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) :

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

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

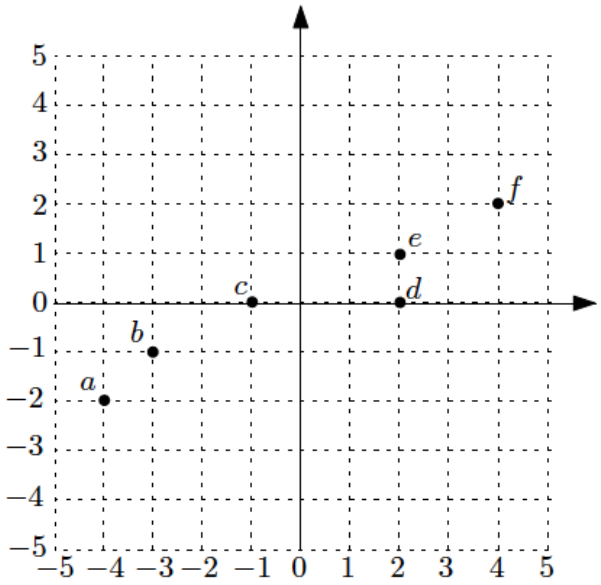
Questions

1. Déterminer la règle de classification associée à ce problème lorsque les deux classes sont équiprobables.
   Réponse : le classifieur Bayésien affecte x à la classe ω_1 (ce que l'on notera $d^*(x) = \omega_1$) si
   $$f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2)$$
   c'est-à-dire, en utilisant l'équiprobabilité des deux classes et le fait que $\sigma_1^2 > \sigma_2^2$
   $$d^*(x) = \omega_1 \Leftrightarrow x^2 \geq a^2 = \frac{2(\sigma_1^2\sigma_2^2)}{\sigma_2^2 - \sigma_1^2} \ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$$
   c'est-à-dire, en remarquant que $x > 0$
   $$d^*(x) = \omega_1 \Leftrightarrow x > a$$
   avec
   $$a = \sqrt{\frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2 - \sigma_1^2} \ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)}$$

2. Déterminer la probabilité d'erreur associée à ce classifieur.
   Réponse : la probabilité d'erreur d'un classifieur est définie par
   $$P_e = P[d^*(x) = \omega_1 | x \in \omega_2]P(x \in \omega_2) + P[d^*(x) = \omega_2 | x \in \omega_1]P(x \in \omega_1)$$
   ce qui donne dans notre cas
   $$P_e = \frac{1}{2}P[x > a | x \in \omega_2] + \frac{1}{2}P[x < a | x \in \omega_1]$$
   soit
   $$P_e = \frac{1}{2}\int_a^\infty \frac{x}{\sigma_2^2} \exp\left(-\frac{x^2}{2\sigma_2^2}\right) dx + \frac{1}{2}\int_0^a \frac{x}{\sigma_1^2} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) dx.$$
   Des calculs élémentaires conduisent à
   $$P_e = \frac{1}{2}\exp\left(-\frac{a^2}{2\sigma_2^2}\right) + \frac{1}{2}\left[1 - \exp\left(-\frac{a^2}{2\sigma_1^2}\right)\right]$$
   avec la valeur de a déterminée précédemment.

Exercice 2 : ACP et Classification

On considère le jeu de données suivant :



Questions

1. Calculer la matrice de variance-covariance Σ.
   Réponse

$$X = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} \quad g = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

$$\Sigma = \frac{1}{6} X_c^T * X_c = \frac{1}{6} X^T * X = \frac{1}{6} \begin{bmatrix} 50 & 21 \\ 21 & 10 \end{bmatrix}$$

2. Calculer le premier vecteur principal et la composante principale correspondante.
   Réponse

$$\chi_{6\Sigma}(\lambda) = det(6\Sigma - \lambda I_2) = \lambda^2 - 60\lambda + 59$$

$$\Delta = 3364 = 58^2 \text{ d'où } \lambda_1 = 1 \text{ et } \lambda_2 = 59$$

Le premier vecteur principal est le vecteur propre de 6Σ associé à la plus grande valeur propre $\lambda_2 = 59$

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

$$x_2 = \frac{3}{7} x_1$$

Calcul de la composante principale associée : $X \begin{bmatrix} 1 \\ 3/7 \end{bmatrix} = \begin{bmatrix} -34/7 \\ -24/7 \\ -1 \\ 2 \\ 17/7 \\ 34/7 \end{bmatrix}$

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 $\frac{20}{7}$.

   Réponse : Matrice des distances entre composantes principales 1D :

$$\begin{bmatrix} dist & a & b & c & d & e & f \\ a & 0 & 10 & 27 & 48 & 51 & 68 \\ b & & 0 & 17 & 38 & 41 & 58 \\ c & & & 0 & 21 & 24 & 41 \\ d & & & & 0 & 3 & 20 \\ e & & & & & 0 & 17 \\ f & & & & & & 0 \end{bmatrix}$$

   (toutes les composantes principales et le seuil ont été multipliés par 7 pour ne pas s'embêter avec les fractions et la distance utilisée est euclidienne)

   Considérons chaque point de l'ensemble de données :
   — a, le point le plus proche est b et $d(a, b) < 20$ donc il y a une classe $C_1 = \{a, b\}$;
   — b, le point le + proche est a ;
   — c, le point le + proche est b et $d(b, c) = 17 < 20$ donc la classe devient $C_1 = \{a, b, c\}$;
   — d, le point le + proche est e et $d(d, e) = 3 < 20$ donc une nouvelle classe $C_2 = \{d, e\}$;
   — e, le point le + proche est d ;
   — f, le point le + proche est e et $d(e, f) = 17 < 20$ donc $C_2 = \{d, e, f\}$.

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, un cerveau 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.

Réponses :

1. On cherche 2 inconnues a et b. Pour que la fonction f passe par les trois points expérimentaux, il faut que a et b suivent trois équations. Or il n'y a en général pas de solution pour un système de 3 équations à 2 inconnues. Il est donc vraisemblable qu'il n'existe pas une fonction de la forme $f(x) = a\sqrt{|x-1|} + bx^2$ passant exactement par les trois points expérimentaux.
2. La fonction modèle est de la forme :

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

Ici, $\beta_1 = a$, $\phi_1 = \sqrt{|x-1|}$ et $\beta_2 = b$, $\phi_2(x) = x^2$. On a donc affaire à un problème des moindres carrés linéaires. On cherche ainsi a et b tels que la quantité suivante $S(a, b)$ soit minimale :

$$S(a,b) = \sum_{i=1}^{3} \left(y_i - a\sqrt{|x_i - 1|} - bx_i^2\right)^2 = (1-a)^2 + (3-b)^2 + (7-a-4b)^2$$

3. Ecriture matricielle sous la forme $A\beta = b$ Par identification, $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 4 \end{bmatrix}$ et $b = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}$ et $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$.

Donc le problème revient à minimiser : $\min_{\beta} \|A\beta - b\|^2$.

La solution est $\widehat{\beta} = (A^T A)^{-1} A^T b$ avec $A^T A = \begin{bmatrix} 2 & 4 \\ 4 & 17 \end{bmatrix}$ (inversible car $det(A^T A) = 18$)

d'où $(A^T A)^{-1} = \frac{1}{18}\begin{bmatrix} 17 & -4 \\ -4 & 2 \end{bmatrix}$. et donc $\widehat{\beta} = \begin{bmatrix} \frac{2}{3} \\ \frac{5}{3} \end{bmatrix}$

On peut d'ailleurs vérifier que la fonction ne passe pas exactement par les points donnés et calculer l'erreur aux moindres carrés :

$$S(a,b) = (1-a)^2 + (3-b)^2 + (7-a-4b)^2 = 2$$

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
$\boldsymbol{x}_1$	soleil	chaud	faible	Oui
$\boldsymbol{x}_2$	soleil	chaud	fort	Oui
$\boldsymbol{x}_3$	couvert	chaud	faible	Non
$\boldsymbol{x}_4$	pluie	froid	faible	Non
$\boldsymbol{x}_5$	pluie	froid	faible	Non
$\boldsymbol{x}_6$	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".
   Réponse : L'indice de Gini de la base d'apprentissage s'écrit
   $$\sum_{i=1}^{2} \frac{n_i}{n}\left(1-\frac{n_i}{n}\right) = \left(\frac{1}{2}\times\frac{1}{2}\right)+\left(\frac{1}{2}\times\frac{1}{2}\right) = \frac{1}{2}.$$

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.

   Soit p une position et soit *test* un test. Si ce test devient l'étiquette du noeud à la position p alors on appelle P_{gauche} (resp. P_{droite})) la proportion d'éléments de l'ensemble des exemples associés à p qui vont sur le noeud en position p_1 (resp. p_2). La réduction d'impureté définie par *test* est définie par :
   $$Gain(p, test) = Gini(p) - [P_{gauche}Gini(p_1) + P_{droite}Gini(p_2)]$$
   On choisit le test qui maximise la quantité Gain(p,test).
   Réponse : La variable "Ciel" coupe la base d'apprentissage en trois sous ensembles associés aux valeurs "Soleil", "Couvert" et "Pluie" qui correspondent aux valeurs de "Jouer" égales à $\{O, O\}$, $\{N\}$ et $\{N, N, O\}$. L'indice de Gini associé à ces trois sous ensembles est
   $$\left(\frac{2}{6}\times 0\right)+\left(\frac{1}{6}\times 0\right)+\left(\frac{3}{6}\times 2\times\frac{2}{3}\times\frac{1}{3}\right) = \frac{2}{9}\approx 0.22.$$
   Pour la variable "Température", on obtient deux sous ensembles $\{O, O, N\}$ et $\{N, N, O\}$, d'où l'indice de Gini
   $$\left(\frac{1}{2}\times 2\times\frac{2}{3}\times\frac{1}{3}\right)\times 2 = \frac{4}{9}\approx 0.44.$$
   Enfin pour la variable "Vent", on obtient $\{O, N, N, N\}$ et $\{O, O\}$ avec un indice de Gini égal à $\frac{2}{3}\times 2\times\frac{1}{4}\times\frac{3}{4} = \frac{1}{4} = 0.25$.
   On en conclut que la variable "Ciel" permet d'obtenir la réduction la plus importante de l'indice de Gini. Ce sera donc cette variable qui sera au premier niveau de l'arbre de décision.

3. Expliquer comment on pourrait procéder si la variable "Température" était une valeur en degrés Celsius.

   Réponse : Dans ce cas, on sépare les valeurs de températures vérifiant $x_i > S$ et $x_i < S$ pour toutes les valeurs possibles des seuils S et on garde à chaque fois la valeur du seuil qui permet d'obtenir la diminution la plus importante de l'indice de Gini.