Corrections Partiel Analyse de Données

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 ?
   Comme les variances des trois densités sont les mêmes, et que les classes sont équiprobables, on sait que le classifieur Bayésien donne la règle de la distance aux barycentres qui est définie par

$$\begin{aligned} d^*(x) = \omega_1 &\Leftrightarrow d(x,0) \leq d(x,4) \text{ et } d(x,0) \leq d(x,6) \Leftrightarrow x \leq 2 \\ d^*(x) = \omega_2 &\Leftrightarrow d(x,4) \leq d(x,0) \text{ et } d(x,4) \leq d(x,6) \Leftrightarrow x \in]2,5[\\ d^*(x) = \omega_3 &\Leftrightarrow d(x,6) \leq d(x,0) \text{ et } d(x,6) \leq d(x,4) \Leftrightarrow x \geq 5. \end{aligned}$$

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 $\boldsymbol{w}^T\boldsymbol{x}^- - b$?
   On sait que les vecteurs supports vérifient $y_i(\boldsymbol{w}^T\boldsymbol{x_i} - b) = 1$ donc $\boldsymbol{w}^T\boldsymbol{x}^+ - b = 1$ et $\boldsymbol{w}^T\boldsymbol{x}^- - b = -1$ ou $\boldsymbol{w}^T\boldsymbol{x}^+ - b = -1$ et $\boldsymbol{w}^T\boldsymbol{x}^- - b = 1$.
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 $w_i(n)$?
   La mise à jour des poids $w_i(n+1)$ en fonction de $w_i(n)$ découle de la règle du gradient :

$$w_i(n+1) = w_i(n) - \mu \left.\frac{\partial e^2(n)}{\partial w_i}\right|_{w_i = w_i(n)}$$

   En utilisant

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

   on obtient

$$\frac{\partial e^2(n)}{\partial w_i} = -2e(n)\frac{\partial y(n)}{\partial w_i} = -2\alpha e(n) x_i(n) y(n)[1 - y(n)]$$

   d'où

$$w_i(n+1) = w_i(n) - \mu' e(n) x_i(n) y(n)[1 - y(n)], \quad i = 1, ..., p.$$

4. (1pt) On désire construire un arbre de décision à partir d'un ensemble de vecteurs dont l'une des composantes appartient à l'ensemble $\{\text{bleu}, \text{blanc}, \text{rouge}\}$. Expliquer comment construire deux branches de cet arbre de décision lorsqu'on choisit cette composante.
Il faut choisir la couleur qui minimise l'indice de Gini ou l'entropie. Tous les vecteurs associés à cette couleur seront dirigés vers une branche de l'arbre tandis que les autres seront dirigés vers l'autre branche de l'arbre.

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) ?
À la première itération, l'ensemble ω_1 contient les points les plus proches de g_1, i.e., $\omega_1 = \{1\}$, tandis que ω_1 contient les points les plus proches de g_2, i.e., $\omega_2 = \{3, 6\}$, et enfin ω_3 contient les points les plus proches de g_3, i.e., $\omega_3 = \{10, 12, 15\}$. Les nouvelles valeurs de g_1, g_2 et g_3 après cette classification sont $g_1 = 1$, $g_2 = \frac{9}{2}$ et $g_3 = \frac{37}{3}$.

Analyse en composantes principales (5 points)

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

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

1. (1pt) Déterminer le tableau centré réduit $\boldsymbol{Y}$ associé à $\boldsymbol{X}$ et montrer que la matrice de covariance de $\boldsymbol{Y}$ est $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -\frac{7}{10} \\ -\frac{7}{10} & 1 \end{pmatrix}$.
Les moyennes et variances des variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ sont $\bar{\boldsymbol{v}}_1 = \frac{20}{5} = 4$, $\bar{\boldsymbol{v}}_2 = \frac{15}{5} = 3$, $\sigma_1^2 = \frac{10}{5} = 2$ et $\sigma_2^2 = \frac{10}{5} = 2$. Le tableau centré est donc

$$\boldsymbol{X}_c = \begin{array}{c|cc} & \boldsymbol{v}_1 & \boldsymbol{v}_2 \\ \bar{x}_1 & -2 & 2 \\ \bar{x}_2 & 0 & -1 \\ \bar{x}_3 & 2 & 0 \\ \bar{x}_4 & -1 & 1 \\ \bar{x}_5 & 1 & -2 \end{array}$$

et le tableau centré réduit est $\boldsymbol{Y} = \frac{1}{\sqrt{2}}\boldsymbol{X}_c$. Des calculs élémentaires permettent d'obtenir la matrice de covariance de $\boldsymbol{Y}$

$$\boldsymbol{\Sigma} = \frac{1}{5}\boldsymbol{Y}^T\boldsymbol{Y} = \begin{pmatrix} 1 & -\frac{7}{10} \\ -\frac{7}{10} & 1 \end{pmatrix}.$$

2. (2pts) 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).
On a

$$\boldsymbol{\Sigma} = \frac{1}{10}\begin{pmatrix} 10 & -7 \\ -7 & 10 \end{pmatrix} = \frac{1}{10}\boldsymbol{M}.$$

Les valeurs propres de $\boldsymbol{M}$ s'obtiennent en cherchant les valeurs de λ solutions de

$$\begin{vmatrix} 10-\lambda & -7 \\ -7 & 10-\lambda \end{vmatrix} = 0 \Leftrightarrow \lambda^2 - 20\lambda + 100 - 49 = 0 \Leftrightarrow (\lambda - 3)(\lambda - 17) = 0.$$

Les deux valeurs propres de $\boldsymbol{M}$ sont donc $\lambda_1 = 17$ et $\lambda_2 = 3$. Un vecteur propre associé à λ_1 vérifie

$$\begin{pmatrix} 10-\lambda_1 & -7 \\ -7 & 10-\lambda_1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Si on choisit un vecteur normé avec une première composante négative, on obtient

$$\boldsymbol{u}_1 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

De la même manière, pour λ_2, on obtient

$$\boldsymbol{u}_2 = \begin{pmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

On en déduit que la matrice $\boldsymbol{M}$ admet les valeurs propres $\mu_1 = \frac{17}{10}$ et $\mu_2 = \frac{3}{10}$ avec les mêmes vecteurs propres que ceux de $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{u}_1$ et $\boldsymbol{u}_2$.

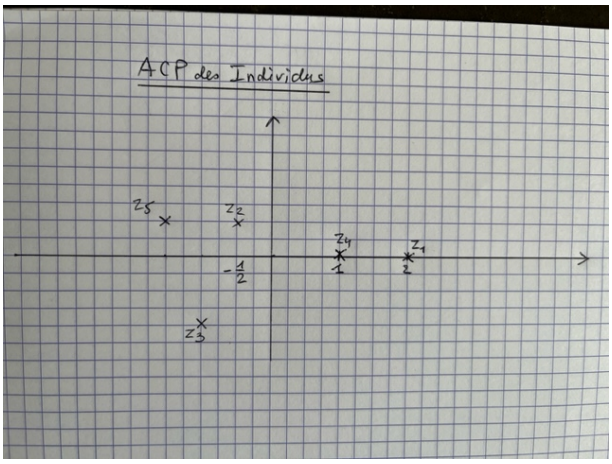
3. (1pt) Représenter l'ACP des individus $\boldsymbol{x}_i$, $i = 1, ..., 5$.
Les projections des individus sur le premier axe principal sont données par

$$\boldsymbol{Y}\boldsymbol{u}_1 = \begin{bmatrix} 2 \\ -1/2 \\ -1 \\ 1 \\ -3/2 \end{bmatrix}.$$

Celles sur le second axe sont

$$\boldsymbol{Y}\boldsymbol{u}_2 = \begin{bmatrix} 0 \\ 1/2 \\ -1 \\ 0 \\ 1/2 \end{bmatrix}.$$

d'où la représentation graphique suivante



4. (1pt) Déduire de la question précédente l'ACP des deux variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$.
   D'après le cours, les coordonnées des deux variables $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ dans le plan des deux premières composantes principales $\boldsymbol{a}_1$ et $\boldsymbol{a}_2$ vérifient

$$\begin{bmatrix} r(\boldsymbol{v}_1, \boldsymbol{a}_1) \\ r(\boldsymbol{v}_1, \boldsymbol{a}_2) \end{bmatrix} = \sqrt{\mu_1}\boldsymbol{u}_1 = \sqrt{\frac{17}{10}}\frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

   et

$$\begin{bmatrix} r(\boldsymbol{v}_2, \boldsymbol{a}_1) \\ r(\boldsymbol{v}_2, \boldsymbol{a}_2) \end{bmatrix} = \sqrt{\mu_2}\boldsymbol{u}_2 = \sqrt{\frac{3}{10}}\frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

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. (1pt) On choisit tout d'abord un modèle linéaire selon l'année t, d'équation $f(t) = at + b$. En posant $\boldsymbol{\beta} = [a, b]^T$, 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 :

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

D'après les données du tableau précédent, on a

$$\boldsymbol{A} = \begin{bmatrix} 1930 & 1 \\ 1950 & 1 \\ 1980 & 1 \\ 2000 & 1 \\ 2022 & 1 \end{bmatrix} \text{ et } \boldsymbol{B} = \begin{bmatrix} 0.001 \\ 0.05 \\ 0.25 \\ 0.04 \\ 0.93 \end{bmatrix}$$

2. (1pt) On considère maintenant un modèle quadratique, d'équation $f(t) = at^2 + bt + c$. En posant cette-fois ci $\boldsymbol{\beta} = [a, b, c]^T$, explicitez à nouveau les matrices $\mathbf{A}$ et $\mathbf{B}$ pour formuler le problème aux moindres carrés associé.
En utilisant le modèle quadratique, on obtoent

$$\boldsymbol{A} = \begin{bmatrix} (1930)^2 & 1930 & 1 \\ (1950)^2 & 1950 & 1 \\ (1980)^2 & 1980 & 1 \\ (2000)^2 & 2000 & 1 \\ (2022)^2 & 2022 & 1 \end{bmatrix} \text{ et } \boldsymbol{B} = \begin{bmatrix} 0.001 \\ 0.05 \\ 0.25 \\ 0.04 \\ 0.93 \end{bmatrix}$$

3. (1pt) Dans le cas général d'un problème aux moindres carrés où la matrice $\mathbf{A} \in \mathbb{R}^{m \times n}$ (avec $m > n$) est de rang n, donnez la solution théorique du problème sans la calculer explicitement.
Nous avons vu en cours que la solution des moindres carrés est

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{B}.$$

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. (2pts) En vous appuyant sur les résultats présentés dans ce tableau, proposez au moins deux méthodes (ou arguments) permettant de justifier quel modèle explique le mieux les données.
Pour déterminer quel modèle explique le mieux les données, il y a plusieurs manières de procéder :
   - On peut tout d'abord calculer l'erreur au sens des moindres carré entre le vecteur de prédiction $\boldsymbol{A}\widehat{\boldsymbol{\beta}}$ et le vecteur $\boldsymbol{B}$ définie par $\|\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{B}\|_2^2 = \sum_{i=1}^{6} [(\boldsymbol{A}\widehat{\boldsymbol{\beta}})_i - \boldsymbol{B}_i]^2$ pour les deux modèles et retenir le modèle qui donne l'erreur la plus petite.
   - On peut aussi calculer l'erreur ℓ_1 entre le vecteur de prédiction $\boldsymbol{A}\widehat{\boldsymbol{\beta}}$ et le vecteur $\boldsymbol{B}$ définie par$\|\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{B}\|_1 = \sum_{i=1}^{6} |(\boldsymbol{A}\widehat{\boldsymbol{\beta}})_i - \boldsymbol{B}_i|$ pour les deux modèles et retenir le modèle qui donne l'erreur la plus petite.
   - Enfin une méthode plus graphique consiste à tracer les courbes des prédictions $(\boldsymbol{A}\widehat{\boldsymbol{\beta}})_i$ en fonction des prix $\boldsymbol{B}_i$ est à retenir le modèle donnant la courbe la plus proche de la droite $y = x$.

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 $b > 0$, $a_2 = a > 0$ et $a_1 = 0$.

1. (1pt) Expliciter le classifieur Bayésien noté d^* lorsque les deux classes sont équiprobables.
   Puisque les deux classes sont équiprobables, le classifieur Bayésien affecte x à la classe ω_1, ce qu'on notera $d^*(x) = \omega_1$ si
   $$f(x|\omega_1) \geq f(x|\omega_2)$$
   c'est-à-dire
   $$1 + \left(\frac{x}{b}\right)^2 \leq 1 + \left(\frac{x-a}{b}\right)^2 \Leftrightarrow x \leq \frac{a}{2}.$$

2. (2pts) Déterminer la probabilité $P = P[d^*(x) = \omega_2 | x \in \omega_1]$ et montrer que cette quantité est la probabilité d'erreur du classifieur.
   On a
   $$P[d^*(x) = \omega_2|x \in \omega_1] = P\left[x > \frac{a}{2} \middle| f(x|\omega_1) = \frac{\frac{1}{\pi b}}{1 + \left(\frac{x}{b}\right)^2}\right] = \int_{\frac{a}{2}}^{\infty} \frac{\frac{1}{\pi b}}{1 + \left(\frac{x}{b}\right)^2} dx.$$
   Si on fait le changement de variables $u = \frac{x}{b}$, on obtient
   $$P[d^*(x) = \omega_2|x \in \omega_1] = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{a}{2b}\right).$$
   La probabilité d'erreur du classifieur Bayésien est
   $$P_e = P[d^*(x) = \omega_2|x \in \omega_1]P(\omega_1) + P[d^*(x) = \omega_1|x \in \omega_2]P(\omega_2).$$
   Les deux classes étant équiprobables, on a $P(\omega_1) = P(\omega_1) = \frac{1}{2}$. On montre que $P[d^*(x) = \omega_2|x \in \omega_1] = P[d^*(x) = \omega_1|x \in \omega_2]$. Donc P est la probabilité d'erreur du classifieur Bayésien.

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.
   Le classifieur Bayésien affecte x à la classe ω_1, ce qu'on notera $d^*(x) = \omega_1$ si
   $$f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2).$$
   Comme $P(\omega_1) = 2P(\omega_2)$ et $P(\omega_1) + P(\omega_2) = 1$, on a $P(\omega_1) = \frac{2}{3}$ et $P(\omega_2) = \frac{1}{3}$ c'est-à-dire
   $$\left[1 + \left(\frac{x}{b}\right)^2\right]\frac{1}{3} \leq \left[1 + \left(\frac{x-a}{b}\right)^2\right]\frac{2}{3}.$$
   La frontière de séparation entre les deux classes est donc définie par
   $$2(x-a)^2 - x^2 + b^2 = 0$$
   qui est bien une équation du second degré en x.