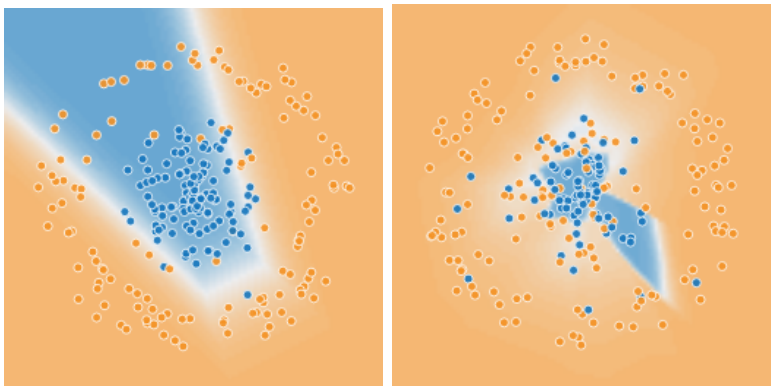
Partiel Analyse de Données

Documents autorisés :
1 feuille A4 Recto/Verso
Durée :
1h30 (+30 min tiers temps)

Questions de cours

1. (1pt) Quelles quantités doit-on connaître pour chaque classe pour mettre en oeuvre le classifieur Bayésien ?
   On doit connaître les probabilités a priori de chaque classe $P(\omega_i)$ et les densités de probabilité conditionnelles à chaque classe $p(\boldsymbol{x}|\omega_i)$.
2. (1pt) Représenter graphiquement un exemple de sous apprentissage et un exemple de sur-apprentissage pour un problème de classification à deux classes.
   Voici deux exemples vus en cours de sous apprentissage (à gauche) et de sur apprentissage (à droite) :



3. (1pt) Lorsque les classes sont linéairement séparables, un des classifieurs SVM vu en cours est défini par le problème suivant :

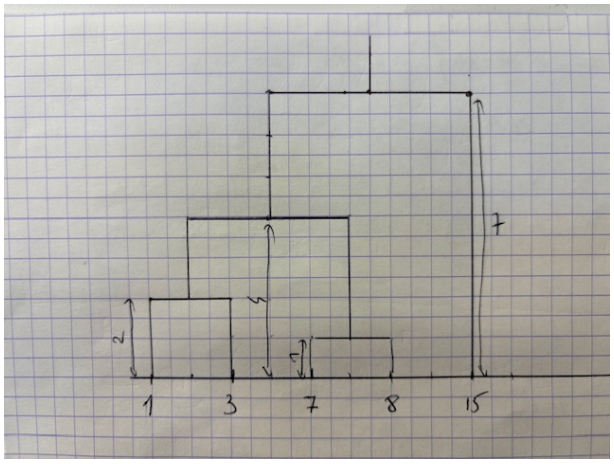
$$\begin{cases} \min\limits_{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}} \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 \right\} \\ \text{s.c.} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq 1, \forall i \in \{1,\dots,n\} \end{cases}$$

   Rappeler l'utilité de la contrainte $y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq 1, \forall i \in \{1,\dots,n\}$.
   Lorsque les classes sont linéairement séparables, on impose la contrainte $y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) = 1$ pour

les vecteurs supports. La contrainte $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right) \geq 1, \forall i \in \{1, \ldots, n\}$ indique que le vecteur $\boldsymbol{x}_i$ est situé du bon coté de l'hyperplan séparateur.

4. (1pt) On considère l'algorithme CART pour classifier des points de $\mathbb{R}^2$ notés $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ de densité continue. Expliquer comment est définie chaque branche de l'arbre.
   À chaque itération de l'algorithme, on choisit la composante x_{i1} ou x_{i2} qui minimise $P_L i_L + P_R i_R$, où i_L est l'indice de Gini de la branche de gauche, i_R est l'indice de Gini de la branche de droite et $P_L = \frac{n_L}{n}$, $P_R = \frac{n_R}{n}$ sont les proportions d'éléments contenus dans les deux branches de l'arbre. Si on choisit la première composante, la branche de l'arbre de droite correspond à $x_{i1} >$ seuil et celle de gauche correspond à $x_{i1} <$ seuil, où "seuil" est la moyenne arithmétique des deux éléments ordonnés maximisant $P_L i_L + P_R i_R$.
5. (1pt) On désire effectuer un clustering de l'ensemble $\mathcal{X} = \{1, 3, 7, 8, 15\}$ à l'aide de la méthode de classification hiérarchique ascendante. Que est le dendogramme obtenu pour la distance

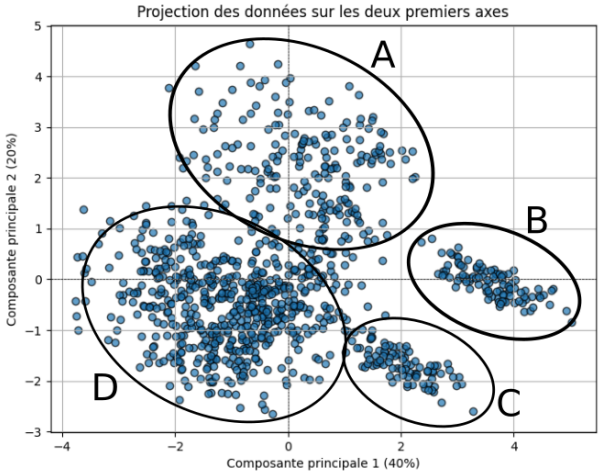
$$d(X_i, X_j) = \min\{d(x, y), x \in X_i, y \in X_j\}?$$

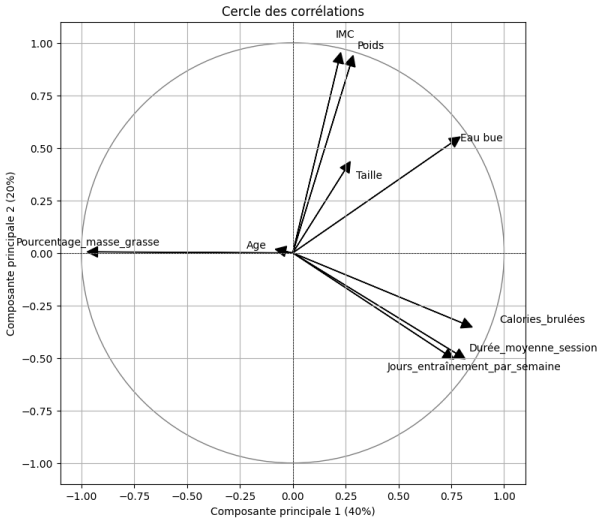


Exercice : ACP

Pour cet exercice, on dispose d'un tableau de données regroupant diverses informations physiologiques et sur la pratique sportive d'un ensemble d'individus. Pour chaque personne, on connaît les variables suivantes : 1) âge, 2) taille (t en m), 3) poids (p en kg), 4) indice de masse corporelle (IMC, calculé par la formule $\frac{p}{t^2}$), 5) pourcentage de masse grasse, 6) quantité d'eau bue en moyenne pendant une séance de sport (en L), 7) nombre de calories brûlées par séance, 8) durée moyenne d'une séance de sport, et 9) nombre de jours d'entraînement par semaine.

On se propose dans cet exercice de faire une Analyse en Composantes Principales de ce tableau de données. Voici une représentation des données projetées sur les 2 premiers axes ainsi que le cercle des corrélations :





6. (1pt) Si on note λ_1 et λ_2 les 2 plus grandes valeurs propres de la matrice de variance-covariance du tableau de données étudié, que peut-on dire ici du rapport $\frac{\lambda_1}{\lambda_2}$?
   L'inertie de l'axe 1 $\frac{\lambda_1}{\sum_j \lambda_j}$ est de 0.4, et l'inertie de l'axe 2 $\frac{\lambda_2}{\sum_j \lambda_j}$ est de 0.2. On en déduit que le rapport $\frac{\lambda_1}{\lambda_2}$, égal au rapport des inerties, est égal à 2.
7. (1pt) Que signifie la très faible longueur de la flèche représentant la variable *Age* sur le cercle des corrélations ?
   Cette représentation de la variable Age signifie qu'elle n'est corrélée ni avec l'axe 1 ni avec l'axe 2 de l'ACP. C'est pour cette raison qu'elle est très mal représentée dans le plan des deux premières composantes principales.
8. (2pts) La projection des individus sur les 2 premiers axes de l'ACP fait apparaître différents profils, correspondant aux clusters étiquetés A, B, C et D. En vous référant au cercle des corrélations, proposez une description des individus appartenant à chacun des clusters.
   L'axe 1 est négativement corrélé au pourcentage de masse grasse, et positivement corrélé avec les variables liées à l'exercice physique. Un individu ayant une haute valeur sur l'axe 1 est donc certainement très athlétique. L'axe 2 est lui fortement positivement corrélé au poids et à l'IMC. Les individus du cluster B sont donc d'un poids moyen et très athlétiques, les individus du cluster C sont légers et athlétiques, ceux du cluster D sont globalement assez légers mais peu athlétiques, et enfin les individus du cluster A sont plutôt lourds et moyennement athlétiques.
9. (2pts) Commentez les assertions suivantes en vous appuyant sur les figures : **justifiez vos réponses !**
   a. Les individus plus lourds ont tendance à avoir un pourcentage de masse grasse plus élevé.
   b. Les individus qui font le plus de séances par semaine ont tendance à faire des séances plus longues.
   c. Les individus qui brulent le plus de calories pendant une séance sont ceux qui boivent le plus.
   d. Les individus plus grands ont tendance à boire plus pendant une séance.

a. La variable Poids est pratiquement orthogonale à la variable Pourcentage de masse grasse, les deux variables sont donc quasiment décorrélées, l'assertion a. est fausse.

b. L'angle entre les 2 variables concernées est très faible, les 2 variables sont donc fortement corrélées, l'assertion b. est vraie.

c. Les variables Eau bue et Calories brûlées sont presque orthogonales, donc l'assertion c. est fausse.

d. La variable Taille est mal représentée sur le cercle des corrélations, on ne peut donc rien conclure sur sa corrélation avec les autres variables.

Exercice 2 : Moindres carrés

On s'intéresse à la durée moyenne du jour (photopériode), en heures, au cours d'une année dans une ville donnée (par exemple Paris). On suppose que la photopériode varie de manière quasi sinusoïdale au cours de l'année.

On souhaite ajuster le modèle suivant :

$$P(t) = a \sin\left(\frac{2\pi}{365} t\right) + b \cos\left(\frac{2\pi}{365} t\right) + c,$$

où t est le numéro du jour dans l'année (1 à 365), $P(t)$ est la durée du jour (en heures), et a, b, c sont des scalaires.

On dispose des mesures suivantes :

Jour t_i	1	30	60	90	120	150	180	210	240	270	300	330	360
Durée du jour P_i (h)	8.6	9.2	10.8	12.5	14.0	15.6	16.0	15.1	13.6	11.9	10.3	9.0	8.7

1. (2 pts) Formulez le problème d'ajustement du modèle sinusoïdal de la photopériode au sens des moindres carrés. Donnez la formulation matricielle en explicitant le vecteur de paramètres $\boldsymbol{\beta}$ et les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ associées.
   Le vecteur des paramètres $\boldsymbol{\beta}$ contient les paramètres du modèle, i.e. $\boldsymbol{\beta} = [a, b, c]^T$.
   Le problème d'estimation de $\boldsymbol{\beta}$ au sens des moindres carrés s'écrit donc :

$$\min_{(a,b,c)\in\mathbb{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(a \sin\left(\frac{2\pi}{365} t_i\right) + b \cos\left(\frac{2\pi}{365} t_i\right) + c - P_i \right)^2,$$

   ce que l'on peut réécrire

$$\min_{\beta\in\mathbb{R}^3} \frac{1}{2} ||\mathbf{A}\boldsymbol{\beta} - \mathbf{B}||^2,$$

   en posant

$$\boldsymbol{A} = \begin{bmatrix} \sin\left(\frac{2\pi}{365} t_1\right) & \cos\left(\frac{2\pi}{365} t_1\right) & 1 \\ \sin\left(\frac{2\pi}{365} t_2\right) & \cos\left(\frac{2\pi}{365} t_2\right) & 1 \\ \dots & \dots & \dots \\ \sin\left(\frac{2\pi}{365} t_n\right) & \cos\left(\frac{2\pi}{365} t_n\right) & 1 \end{bmatrix} \text{ et } \boldsymbol{B} = \begin{bmatrix} P_1 \\ P_2 \\ \dots \\ P_n \end{bmatrix},$$

   avec $n = 13$.

2. (1 pt) Quelle est la dimension de la matrice $\boldsymbol{A}$? De quel rang doit être la matrice pour que le problème admette une solution ?
   La matrice $\boldsymbol{A}$ est de dimension 13×3, et doit être de rang 3 afin que la matrice $\boldsymbol{A}^T\boldsymbol{A}$ soit inversible.

3. (1 pt) Donnez l'expression analytique de la solution à partir des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$.

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

   où $(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ est la matrice pseudo-inverse de $\boldsymbol{A}$.

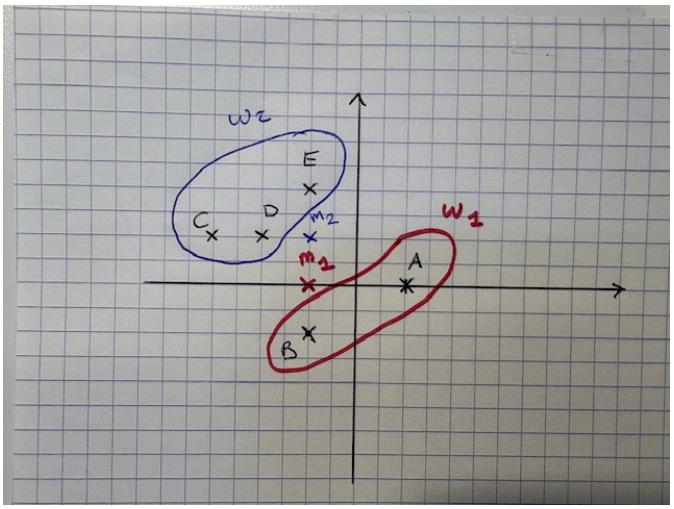
Exercice 3 : Algorithme K-means et algorithme EM

On considère 5 points de $\mathbb{R}^2$ définis comme suit

$$A = \begin{bmatrix}1\\0\end{bmatrix}, B = \begin{bmatrix}-1\\-1\end{bmatrix}, C = \begin{bmatrix}-3\\1\end{bmatrix}, D = \begin{bmatrix}-2\\1\end{bmatrix}, E = \begin{bmatrix}-1\\2\end{bmatrix}$$

que l'on désire regrouper en deux classes ω_1 et ω_2.

1. **Algorithme kmeans** : Représenter ces 5 points et expliquer quel sera le résultat de la première itération de l'algorithme k-means si les représentants initiaux des deux classes sont $\boldsymbol{m}_1 = (-1, 0)^T$ et $\boldsymbol{m}_2 = (-1, 1)^T$ (on entourera les points associés à ω_1 et ceux associés à ω_2). Déterminer les représentants utilisés à la seconde itération. (1pt)



Les représentants utilisés à la seconde itération seront donc

$$\boldsymbol{m}_1^1 = \frac{\boldsymbol{A}+\boldsymbol{B}}{2} = \begin{bmatrix}0\\-\frac{1}{2}\end{bmatrix} \text{ et } \boldsymbol{m}_2^1 = \frac{\boldsymbol{C}+\boldsymbol{D}+\boldsymbol{E}}{3} = \begin{bmatrix}-2\\\frac{4}{3}\end{bmatrix}.$$

2. **Algorithme EM** : Une autre manière de regrouper les 5 points en deux classes est de modéliser la loi de ces points à l'aide d'un mélange de deux lois gaussiennes dont les paramètres peuvent être estimés à l'aide de l'algorithme EM.
   (a) (2pts) Déterminer la probabilité d'associer le point E à la classe ω_1 lors d'une itération de l'algorithme EM, si les paramètres de l'itération précédente sont $\boldsymbol{m}_1 = (-1, 0)^T$, $\boldsymbol{m}_2 = (-1, 1)^T$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbb{I}_2$ (où $\mathbb{I}_2$ est la matrice identité de taille 2×2) et $\pi_1 = \pi_2 = \frac{1}{2}$. Vérifier que cette probabilité est inférieure à 0.5.
   D'après le cours, cette probabilité est

$$P(y_E = \omega_1 | E, \boldsymbol{\theta}) = \frac{\pi_1 p(E|y_E = \omega_1, \boldsymbol{\theta})}{\sum_{k=1}^{2} \pi_k p(E|y_E = \omega_k, \boldsymbol{\theta})}. \quad (1)$$

Comme les matrices de covariance des densités $p(E|y_E = \omega_1, \boldsymbol{\theta})$ et $p(E|y_E = \omega_2, \boldsymbol{\theta})$ sont égales à la matrice identité, on a

$$p(E|y_E = \omega_1, \boldsymbol{\theta}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(E - \boldsymbol{m}_1)^T(E - \boldsymbol{m}_1)\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\|E - \boldsymbol{m}_1\|^2\right),$$

et donc

$$p(E|y_E = \omega_2, \boldsymbol{\theta}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\|E - \boldsymbol{m}_2\|^2\right).$$

Mais $\|E - \boldsymbol{m}_1\|^2 = 4$ et $\|E - \boldsymbol{m}_2\|^2 = 1$, donc

$$p(E|y_E = \omega_1, \boldsymbol{\theta}) = \frac{1}{2\pi} \exp(-2) \text{ et } p(E|y_E = \omega_2, \boldsymbol{\theta}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\right).$$

Donc

$$p(y_E = \omega_1|E, \boldsymbol{\theta}) = \frac{\exp(-2)}{\exp(-2) + \exp\left(-\frac{1}{2}\right)} = \frac{1}{1 + \exp\left(\frac{3}{2}\right)} < \frac{1}{2}.$$

(b) (2pts) On suppose qu'à une itération de l'algorithme EM, les responsabilités associées à la classe ω_1 sont $\delta(1|A) = \delta(1|B) = 0.8$ et $\delta(1|C) = \delta(1|D) = \delta(1|E) = 0.3$. Déterminer $\widehat{\pi}_1$ l'estimation de la probabilité a priori de la classe ω_1 et $\widehat{\mu}_1$ l'estimation du vecteur moyenne de la classe ω_1 issues de ces responsabilités.
D'après le cours, on a :

$$\widehat{\pi}_j = \frac{\widehat{n}_j}{n} \qquad \text{avec} \qquad \widehat{n}_j = \sum_{i=1}^{n} \delta(j|i),$$

et

$$\widehat{\boldsymbol{\mu}}_j = \frac{1}{\widehat{n}_j} \sum_{i=1}^{n} \delta(j|i)\boldsymbol{x_i}.$$

Donc

$$\widehat{\pi}_1 = \frac{2 \times 0.8 + 3 \times 0.3}{5} = \frac{1}{2},$$

et

$$\widehat{\boldsymbol{\mu}}_1 = \frac{2}{5}\begin{bmatrix} 0.8 - 0.8 - 0.9 - 0.6 - 0.3 \\ -0.8 + 0.3 + 0.3 + 0.6 \end{bmatrix} = \begin{bmatrix} -18/25 \\ 4/25 \end{bmatrix}.$$