TD – Analyse de données

Exercice 1 : Classification Bayésienne

On considère un problème de classification à deux classes ω_1 et ω_2 de densités (lois de Rayleigh) :

$$f(x|\omega_i) = \frac{x}{\sigma_i^2} exp(-\frac{x^2}{2\sigma_i^2}) \mathbb{I}_{\mathbb{R}^+}, \forall i = \{1, 2\} \quad (1)$$

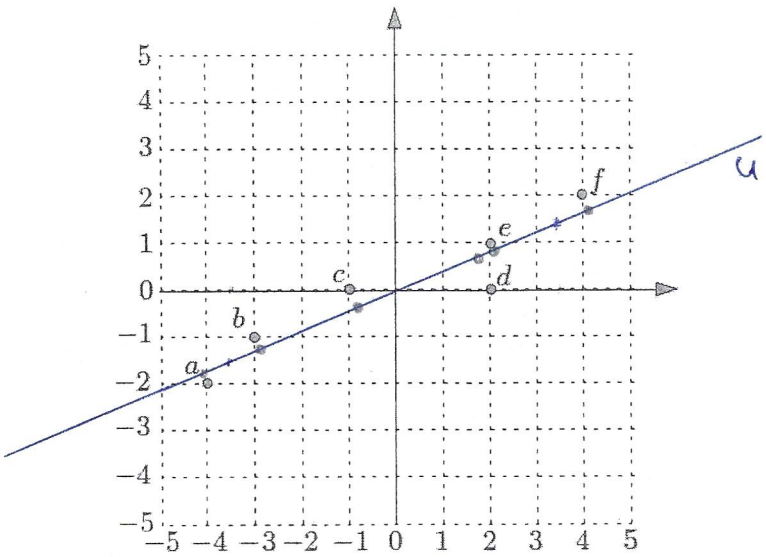
où $\mathbb{I}_{\mathbb{R}^+}$ est la fonction indicatrice sur $\mathbb{R}^+$ ($\mathbb{I}_{\mathbb{R}^+} = 1$ si $x > 0$ et 0 sinon) et $\sigma_1^2 > \sigma_2^2$.

Questions

1. Déterminer la règle de classification associée à ce problème avec la fonction de coût 0-1 et lorsque les deux classes sont équiprobables.
2. Déterminer la probabilité d'erreur associée à ce classifieur.

Exercice 2 : ACP et Classification

On considère le jeu de données suivant :



Questions

1. Caculer la matrice de variance-covariance Σ.
2. Calculer le premier vecteur principal et les composantes principales correspondantes.
3. Appliquer, sur les composantes principales 1D, l'algorithme des k-plus proches voisins pour $k = 1$ en supposant que le seuil est égal à $\frac{20}{7}$.

Exercice 3 : Cerveaux le retour !

On observe le nombre de cerveaux éveillés lors d'une séance de TP un vendredi matin à 8h sur un groupe de 10 étudiants : on remarque qu'au début de la séance, aucun cerveau n'est éveillé, au bout d'une heure, seulement 3 cerveaux sont éveillés et à 10h, à la pause, 7 cerveaux sont éveillés. On essaie de modéliser ces observations par une fonction f dont 3 points du graphique seraient connus ($f(0) = 1$, $f(1) = 3$ et $f(2) = 7$). On propose de chercher f dépendante du temps t exprimé en heure dans la famille des polynômes. Cependant, on considère que les mesures sont entachées d'erreurs ($f(0) \approx 1$, $f(1) \approx 3$ et $f(2) \approx 7$) et on cherche une fonction f plutôt de la forme :

$$f(t) = a\sqrt{|t-1|} + bt^2. \tag{2}$$

Questions

1. A priori, peut-on trouver une fonction de la forme $f(t) = a\sqrt{|t-1|} + bt^2$, qui passe exactement par les trois points expérimentaux ?
2. Ecrire le problème de minimisation qui détermine les coefficients a et b au sens des moindres carrés.
3. Résoudre le problème aux moindres carrés. En déduire l'erreur aux moindres carrés associée à cette approximation.

Exercice 4 : Arbre de décision

On cherche à construire un arbre de décision permettant de décider si un individu doit jouer au tennis ou non. Une base d'apprentissage a été construite comme suit.

	Ciel	Température	Vent	Jouer
x_1	soleil	chaud	faible	Oui
x_2	soleil	chaud	fort	Oui
x_3	couvert	chaud	faible	Non
x_4	pluie	froid	faible	Non
x_5	pluie	froid	faible	Non
x_6	pluie	froid	fort	Oui

Questions

1. Déterminer l'indice de Gini associé à cette base d'apprentissage vis-à-vis des deux classes "Jouer au Tennis" et "Ne pas jouer au Tennis".
2. Déterminer la variation de l'indice de Gini lorsqu'on coupe les données à l'aide des variables "Ciel", "Température" et "Vent". En déduire la variable qui sera utilisée au premier niveau de l'arbre de décision.
3. Expliquer comment on pourrait procéder si la variable "Température" était une valeur en degrés celsius.

TD 1
AD.

Analyse de Données

Exercice 1 : Classification Bayésienne

1/ $d^*(x) = \omega_1 \iff P(\omega_1|x) \geq P(\omega_2|x)$

$\iff f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2)$

Il y a équiprobabilité : $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ (à priori)

$\iff f(x|\omega_1) \geq f(x|\omega_2)$

$\iff \frac{x}{\sigma_1^2}\exp\left(-\frac{x^2}{2\sigma_1^2}\right) \geq \frac{x}{\sigma_2^2}\exp\left(-\frac{x^2}{2\sigma_2^2}\right)$

$\iff \sigma_2^2 \exp\left(-\frac{x^2}{2\sigma_1^2}\right) - \sigma_1^2 \exp\left(-\frac{x^2}{2\sigma_2^2}\right) \geq 0$

$\iff \ln(\sigma_2^2) - \ln(\sigma_1^2) - \frac{x^2(\sigma_2^2 - \sigma_1^2)}{2\sigma_1^2\sigma_2^2} \geq 0$

$\sqrt{\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \times \frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2 - \sigma_1^2}} \leq x$

$\iff x \geq \sqrt{\frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2 - \sigma_1^2}\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)}$

2/ Probabilité d'erreur (2 classes)

$P_e = P[d^*(x) = \omega_1 | x \in \omega_2]\, P(x \in \omega_2) + P[d^*(x) = \omega_2 | x \in \omega_1]\, P(x \in \omega_1)$

$= \frac{1}{2} P[x > a | x \in \omega_2] + \frac{1}{2} P[x < a | x \in \omega_1]$

$= \frac{1}{2}\int_a^{+\infty} \frac{x}{\sigma_2^2}\exp\left(-\frac{x^2}{2\sigma_2^2}\right)dx + \frac{1}{2}\int_0^a \frac{x}{\sigma_1^2}\exp\left(\frac{-x^2}{2\sigma_1^2}\right)dx$

$= \frac{1}{2}\left(\left[-\exp\left(\frac{-x^2}{2\sigma_2^2}\right)\right]_a^{+\infty} + \left[-\exp\left(\frac{-x^2}{2\sigma_1^2}\right)\right]_0^a\right)$

$= \frac{1}{2}\left[\exp\left(\frac{-a^2}{2\sigma_2^2}\right) + 1 - \exp\left(\frac{-a^2}{2\sigma_1^2}\right)\right]$

Exercice 2 : ACP et Classification

1/ Calcul de la matrice de variance/covariance

$$X = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} \quad X_{moy} = (0,0) \quad \text{Donc } X^c = X$$

$$\text{Donc } \Sigma = \frac{1}{n} X^{cT} X^c = \frac{1}{n} X^T X \quad \text{avec } n = 6$$

$$\frac{1}{6} \begin{bmatrix} -4 & -3 & -1 & 2 & 2 & 4 \\ -2 & -1 & 0 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 50 & 21 \\ 21 & 10 \end{bmatrix}$$

2/ On peut travailler avec la matrice 6Σ, en effet, elle aura les mêmes vecteurs propres, mais des valeurs propres seront multipliées par 6.

$$tr(6\Sigma) = 60 = \lambda_1 + \lambda_2$$
$$det(6\Sigma) = 59 = \lambda_1 \times \lambda_2$$

$$\lambda_1 + \lambda_2 = 60 \Rightarrow \lambda_1 = 60 - \lambda_2$$

$$59 = (60 - \lambda_2) \times \lambda_2 \Leftrightarrow -\lambda_2^2 + 60\lambda_2 - 59 = 0$$

$$\Rightarrow \lambda_2 = 1 \text{ ou } \lambda_2 = 59$$
$$\lambda_1 = 59 \qquad \lambda_2 = 1.$$

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \vec{v}$ tq $6\Sigma\vec{v} = \lambda_2 \vec{v}$

$$\begin{bmatrix} 50 & 21 \\ 21 & 10 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 59 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{aligned} 50x_1 + 21x_2 &= 59x_1 \\ 21x_1 + 10x_2 &= 59x_2 \end{aligned} \quad \Rightarrow x_2 = \frac{9}{21} x_1 = \frac{3}{7} x_1$$
$$x_2 = \frac{21}{49} x_1 = \frac{3}{7} x_1.$$

$$u = \begin{bmatrix} 7 \\ 3 \end{bmatrix}$$

$$C = Xu = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} 7 \\ 3 \end{bmatrix} \quad C = \begin{bmatrix} -34 \\ -24 \\ -7 \\ 14 \\ 17 \\ 34 \end{bmatrix}$$

TD 1.2
AD

3/ Matrice des distances $D =$

	a	b	c	d	e	f
a	0	10	27	48	51	68
b	10	0	17	38	41	58
c	27	17	0	21	24	41
d	48	38	21	0	3	20
e	51	41	24	3	0	17
f	68	58	41	20	17	0

$dist(a,b) = |a_A - a_B|$

0) {a} b ppv a dist(a,b) < 20

→ b et a ds classe / cluster.

1) {a,b} c ppv b dist(b,c) < 20

2) {a,b,c} d ppv c mais dist(c,d) > 20.

3) {a,b,c}, {d} ...

4)

5) {a,b,c}, {d,e,f}.

Exercice n°3 : Cerveaux le retour !

1/ On ne pourra pas passer exactement par les 3 points.
Car 3 points mais que 2 inconnues

2/
$$\begin{matrix} t & y \end{matrix}$$
$$\begin{bmatrix} 0 & 1 \\ 1 & 3 \\ 2 & 7 \end{bmatrix}$$

$f(t) = a\sqrt{|t-1|} + bt^2$

Fonction modèle : $f(x,\beta) = \sum_{k=1}^{m} \beta_k \phi_k(x)$

$\beta_1 = a \quad \beta_2 = b \quad \phi_1(x) = \sqrt{|x-1|} \quad \phi_2(x) = x^2$.

$$S(\beta) = \sum_{i=1}^{3} (y_i - f(t_i,\beta))^2$$
$$= (1-a)^2 + (3-b)^2 + (7-a-4b)^2$$
$$= 1 - 2a + a^2 + 9 - 6b + b^2 + 49 + a^2 + 16b^2 - 56b - 14a + 8ab$$
$$= 59 - 16a - 62b + 8ab + 2a^2 + 17b^2$$

$\underset{\beta}{\operatorname{argmin}}\ S(\beta)$ → forme matricielle $d = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}$ $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 4 \end{bmatrix}$

$\operatorname{argmin} \|d - A\beta\|^2$

Donc $\hat{\beta} = (A^TA)^{-1}A^Td$ $\quad (A^TA)^{-1} = \frac{1}{18}\begin{bmatrix} 17 & -4 \\ -4 & 2 \end{bmatrix}$.

$A^TA = \begin{bmatrix} 2 & 4 \\ 4 & 17 \end{bmatrix}$

Exercice 4 : Arbre de décision

2 classes $\begin{cases} 1 \text{ jouer au tennis} & n_1 = 3 \\ 2 \text{ ne pas jouer} & n_2 = 3 \end{cases}$

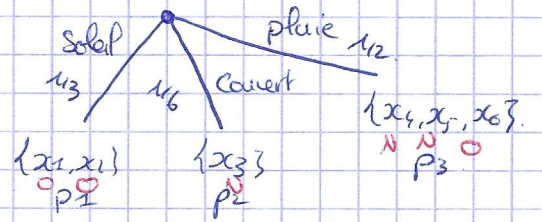
1/ L'indice de Gini de la base d'apprentissage

$$\sum_{i=1}^{2} \frac{n_i}{n}\left(1 - \frac{n_i}{n}\right) = \frac{1}{2} = \text{Gini}(p)$$

$$\text{Gain}(p, test) = \text{Gini}(p) - (P_{gauche}\,\text{Gini}(p_1) + P_{droite}\,\text{Gini}(p_2))$$

$p = \{x_1, \dots, x_6\}$ test = Ciel.

$$\text{Gain}(p, test) \quad \frac{1}{3} \times 0 + \frac{1}{6} \times 0 + \frac{1}{2}\left(2 \times \left(\frac{1}{3} \times \frac{2}{3}\right)\right)$$



$$= \frac{2}{9} \simeq 0,22$$

Pour température $G = \frac{4}{9} \simeq 0,44$

Vent $G = \frac{1}{4} = 0,25$.

C'est donc le ciel qui maximise le gain