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

1. (1pt) Expliquer ce qu'est une matrice de confusion pour un classifieur travaillant en mode supervisé. Donner un exemple pour quatre classes en expliquant les différents termes de la matrice.
2. (2pt) Expliquer le principe du classifieur Bayésien. Que doit-on connaître pour mettre en oeuvre ce classifieur ? Expliquer avec soin votre réponse.
3. (1pt) Expliquer le principe de l'algorithme des k moyennes (k-means).
4. (1pt) Dans un problème de classification à deux classes ω_1 et ω_2, quelle est la valeur de l'indice de Gini d'un ensemble possédant 4 éléments de ω_1 et 6 éléments de ω_2 ? Comment utilise-t-on cet indice pour scinder une branche d'un arbre de classification en deux parties ?

Exercice 2 : Moindres carrés

La quantification de l'abondance d'une espèce dans son habitat est un problème essentiel en écologie. On considère généralement qu'une relation lie le nombre d'individus N à la surface S de l'habitat en question par la formule suivante :

$$N = aS^b$$

On dispose de k couples d'observations (N_i, S_i), obtenus en comptant les représentants d'espèces de fleurs dans des prairies des Pyrénées, et en mesurant la surface de ces prairies. On voudrait ainsi estimer la valeur des paramètres a et b à partir de ces observations.

1. (2pt) Formulez le problème d'estimation des paramètres a et b au sens des moindres carrés et donnez-en la formulation matricielle en explicitant les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ associées.
2. (1pt) Expliquez comment estimer les paramètres a et b à partir des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$.

Exercice 3 : ACP

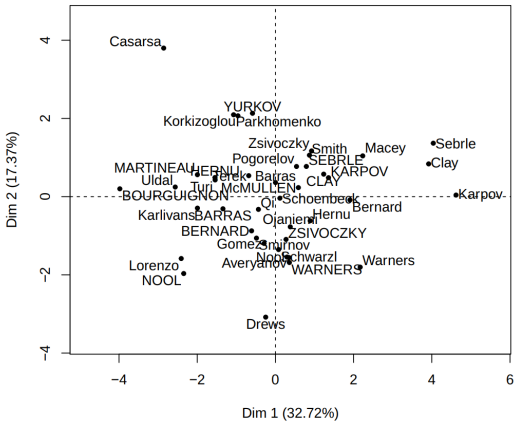
On dispose d'une base de données répertoriant les performances de plusieurs athlètes lors de deux épreuves : le décathlon des jeux olympiques d'Athènes (août 2004) et le Decastar organisé à Talence en Gironde (septembre 2004). Une partie des données est présentée dans le tableau ci-dessous :

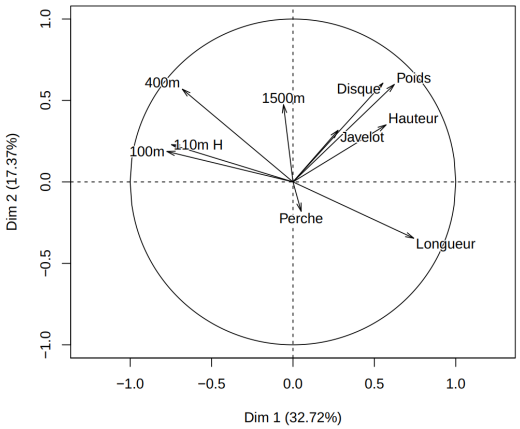
Athlète	100m	Longueur	Poids	Hauteur	400m	110m H	Disque	Perche	Javelot	1500m	Points
...											
A	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5	51.53	274.21	7926
B	11.36	6.68	14.92	1.94	53.2	15.39	48.66	4.4	58.62	296.12	7404
C	10.5	7.81	15.93	2.09	46.81	13.97	51.65	4.6	55.54	278.11	8725
D	11.1	7.03	13.22	1.85	49.34	15.38	40.22	4.5	58.36	263.08	7592
...											

Ce tableau de données compte au total 41 lignes, correspondant aux performances de 41 athlètes (certains athlètes ont participé aux 2 épreuves et apparaissent donc 2 fois ; les participants au Decastar ont leur nom écrit en lettres majuscules). En plus des résultats aux 10 épreuves du décathlon, on dispose d'une colonne supplémentaire précisant le nombre de points obtenus par l'athlète lors de l'épreuve.

1. (1pt) Quelles sont les dimensions de la matrice $\boldsymbol{X}$ des données du problème, et de la matrice de variance-covariance $\boldsymbol{\Sigma}$ associée ?
2. (1pt) Expliquez pourquoi il est important de centrer-réduire les données de cette base.

Voici une représentation des données projetées sur les 2 premiers axes ainsi que le cercle de corrélation des variables :





3. (1pt) Certaines flèches (par exemple, 100m et 110m H) pointent dans une direction très proche. Expliquez ce que cela signifie.
4. (1pt) La variable additionnelle "Points" a une corrélation de 0.92 avec l'axe 1 de l'ACP et de -0.03 avec l'axe 2. Représentez la variable sur le cercle des corrélations.
5. (1pt) D'après vous, quelle est l'information portée par l'axe 1 de l'ACP ?

6. (1pt) Au vu du graphe des individus, diriez-vous que les meilleures performances ont été réalisées lors du Decastar ou des Jeux Olympiques (on rappelle que les données du Decastar sont représentées en lettres majuscules et que celles des jeux olympiques le sont en lettres minuscules) ? Justifiez votre réponse en vous appuyant sur le graphique.
7. (1pt) Les athlètes A, B, C et D ont été anonymisés : il s'agissait de Lorenzo, Karpov, Casarsa et Drews. En vous appuyant sur le graphe des individus et le cercle des corrélations, associez chaque athlète a la ligne correspondante dans le tableau (A, B, C ou D).

Exercice 4 : machines à vecteurs supports

On considère un problème de classification supervisée à deux classes avec une base d'apprentissage $\mathcal{B} = \{(\boldsymbol{x}_i, y_i), i = 1, ..., n\}$ avec $\boldsymbol{x}_i \in \mathbb{R}^d$ et $y_i \in \{-1, +1\}$. Le classifieur SVM cherche une fonction de décision de la forme $f(\boldsymbol{x}) = \text{sign}\left[\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b\right]$ avec $b \in \mathbb{R}$. Le vecteur $\boldsymbol{w}$ s'obtient par résolution du problème

$$\begin{cases} \min\limits_{\substack{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R} \\ (\xi_1,\ldots,\xi_n)\in\mathbb{R}^n}} \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i \right\} \\ \textbf{sous les contraintes} \begin{cases} y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq 1 - \xi_i & \forall i \in \{1,\ldots,n\} \\ \xi_i \geq 0 & \forall i \in \{1,\ldots,n\} \end{cases} \end{cases}$$

1. Expliquer l'intérêt du terme $C\sum\limits_{i=1}^{n}\xi_i$.
2. Comment choisit-on la valeur de C ?
3. La solution du problème ci-dessus est $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i)$ où les réels α_i sont positifs ou nuls. Expliquer pour quels vecteurs on a $\alpha_i > 0$ et $\alpha_i = 0$.
4. Donner la règle de classification lorsqu'on utilise un noyau κ tel que $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$.
5. Donner un exemple de noyau vu en cours.