$\min(1-p, p) = 1/4$

Exercice 5 Classifieur bayésien (suite)

Soit f_0 (resp. f_1) la densité de X conditionnellement à $Y=0$ (resp. $Y=1$) c'est-à-dire que pour $j=0,1$, pour toute fonction borélienne φ, $\mathbb{E}[\varphi(X)|Y=j] = \int_{\mathbb{R}^d} \varphi(x) f_j(x) dx$. Soit $p = \mathbb{P}(Y=1)$.

0. Montrer que

   (a)
   $$f(x) = p f_1(x) + (1-p) f_0(x).$$

   (b) En déduire une écriture de $\eta_P^*(x)$ en fonction de f_0, f_1 et p.

1. Écrire R_P^* en fonction de f_0, f_1 et p.
2. Dans le cas $p = 1/2$, en déduire que

$$R_P^* = \frac{1}{2} - \frac{1}{4} \int_{\mathbb{R}^d} |f_0(x) - f_1(x)| dx.$$

On pourra utiliser la formule $\min\{a,b\} = \frac{a+b}{2} - \frac{|a-b|}{2}$.

Answer of exercise 5

0. (a) **Méthode 1**
   Soit $\varphi : \mathbb{R}^d \to \mathbb{R}$ une fonction continue à support compact. Nous avons

$$\begin{aligned} \mathbb{E}[\varphi(X)] &= \mathbb{E}[\varphi(X)|Y=1]\mathbb{P}(Y=1) + \mathbb{E}[\varphi(X)|Y=0]\mathbb{P}(Y=0) \\ &= \int_{\mathbb{R}^d} p\varphi(x) f_1(x) + (1-p)\varphi(x) f_0(x) dx, \end{aligned}$$

   d'où le résultat.

Méthode 2

$$P(X \leq x, Y = y) = P(X \leq x | Y = y)P(Y = y) = F_X^{Y=y}(x)p^y(1-p)^{1-y}$$

$$\Rightarrow F_X(x) = \sum_y P(X \leq x, Y = y) = P(X \leq x, Y = 1) + P(X \leq x, Y = 0) = F_X^{Y=1}(x)p + F_X^{Y=0}(x)(1-p)$$

$$\Rightarrow f_X(x) = \frac{\partial F_X(x)}{\partial x} = pf_1(x) + (1-p)f_0(x).$$

(b) Par définition de l'espérance conditionnelle, la fonction η est la seule fonction mesurable telle que, pour toute fonction mesurable $\varphi : \mathbb{R}^d \to \mathbb{R}$ nous avons

$$\mathbb{E}[\varphi(X)\eta(X)] = \mathbb{E}[\varphi(X)\mathbf{1}_{\{Y=1\}}].$$

Or

$$\mathbb{E}[\varphi(X)\mathbf{1}_{\{Y=1\}}] = \mathbb{E}[\varphi(X)|Y = 1]\mathbb{P}(Y = 1) = p\int_{\mathbb{R}^d} \varphi(x)f_1(x)dx.$$

Comme

$$\mathbb{E}[\varphi(X)\eta(X)] = \int_{\mathbb{R}^d} \varphi(x)\eta(x)f(x)dx,$$

et que le résultat est vrai pour toute fonction borélienne. Cela nous donne

$$\eta(x)f(x) = pf_1(x) \text{ ou encore } \eta(x) = \frac{pf_1(x)}{f(x)},$$

en tout point x où f ne s'annule pas. C'est la formule de Bayes appliquée à notre cas particulier.

1. Par le résultat de la question 0.(ii),

$$R_P^* = \int_{\mathbb{R}^d} \min\{\eta(x), 1-\eta(x)\}f(x)dx = \int_{\mathbb{R}^d} \min\left\{\frac{pf_1(x)}{f(x)}, 1 - \frac{pf_1(x)}{f(x)}\right\} f(x)dx$$

Puis par 0.(i), $1 - \frac{pf_1(x)}{f(x)} = \frac{(1-p)f_0(x)}{f(x)}$. D'où

$$R_P^* = \int_{\mathbb{R}^d} \min\left\{pf_1(x), (1-p)f_0(x)\right\} dx.$$

2. Dans le cas $p = 1/2$, nous avons

$$R_P^* = \frac{1}{2}\int_{\mathbb{R}^d} \min\left\{f_1(x), f_0(x)\right\} dx = \frac{1}{2}\int_{\mathbb{R}^d} \frac{f_0(x) + f_1(x)}{2} - \frac{|f_0(x) - f_1(x)|}{2} dx,$$

et le résultat provient du fait que, si $p = 1/2$, $\frac{f_0(x)+f_1(x)}{2} = f(x)$.
Le risque de Bayes est donc lié à la distance entre f_0 et f_1 : plus la distance L^1 entre f_0 et f_1 est élevée et plus le risque de Bayes est petit.

Exercice 6 Un classifieur universellement consistant

Soit (X, Y) un couple de loi $\mathbb{P}$ avec $X \in \{1, ..., 5\}$ et $Y \in \{0, 1\}$. On note toujours g_P^* le classifieur de Bayes et R_P^* son risque de classification. On se propose d'étudier la règle de classification suivante, étant donné $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$,

$$\hat{g}(x) = \mathbf{1}_{\{\hat{\eta}(x) > \frac{1}{2}\}},$$

avec

$$\hat{\eta}(x) = \frac{1}{\text{card}\{i, X_i = x\}} \sum_{i, X_i = x} Y_i$$

et $\hat{\eta}(x) = 0$ si $\text{card}\{i, X_i = x\} = 0$.

1. Déterminer $\hat{\eta}$ et $\hat{g}$ sur les données suivantes $\mathcal{D}_7$:

i	1	2	3	4	5	6	7
X_i	3	4	4	3	3	4	5
Y_i	0	1	0	1	0	1	0

2. Montrer que $\hat{g}$ est universellement consistante.

Answer of exercise 6

1. $\hat{\eta}(x, \mathcal{D}_n)$ est en fait la moyenne empirique de la variable réponse Y quand la variable explicative X vaut x. Notons $n_x = \text{card}\{i, X_i = x\} \leq n$ par construction et $\sum_x n_x = n$. Typiquement pour le tableau suivant, on obtient

i	X_i	Y_i
1	3	0
2	4	1
3	4	0
4	3	1
5	3	0
6	4	1
7	5	0

$$\Rightarrow \hat{\eta}(x, \mathcal{D}_7) = \begin{cases} 1/3 & \text{si } x = 3 \\ 2/3 & \text{si } x = 4 \\ 0 & \text{si } x = 5 \end{cases} \Rightarrow g(x, \mathcal{D}_7) = \begin{cases} 0 & \text{si } x = 3 \\ 1 & \text{si } x = 4 \\ 0 & \text{si } x = 5 \end{cases}$$

2. Par la loi des grands nombres, quand $P(X = x) > 0$, i.e. $P(X = 0) \in (0, 1), P(X = 1) \in (0, 1)$ on a

$$\frac{1}{n}\sum_{i=1}^{n} 1_{X_i = x} Y_i \xrightarrow[n \to +\infty]{p.s.} E(Y 1_{X=x}),\ \frac{1}{n}\sum_{i=1}^{n} 1_{X_i = x} \xrightarrow[n \to +\infty]{p.s.} E(1_{X=x}) = P(X = x)$$

Or $E(Y 1_{X=x}) = E(1_{X=x})E(Y|X = x) = P(X = x)\eta(x)$. En effet pour A un évènement tel que $P(A) > 0$, la variable conditionnelle $Z|A$ a pour fonction de répartition

$$P(Z \leq x|A) = \frac{P(Z \leq x \cap A)}{P(A)} = F_Z^A(x),$$

et pour espérance

$$\begin{cases} E(Z|A) = \int_{\mathbb{R}} x dF_Z^A(x) = \int_{\mathbb{R}} x \frac{dP(Z \leq x \cap A)}{P(A)} = \int_{\mathbb{R}} x 1_A(x) \frac{dF_Z(x)}{P(A)} \\ E(Z 1_A) = \int_{\mathbb{R}} x 1_A(x) dF_Z(x) \end{cases} \Rightarrow E(Z|A) = \frac{E(Z 1_A)}{P(A)}.$$

On a bien $E(Z 1_A) = P(A)E(Z|A)$ qui est vrai aussi pour $P(A) = 0$.

Notons $n_X = \text{card}\{i = 1, \ldots, n, X_i = X\} \leq n$ le cardinal (aléatoire) du nombres d'observations tel que la variable explicative vaut X.

Etudions l'écart absolu moyen

$$\begin{aligned} E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|) &= E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \times 1_{n_X \neq 0}) + E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \times 1_{n_X = 0}) \\ &= E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \times 1_{n_X \neq 0}) + E(\eta(X) \times 1_{n_X = 0}) \\ &\leq E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|\ |n_X \neq 0) E(1_{n_X \neq 0}) + E(1_{n_X = 0}) \\ &= E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|\ |n_X \neq 0)(1 - P(n_X = 0)) + P(n_X = 0) \end{aligned}$$

Or $P(n_X = 0) \underset{n \to +\infty}{\longrightarrow} 0$ et $E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|\ |n_X \neq 0) \underset{n \to +\infty}{\longrightarrow} 0$ donc

$$E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|) \leq E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|\ |n_X \neq 0)(1 - P(n_X = 0)) + P(n_X = 0) \underset{n \to +\infty}{\longrightarrow} 0.$$

Exercice 7 Classification et maximum de vraisemblance

On suppose que la loi P du couple (X, Y) est donnée de la manière suivante : $Y \sim \mathcal{B}(p)$ pour un paramètre $p \in]0, 1[$ puis $X \in \mathbb{R}^d$ est donnée par sa loi conditionnelle sachant Y :

$$X|Y \sim \mathcal{N}(V_Y, \Sigma)$$

où Σ est une matrice définie positive et V_0, V_1 sont deux vecteurs distincts de $\mathbb{R}^d$.

1. Donner la loi jointe de (X, Y) ainsi que les lois marginales.
2. Déterminer la fonction de régression η_P^*.

3. En déduire la forme du classifieur de Bayes, g_P^*, et montrer que son risque de classification s'écrit

$$R_P^* = pP\left(Z > \delta/2 + \delta^{-1}\log(\frac{p}{1-p})\right) + (1-p)P\left(Z < -\delta/2 + \delta^{-1}\log(\frac{p}{1-p})\right)$$

où $\delta = \left\|\Sigma^{-1/2}(V_1 - V_0)\right\|$ et $Z \sim \mathcal{N}(0,1)$.

4. En pratique, on dispose d'un n-échantillon (X_i, Y_i), $1 \leq i \leq n$, de la loi P. On suppose que l'on est dans un cas où l'on connaît p et Σ et on se propose d'estimer V_0 et V_1 par maximum de vraisemblance. Donner la forme des estimateurs $\hat{V}_0$ et $\hat{V}_1$ ainsi obtenus.
5. En déduire un estimateur de la fonction de régression, $\hat{\eta}$ et une règle de classification, $\hat{g}$.
6. Montrer que $R(\hat{g}) \to R_P^*$ en probabilité, quand $n \to \infty$.
7. Montrer que $\hat{g}$ n'est pas universellement consistante. Il suffira de fabriquer une autre loi P' telle que si (X_i, Y_i) et (X, Y) sont iid de loi P', alors $R(\hat{g})$ ne converge pas vers R_P^*.

Answer of exercise 7

1. La loi jointe de (X, Y) est caractérisée par sa densité par rapport à la mesure $\lambda_d \otimes \mu$ où λ_d est la mesure de Lebesgue sur $\mathbb{R}^d$ et μ est la mesure de comptage sur $\{0,1\}$ ($\mu(S) = \text{card}(S)$ pour tout $S \subseteq \{0,1\}$). Cette densité est une fonction $f_{X,Y} : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}$ telle que, pour toute fonction test mesurable $\varphi : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[\varphi(X,Y)] &= \int_{\mathbb{R}^d \times \{0,1\}} \varphi(x,y) f_{X,Y}(x,y) d(\lambda_d \otimes d\mu)(x,y) \\
&= \sum_{j \in \{0,1\}} \int_{\mathbb{R}^d} \varphi(x,j) f_{X,Y}(x,j) dx \\
&= \int_{\mathbb{R}^d} \varphi(x,0) f_{X,Y}(x,0) dx + \int_{\mathbb{R}^d} \varphi(x,1) f_{X,Y}(x,1) dx.
\end{aligned}$$

Or

$$\begin{aligned}
\mathbb{E}[\varphi(X,Y)] &= \mathbb{E}[\varphi(X,Y)|Y=0] \times \mathbb{P}(Y=0) + \mathbb{E}[\varphi(X,Y)|Y=1] \times \mathbb{P}(Y=1) \\
&= \int_{\mathbb{R}^d} \varphi(x,0) f_X^{Y=0}(x) dx \times (1-p) + \int_{\mathbb{R}^d} \varphi(x,1) f_X^{Y=0}(x) dx \times p
\end{aligned}$$

où

$$f_X^{Y=y}(x) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{(x-V_y)^T \Sigma^{-1}(x-V_y)}{2}\right)$$

est la loi densité de X conditionnellement à $Y = y$ (pour $y \in \{0,1\}$). Ce qui nous donne

$$f_{X,Y}(x,y) = \begin{cases} (1-p) f_X^{Y=0}(x) & \text{si } y = 0 \\ p f_X^{Y=1}(x) & \text{si } y = 1. \end{cases}$$

Pour obtenir les lois marginales il suffit d'intégrer la densité jointe. Intégrer par rapport à la mesure de comptage revient à sommer sur toutes les valeurs possibles. On obtient

$$f_X(x) = (1-p) f_X^{Y=0}(x) + p f_X^{Y=1}(x)$$

pour la loi de X. Pour Y, d'après l'énoncé, nous avons,

$$P(Y=1) = p, P(Y=0) = 1-p.$$

2. pour un problème de classification binaire, on a, d'après la formule de Bayes,

$$\eta(x) = E(Y|X=x) = P(Y=1|X=x) = \frac{p f_X^{Y=1}(x)}{f_X^{Y=1}(x)p + f_X^{Y=0}(x)(1-p)}$$

3. dans ce cadre le classificateur de Bayes est

$$g^\star(x) = 1_{\eta(x)>1/2} = 1_{f_X^{Y=1}(x)p > f_X^{Y=0}(x)(1-p)}$$

Son risque vaut

$$\begin{aligned} R^\star &= E(\min(\eta(X), 1-\eta(X))) = E\left(\min\left(\frac{pf_X^{Y=1}(X)}{f_X(X)}, 1-\frac{pf_X^{Y=1}(X)}{f_X(X)}\right)\right) \\ &= E\left(\min\left(\frac{pf_X^{Y=1}(X)}{f_X(X)}, \frac{(1-p)f_X^{Y=0}(X)}{f_X(X)}\right)\right) = E\left(\frac{\min\left(pf_X^{Y=1}(X), (1-p)f_X^{Y=0}(X)\right)}{f_X(X)}\right) \\ &= \int \frac{\min\left(pf_X^{Y=1}(x), (1-p)f_X^{Y=0}(x)\right)}{f_X(x)} f_X(x)dx = \int \min\left(pf_X^{Y=1}(x), (1-p)f_X^{Y=0}(x)\right) dx \end{aligned}$$

Calculons le minimum

$$\begin{gathered} pf_X^{Y=1}(x) < (1-p)f_X^{Y=0}(x) \\ \Leftrightarrow p\frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\left(-\frac{(x-V_1)^T\Sigma^{-1}(x-V_1)}{2}\right) < (1-p)\frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\left(-\frac{(x-V_0)^T\Sigma^{-1}(x-V_0)}{2}\right) \\ \Leftrightarrow \log\frac{p}{1-p} - \frac{(x-V_1)^T\Sigma^{-1}(x-V_1)}{2} < -\frac{(x-V_0)^T\Sigma^{-1}(x-V_0)}{2} \\ \Leftrightarrow 2\log\frac{p}{1-p} < (x-V_1)^T\Sigma^{-1}(x-V_1) - (x-V_0)^T\Sigma^{-1}(x-V_0) \end{gathered}$$

Or

$$\begin{aligned} (x-V_0)^T\Sigma^{-1}(x-V_0) &= (x-V_1+V_1-V_0)^T\Sigma^{-1}(x-V_1+V_1-V_0) \\ &= (x-V_1)^T\Sigma^{-1}(x-V_1+V_1-V_0) + (V_1-V_0)^T\Sigma^{-1}(x-V_1+V_1-V_0) \\ &= (x-V_1)^T\Sigma^{-1}(x-V_1) + 2(x-V_1)^T\Sigma^{-1}(V_1-V_0) + \underbrace{(V_1-V_0)^T\Sigma^{-1}(V_1-V_0)}_{\delta^2} \end{aligned}$$

Ainsi

$$pf_X^{Y=1}(x) < (1-p)f_X^{Y=0}(x) \Leftrightarrow 2\log\frac{p}{1-p} < -2(x-V_1)^T\Sigma^{-1}(V_1-V_0) - \delta^2$$

Autrement dit

$$\begin{aligned} R_P^* &= \int 1_{2\log\frac{p}{1-p}+\delta^2<-2(x-V_1)^T\Sigma^{-1}(V_1-V_0)} pf_X^{Y=1}(x)dx + \int 1_{2\log\frac{p}{1-p}+\delta^2>-2(x-V_1)^T\Sigma^{-1}(V_1-V_0)}(1-p)f_X^{Y=0}(x)dx \\ &= pE\left(1_{2\log\frac{p}{1-p}+\delta^2<-2(X-V_1)^T\Sigma^{-1}(V_1-V_0)}|Y=1\right) + (1-p)E\left(1_{2\log\frac{p}{1-p}+\delta^2>-2(X-V_1)^T\Sigma^{-1}(V_1-V_0)}|Y=0\right) \\ &= \underbrace{p\,P(2\log\frac{p}{1-p}+\delta^2 < -2(X-V_1)^T\Sigma^{-1}(V_1-V_0)|Y=1)}_{p_{1,X}} \\ &\quad \underbrace{+(1-p)\,P(2\log\frac{p}{1-p}+\delta^2 > -2(X-V_1)^T\Sigma^{-1}(V_1-V_0)|Y=0)}_{p_{0,X}} \end{aligned}$$

Comme $X|Y=1 \sim \mathcal{N}(V_1, \Sigma)$ on a

$(V_1-V_0)^T\Sigma^{-1}(X-V_1)|Y=1 \sim \mathcal{N}((V_1-V_1)^T\Sigma^{-1}(V_1-V_0), (V_1-V_0)^T\Sigma^{-1}\Sigma\Sigma^{-1}(V_1-V_0)) = \mathcal{N}(0,\delta^2)$.

Donc $(V_1-V_0)^T\Sigma^{-1}(X-V_1)|Y=1$ a même loi que δZ avec $Z \sim \mathcal{N}(0,1)$ et

$$\begin{aligned} p_{1,X} &= P(2\log(\frac{p}{1-p}) + \delta^2 < -2\delta Z) = P(\log(\frac{p}{1-p})/\delta + \delta/2 < -Z) \\ &= P(\log(\frac{p}{1-p})/\delta + \delta/2 < Z) \end{aligned}$$

Comme $(V_1-V_0)^T\Sigma^{-1}(X-V_1) = (V_1-V_0)^T\Sigma^{-1}(X-V_0) - \delta^2$ on a

$$(V_1-V_0)^T\Sigma^{-1}(X-V_1)|Y=0 = (V_1-V_0)^T\Sigma^{-1}(X-V_0) - \delta^2|Y=0 \sim \mathcal{N}(-\delta^2, \delta^2).$$

Donc $(V_1-V_0)^T\Sigma^{-1}(X-V_1)|Y=0$ a même loi que $-\delta^2 + \delta Z$ avec $Z \sim \mathcal{N}(0,1)$ et

$$\begin{aligned} p_{0,X} &= P(2\log\frac{p}{1-p} + \delta^2 > 2\delta^2 - 2\delta Z)) = P(2\log\frac{p}{1-p} - \delta^2 > -2\delta Z)) \\ &= P(\log\frac{p}{1-p}/\delta - \delta/2 > -Z)) = P(\log\frac{p}{1-p}/\delta - \delta/2 > Z)) \end{aligned}$$

Ainsi

$$R_P^* = pP\left(Z > \delta/2 + \delta^{-1}\log(\frac{p}{1-p})\right) + (1-p)P\left(Z < -\delta/2 + \delta^{-1}\log(\frac{p}{1-p})\right)$$

4. Nous pouvons réécrire la densité du couple (X,Y) de façon suivante

$$\begin{aligned} f_{X,Y}(x,y) &= \left(f_X^{Y=1}(x)p\right)^y \left(f_X^{Y=0}(x)(1-p)\right)^{1-y} \\ &= \left(\frac{p}{\sqrt{2\pi\det(\Sigma)}}\exp\left(-\frac{(x-V_1)^T\Sigma^{-1}(x-V_1)}{2}\right)\right)^y \left(\frac{1-p}{\sqrt{2\pi\det(\Sigma)}}\exp\left(-\frac{(x-V_0)^T\Sigma^{-1}(x-V_0)}{2}\right)\right)^{1-y} \\ &= \frac{p^y(1-p)^{1-y}}{\sqrt{2\pi\det(\Sigma)}}\exp\left(-y\frac{(x-V_1)^T\Sigma^{-1}(x-V_1)}{2} - (1-y)\frac{(x-V_0)^T\Sigma^{-1}(x-V_0)}{2}\right) \end{aligned}$$

La log-vraisemblance s'écrit donc

$$\begin{aligned} \mathcal{L}(V_0,V_1) &= -n\log(\sqrt{2\pi\det(\Sigma)}) + \sum_{i=1}^n y_i\log(p) + \sum_{i=1}^n(1-y_i)\log(1-p) \\ &\quad - \sum_{i=1}^n y_i\frac{(x_i-V_1)^T\Sigma^{-1}(x_i-V_1)}{2} - \sum_{i=1}^n(1-y_i)\frac{(x_i-V_0)^T\Sigma^{-1}(x_i-V_0)}{2} \end{aligned}$$

$$\frac{\partial\mathcal{L}(V_0,V_1)}{\partial V_0} = -\sum_{i=1}^n(1-y_i)\Sigma^{-1}(x_i-V_0) = -\Sigma^{-1}\sum_{i=1}^n(1-y_i)(x_i-V_0)$$

Annulons la dérivée en utilisant Σ^{-1} est inversible (puisque c'est l'inverse de Σ)

$$\frac{\partial\mathcal{L}(V_0,V_1)}{\partial V_0} = 0 \Leftrightarrow \sum_{i=1}^n(1-y_i)(x_i-V_0) = 0 \Leftrightarrow V_0 = \frac{\sum_{i=1}^n(1-y_i)x_i}{n-\sum_{i=1}^n y_i}$$

De même

$$\frac{\partial\mathcal{L}(V_0,V_1)}{\partial V_1} = -\Sigma^{-1}\sum_{i=1}^n y_i(x_i-V_1) \Rightarrow V_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n y_i}$$

Les deux estimateurs sont donc

$$\hat{V}_0 = \frac{\sum\limits_{i=1}^n(1-Y_i)X_i}{n-\sum\limits_{i=1}^n Y_i} = \frac{\sum\limits_{i,Y_i=0} X_i}{n-\sum\limits_{i=1}^n Y_i},\ \hat{V}_1 = \frac{\sum\limits_{i=1}^n Y_iX_i}{\sum\limits_{i=1}^n Y_i} = \frac{\sum\limits_{i,Y_i=1} X_i}{\sum\limits_{i,Y_i=1} Y_i}.$$

5. en utilisant la question 1, on a la fonction de régression

$$\hat{\eta}(x) = \frac{pf_X^{Y=1}(x;\hat{V}_1)}{f_X^{Y=1}(x;\hat{V}_1)p + f_X^{Y=0}(x;\hat{V}_0)(1-p)} = \frac{p\exp\left(-\frac{(x-\hat{V}_1)^T\Sigma^{-1}(x-\hat{V}_1)}{2}\right)}{p\exp\left(-\frac{(x-\hat{V}_1)^T\Sigma^{-1}(x-\hat{V}_1)}{2}\right) + (1-p)\exp\left(-\frac{(x-\hat{V}_0)^T\Sigma^{-1}(x-\hat{V}_0)}{2}\right)}$$

et le classifieur

$$\hat{g}(x) = 1_{\hat{\eta}(x)>1/2} = 1_{p\exp\left(-\frac{(x-\hat{V}_1)^T\Sigma^{-1}(x-\hat{V}_1)}{2}\right) > (1-p)\exp\left(-\frac{(x-\hat{V}_0)^T\Sigma^{-1}(x-\hat{V}_0)}{2}\right)}$$

6. par la loi des grands nombres, on a

$$\hat{V}_1 = \frac{1/n\sum\limits_{i=1}^n Y_iX_i}{1/n\sum\limits_{i=1}^n Y_i} \xrightarrow[n\to+\infty]{p.s.} \frac{E(YX)}{E(Y)} = \frac{E(YX|Y=1)P(Y=1) + E(YX|Y=0)P(Y=0)}{P(Y=1)} = E(X|Y=1)+0$$

De même $\hat{V}_0 \xrightarrow[n\to+\infty]{p.s.} E(X|Y=0)$. Etant une fonction continue, la fonction de régression converge aussi $\hat{\eta}(X,\mathcal{D}_n) \xrightarrow[n\to+\infty]{p.s.} \eta(X)$. De plus

$$0 \le |\hat{\eta}(X,\mathcal{D}_n) - \eta(X)| \le |\hat{\eta}(X,\mathcal{D}_n)| + |\eta(X)| \le 2$$

Par le théorème de convergence dominée, $E(|\hat{\eta}(X,\mathcal{D}_n) - \eta(X)|) \xrightarrow[n\to+\infty]{} 0$. En utilisant l'exercice 2, on a le résultat souhaité.

7. Considérons $X \sim \mathcal{N}(V_0, \Sigma)$ indépendamment de $Y \sim \mathcal{B}(p)$ avec $0 < p < 1/2$ par exemple $1/3$. Par indépendance on a $\eta(x) = E(Y) = p$ et $g^\star(x) = 1_{\eta(x)>1/2} = 0$ donc

$$R_P^* = P(g^\star(X) \neq Y) = P(Y \neq 0) = P(Y = 1) = p.$$

Pour le classifieur, on a par indépendance

$$R(\hat{g}) = P(\hat{g}(X) \neq Y) = P(\hat{g}(X) = 0, Y = 1) + P(\hat{g}(X) = 1, Y = 0) = pP(\hat{g}(X) = 0) + (1-p)P(\hat{g}(X) = 1).$$

Or

$$P(\hat{g}(X) = 0) = P\left(p \exp\left(-\frac{(X - \hat{V}_1)^T \Sigma^{-1} (X - \hat{V}_1)}{2}\right) \leq (1-p) \exp\left(-\frac{(X - \hat{V}_0)^T \Sigma^{-1} (X - \hat{V}_0)}{2}\right)\right)$$

La convergence presque sûre des estimateurs garantit que

$$P(\hat{g}(X) = 0) \xrightarrow[n \to +\infty]{p.s.} P\left(Z - \delta/2 \leq \delta^{-1} \log\left(\frac{1-p}{p}\right)\right) =: p_{\delta,p}$$

Pour $P(\hat{g}(X) = 1)$, on trouve

$$P(\hat{g}(X) = 1) = 1 - P(\hat{g}(X) = 0) = 1 - p_{\delta,p}.$$

car $\{\widehat{g}(X) = 0\}^c = \{\widehat{g}(X) = 1\}$. Donc

$$R(\hat{g}) \xrightarrow[n \to +\infty]{p.s.} p(1 - p_{\delta,p}) + (1-p)p_{\delta,p}.$$

S'il était universellement consistant alors les deux limites devraient être égales

$$p(1 - p_{\delta,p}) + (1-p)p_{\delta,p} = p$$

C'est impossible avec $p < 1/2$.

Exercice 8 Risque de classification pondéré

Soit $\mathcal{D}_n = \{(X_i, Y_i), i = 1, ..., n\}$ un échantillon de copies du couple de variables aléatoires (X, Y) avec $X \in \mathbb{R}^d$ et $Y \in \{0, 1\}$. Soient $\omega_0, \omega_1 > 0$ fixés et

$$\ell(y, y') = \omega_0 \mathbf{1}_{\{y=0; y'=1\}} + \omega_1 \mathbf{1}_{\{y=1; y'=0\}}.$$

,

0. Vérifier que le risque de classification pondéré R_ω d'un classifieur h s'écrit

$$R_\omega(h) = \omega_0 \mathbb{P}(Y = 0, h(X) = 1) + \omega_1 \mathbb{P}(Y = 1, h(X) = 0).$$

1. Soit $\eta(X) = \mathbb{E}[Y|X = x]$ la fonction de régression. Montrer que

$$R_\omega(h) \geq \mathbb{E}\left[\min\left\{\omega_0(1 - \eta(X)), \omega_1 \eta(X)\right\}\right].$$

2. Soit

$$h_\omega^*(x) = \mathbf{1}_{\left\{\eta(x) > \frac{\omega_0}{\omega_0 + \omega_1}\right\}},$$

calculer $R_\omega(h_\omega^*)$ et commenter.

3. Montrer que, pour tout classifieur h

$$R_\omega(h) - R_\omega(h_\omega^*) = \mathbb{E}\left[|\omega_1 \eta(X) - \omega_0(1 - \eta(X)| \mathbf{1}_{\{h(X) \neq h_\omega^*(X)\}}\right].$$

4. Soit $\widehat{\eta}$ un estimateur de la fonction η_P^* et

$$\widehat{h}_\omega(x) = \mathbf{1}_{\left\{\widehat{\eta}(x) > \frac{\omega_0}{\omega_0 + \omega_1}\right\}}.$$

(a) Montrer que $\widehat{h}_\omega(x) \neq h_\omega^*(x)$ implique $|\widehat{\eta}(x) - \eta(x)| \geq \left|\eta(x) - \frac{\omega_0}{\omega_0 + \omega_1}\right|$.

(b) En déduire que

$$\mathbb{E}[R_\omega(\widehat{h}_\omega)] - R_\omega(h^*_\omega) \leq (\omega_0+\omega_1)\mathbb{E}\left[|\widehat{\eta}(x)-\eta(x)|\right].$$

Answer of exercise 8

1.

$$\mathbb{E}[R_\omega(h)] = \omega_0\mathbb{P}(Y=0,h(X)=1)+\omega_1\mathbb{P}(Y=1,h(X)=0)$$

Comme $h(X)$ est $(X,\mathcal{D}_n)$-mesurable, et Y indépendant de $\mathcal{D}_n$, nous avons (en remarquant que $\mathbf{1}_{\{h(X)=1\}} = h(X)$),

$$\begin{aligned}\mathbb{P}(Y=0,h(X)=1) &= \mathbb{E}\left[\mathbb{P}(Y=0,h(X)=1|X,\mathcal{D}_n)\right] = \mathbb{E}\left[\mathbf{1}_{\{h(X)=1\}}\mathbb{P}(Y=0|X,\mathcal{D}_n)\right]\\ &= \mathbb{E}\left[h(X)\mathbb{P}(Y=0|X)\right] = \mathbb{E}\left[h(X)(1-\mathbb{P}(Y=1|X))\right]\\ &= \mathbb{E}\left[h(X)(1-\eta(X))\right].\end{aligned}$$

De même,

$$\mathbb{P}(Y=1,h(X)=0) = \mathbb{E}[(1-h(X))\eta(X)].$$

Nous avons donc

$$\begin{aligned}\mathbb{E}[R_\omega(h)] &= \mathbb{E}[\omega_0 h(X)(1-\eta(X))+\omega_1(1-h(X))\eta(X)]\\ &\geq \mathbb{E}[h(X)\min\{\omega_0(1-\eta(X)),\omega_1\eta(X)\}+(1-h(X))\min\{\omega_0(1-\eta(X)),\omega_1\eta(X)\}]\\ &= \mathbb{E}[\min\{\omega_0(1-\eta(X)),\omega_1\eta(X)\}].\end{aligned}$$

Attention ici au fait que R_ω n'est en général pas égal au risque de classification vu en cours (sauf dans le cas $\omega_0=\omega_1=1$). Aucun résultat du cours ne permet donc d'affirmer que le classifieur $h^*(x) = \mathbf{1}_{\{\eta(x)>1=/2\}}$ (classifieur de Bayes) minimise R_ω.

2. Comme h^*_ω ne dépend pas de $\mathcal{D}_n$,

$$R_\omega(h^*_\omega) = \omega_0\mathbb{P}(Y=0,h^*_\omega(X)=1)+\omega_1\mathbb{P}(Y=1,h^*_\omega(X)=0).$$

Nous avons,

$$\mathbb{P}(Y=0,h^*_\omega(X)=1) = \mathbb{E}[\mathbf{1}_{\{h^*_\omega(X)=1\}}\mathbb{P}(Y=0|X)] = \mathbb{E}\left[\mathbf{1}_{\left\{\eta(X)>\frac{\omega_0}{\omega_0+\omega_1}\right\}}(1-\eta(X))\right],$$

et de même,

$$\mathbb{P}(Y=1,h^*_\omega(X)=0) = \mathbb{E}\left[\mathbf{1}_{\left\{\eta(X)\leq\frac{\omega_0}{\omega_0+\omega_1}\right\}}\eta(X)\right].$$

Or $\eta(X) > \frac{\omega_0}{\omega_0+\omega_1} \Leftrightarrow \omega_1\eta(X) > \omega_0(1-\eta(X))$ d'où

$$\begin{aligned}R_\omega(h^*_\omega) &= \mathbb{E}\left[\omega_0(1-\eta(X))\mathbf{1}_{\{\omega_1\eta(X)>\omega_0(1-\eta(X))\}}+\omega_1\eta(X)\mathbf{1}_{\{\omega_1\eta(X)\leq\omega_0(1-\eta(X))\}}\right]\\ &= \mathbb{E}[\min\{\omega_0(1-\eta(X)),\omega_1\eta(X)\}].\end{aligned}$$

D'après 1., tout classifieur h vérifie $\mathbb{E}[R_\omega(h)] \geq R_\omega(h^*_\omega)$, h^*_ω est donc le classifieur minimisant le risque pondéré. C'est le classifieur de Bayes pour ce risque-là.

3. Soit h un classifieur, en conditionnant par X à l'intérieur de l'espérance et en remarquant que $\mathbf{1}_{\{h(X)=1\}} = h(X)$ et $\mathbf{1}_{\{h(X)=0\}} = 1-h(X)$ (de même en remplaçant h par h^*_ω), nous obtenons,

$$\begin{aligned}\mathbb{E}[R_\omega(h)]-R_\omega(h^*_\omega) &= \mathbb{E}\left[\omega_0\mathbf{1}_{\{Y=0\}}\left(\mathbf{1}_{\{h(X)=1\}}-\mathbf{1}_{\{h^*_\omega(X)=1\}}\right)+\omega_1\mathbf{1}_{\{Y=1\}}\left(\mathbf{1}_{\{h(X)=0\}}-\mathbf{1}_{\{h^*_\omega(X)=0\}}\right)\right]\\ &= \mathbb{E}\left[\omega_0(1-\eta(X))\left(h(X)-h^*_\omega(X)\right)+\omega_1\eta(X)\left((1-h(X))-(1-h^*_\omega(X))\right)\right]\\ &= \mathbb{E}\left[(\omega_0(1-\eta(X))-\omega_1\eta(X))\left(h(X)-h^*_\omega(X)\right)\right]\\ &= \mathbb{E}\left[(\omega_0(1-\eta(X))-\omega_1\eta(X))(-\mathbf{1}_{\{h(X)\neq h^*_\omega(X),h^*_\omega(X)=1\}})\right]\\ &\quad +\mathbb{E}\left[(\omega_0(1-\eta(X))-\omega_1\eta(X))\mathbf{1}_{\{h(X)\neq h^*_\omega(X),h^*_\omega(X)=0\}}\right].\end{aligned}$$

Or $h^*_\omega(X)=1 \Leftrightarrow \omega_0(1-\eta(X))-\omega_1\eta(X)) < 0$ d'où le résultat.

4. (a) Soit x tel que $h_\omega(x) \neq h^*_\omega(x)$, deux cas sont possibles.

Soit $h_\omega(x) = 1$ **et** $h^*_\omega(x) = 0$ dans ce cas-là, $\widehat{\eta}(x) > \frac{\omega_0}{\omega_0+\omega_1}$ et $\eta(x) \leq \frac{\omega_0}{\omega_0+\omega_1}$. Nous avons donc

$$|\widehat{\eta}(x) - \eta(x)| = \widehat{\eta}(x) - \eta(x) > \frac{\omega_0}{\omega_0+\omega_1} - \eta(x) = \left|\eta(x) - \frac{\omega_0}{\omega_0+\omega_1}\right|.$$

Soit $h_\omega(x) = 0$ **et** $h^*_\omega(x) = 1$ dans ce cas-là, $\widehat{\eta}(x) \leq \frac{\omega_0}{\omega_0+\omega_1}$ et $\eta(x) > \frac{\omega_0}{\omega_0+\omega_1}$. Nous avons donc

$$|\widehat{\eta}(x) - \eta(x)| = \eta(x) - \widehat{\eta}(x) \geq \eta(x) - \frac{\omega_0}{\omega_0+\omega_1} = \left|\eta(x) - \frac{\omega_0}{\omega_0+\omega_1}\right|.$$

(b) Par 3.

$$\begin{aligned}
\mathbb{E}[R_\omega(h_\omega)] - R_\omega(h^*_\omega) &= \mathbb{E}[|\omega_0(1-\eta(X)) - \omega_1\eta(X)|\mathbf{1}_{\{h_\omega(X)\neq h^*_\omega(X)\}}] \\
&= \mathbb{E}[|\omega_0 - (\omega_0+\omega_1)\eta(X)|\mathbf{1}_{\{h_\omega(X)\neq h^*_\omega(X)\}}] \\
&\leq (\omega_0+\omega_1)\mathbb{E}[|\widehat{\eta}(X) - \eta(X)|\mathbf{1}_{\{h_\omega(X)\neq h^*_\omega(X)\}}] \leq (\omega_0+\omega_1)\mathbb{E}[|\widehat{\eta}(X) - \eta(X)|],
\end{aligned}$$

où l'avant-dernière inégalité provient de (a).