Partiel Analyse de Données

Documents autorisés :
planches de cours, sujets de TD/TP, notes MANUSCRITES PERSONNELLES de cours/TD (PAS de PHOTOCOPIES), pas de calculatrice.
Durée :
1h30 (+30 min tiers temps)

Questions de cours (5 points)

1. (1pt) On considère trois classes équiprobables notées ω_1, ω_2 et ω_3 de densités $f(x|\omega_1) = \mathcal{N}(0,1)$, $f(x|\omega_2) = \mathcal{N}(4,1)$ et $f(x|\omega_3) = \mathcal{N}(6,1)$. Quel est le classifieur Bayésien pour ce problème ?
2. (1pt) On considère un problème de classification supervisée à deux classes ω_1 et ω_2 avec un ensemble d'apprentissage constitué de vecteurs $\boldsymbol{x}_i$ avec leurs étiquettes $y_i = \pm 1$. On applique le classifieur SVM (machines à vecteurs supports) et on obtient deux vecteurs supports notés $\boldsymbol{x}^+ \in \omega_1$ et $\boldsymbol{x}^- \in \omega_2$. Si on note $\boldsymbol{w}^T\boldsymbol{x} - b$ la frontière séparatrice de ce classifieur, quelles sont les valeurs de $\boldsymbol{w}^T\boldsymbol{x}^+ - b$ et $\boldsymbol{w}^T\boldsymbol{x}^- - b$?
3. (1pt) On considère un réseau de neurones à une couche et une sortie dont le vecteur d'entrée à l'itération n est noté $\boldsymbol{x}(n) = (x_1(n), ..., x_p(n))^T$ avec $x_p(n) = -1$ et de sortie

$$y(n) = f\left[\sum_{i=1}^{p} w_i x_i(n)\right]$$

   avec $f(u) = \frac{1}{1+\exp(-\alpha u)}, u \in \mathbb{R}, \alpha > 0$. Comment obtenir la règle de mise à jour des poids $w_i(n+1)$ en fonction de $w_i(n)$?
4. (1pt) On désire construire un arbre de décision à partir d'un ensemble de données à valeurs dans $\{\text{bleu}, \text{blanc}, \text{rouge}\}$. Expliquer comment construire les deux premières branches de cet arbre de décision.
5. (1pt) On considère l'ensemble $\mathcal{X} = \{1, 3, 6, 10, 12, 15\}$. Que donne la première étape de l'algorithme k-means pour séparer cet ensemble en trois classes ω_1, ω_2 et ω_3 lorsque les points initiaux de cet algorithme sont $g_1 = 0$, $g_2 = 5$ et $g_3 = 11$ (on précisera les 3 classes obtenues et les nouvelles valeurs de g_1, g_2 et g_3 après cette classification) ?

Analyse en composantes principales (5 points)

On considère le tableau de données suivant noté $\boldsymbol{X}$ constitué de 5 individus $\boldsymbol{x}_i$, $i = 1, ..., 5$ avec 2 variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ qui représentent leurs consommations d'eau et de gaz.

	$\boldsymbol{v}_1$	$\boldsymbol{v}_2$
$\boldsymbol{x}_1$	2	5
$\boldsymbol{x}_2$	4	2
$\boldsymbol{x}_3$	6	3
$\boldsymbol{x}_4$	3	4
$\boldsymbol{x}_5$	5	1

1. (2pts) Déterminer le tableau centré réduit $\boldsymbol{Y}$ associé à $\boldsymbol{X}$ et montrer que la matrice de covariance de $\boldsymbol{Y}$ est $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -\frac{7}{10} \\ -\frac{7}{10} & 1 \end{pmatrix}$.
2. (1pts) Déterminer les valeurs propres et les vecteurs propres de $\boldsymbol{M} = 10\boldsymbol{\Sigma}$. En déduire que les valeurs propres et les vecteurs propres de $\boldsymbol{\Sigma}$ sont $\mu_1 = \frac{17}{10}$ et $\mu_2 = \frac{3}{10}$ et $\boldsymbol{u}_1 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \boldsymbol{u}_2 = \begin{pmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$ (en s'assurant que la première composante de ces vecteurs est négative).
3. (1pt) Représenter l'ACP des individus $\boldsymbol{x}_i$, $i = 1, ..., 5$.
4. (1pt) Déduire de la question précédente l'ACP des deux variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$.

Moindres carrés (5 points)

Le tableau ci-dessous représente l'évolution du prix d'une baguette de pain (rapporté en euros actuels) depuis environ un siècle.

Année	1930	1960	1980	2000	2022
Prix (en euros)	0.001	0.05	0.25	0.64	0.93

On cherche à estimer les paramètres d'un modèle pour prédire le prix du pain en une année donnée.

1. On choisit tout d'abord un modèle linéaire selon l'année t, d'équation $f(t) = at + b$. En posant $\beta = [a, b]^T$, écrivez matriciellement le problème d'estimation aux moindres carrés à résoudre à partir des données de l'énoncé, c'est-à-dire définissez les matrices $\mathbf{A}$ et $\mathbf{B}$ telles que le problème aux moindres carrés puisse s'écrire :

$$\min_{\beta \in \mathbb{R}^2} \frac{1}{2} ||\mathbf{A}\beta - \mathbf{B}||^2$$

2. On considère maintenant un modèle quadratique, d'équation $f(t) = at^2 + bt + c$. En posant cette-fois ci $\beta = [a, b, c]^T$, explicitez à nouveau les matrices $\mathbf{A}$ et $\mathbf{B}$ pour formuler le problème aux moindres carrés associé.

3. Dans le cas général d'un problème aux moindres carrés où la matrice $\mathbf{A} \in \mathbb{R}^{m \times n}$, avec $m > n$, donnez la solution théorique du problème sans la calculer explicitement.

   Le tableau ci-dessous résume les prédictions des modèles linéaire et quadratique sur le jeu de données de départ, ainsi que sur une nouvelle donnée (année 1970).

Année	1930	1960	1980	2000	2022	1970
Prix (en euros)	0.001	0.05	0.25	0.64	0.93	0.09
Prédictions du modèle linéaire	-0.14	0.18	0.39	0.60	0.84	0.28
Prédictions du modèle quadratique	-0.02	0.09	0.28	0.56	0.97	0.17

4. En vous appuyant sur les résultats présentés dans ce tableau, de quel modèle diriez-vous qu'il explique le mieux les données ?

Classification Bayésienne (5 points)

On considère un problème de classification à deux classes ω_1 et ω_2 de densités

$$f(x|\omega_i) = \frac{\frac{1}{\pi b}}{1 + \left(\frac{x - a_i}{b}\right)^2}, \quad x \in \mathbb{R}, i \in \{1, 2\} \tag{1}$$

avec $b > 0$ et $a_2 > a_1 = 0$.

1. (1pt) Expliciter le classifieur Bayésien noté d^* lorsque les deux classes sont équiprobables.
2. (2pts) Déterminer la probabilité $P = P[d^*(x) = \omega_2 | x \in \omega_1]$. Que représente cette probabilité ?
3. (2pts) Reprendre la première question lorsque $P(\omega_1) = 2P(\omega_2)$ et montrer que la frontière de séparation entre les deux classes est une équation du second degré en x qu'on ne cherchera pas à résoudre.