Partiel Analyse de Données

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

Questions de cours

1. (1pt) Expliquer ce qu'on entend par "classes linéairement séparables".
2. (1pt) Qu'appelle-t-on méthode du "leave-one out" ?
3. (1pt) Dans le classifieur "soft-margin" SVM chaque vecteur $\boldsymbol{x}_i$ de classe y_i doit vérifier la contrainte $y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 1 - \xi_i$. Quelle est l'utilité de la variable ξ_i ?
4. (1pt) Donner au moins deux inconvénients de l'algorithme K-means.
5. (1pt) Expliquer ce que sont les points "core", "border" et "noise" dans l'algorithme DBSCAN.

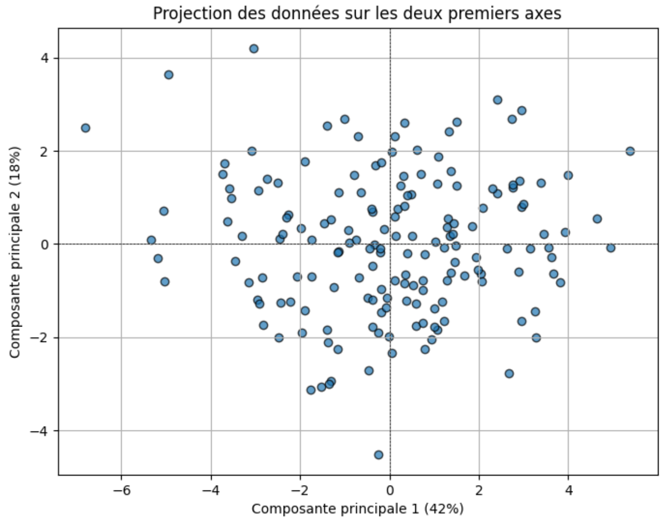
Exercice 1 : ACP

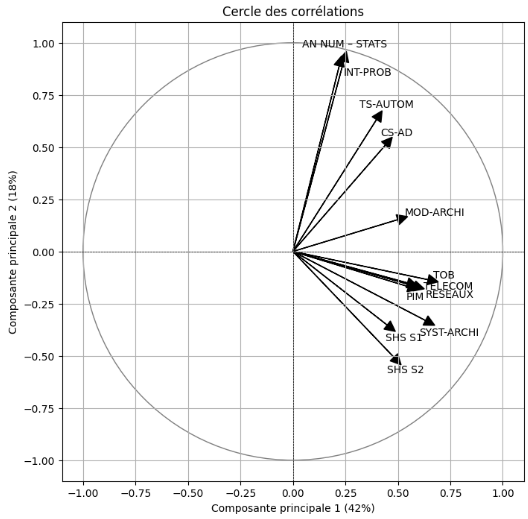
Pour cet exercice, on dispose d'un bordereau des notes obtenues par les 1SN pendant l'année 2022-2023. Chaque étudiant.e a obtenu 12 notes, une pour chacune des UEs suivantes : Sciences Humaines et Sociales Semestres 1 et 2, Architecture des ordinateurs et Système (`SYST-ARCHI`), Programmation Impérative (`PIM`), Technologies Objet (`TOB`), Telecommunications, Réseaux, Modélisation et Architecture (`MOD-ARCHI`), Calcul Scientifique et Analyse de Données (`CS-AD`), Traitement du Signal et Automatique (`TS-AUTOM`), Intégration et Probabilités (`INT-PROB`), et Analyse Numérique et Statistiques (`AN NUM-STATS`).

On se propose dans cet exercice de faire une Analyse en Composantes Principales de ce tableau de données.

1. (1pt) D'après vous, est-il nécessaire de centrer et/ou réduire ces données pour appliquer l'ACP ? Justifiez votre réponse.
2. (1pt) Combien la matrice de variance-covariance de ces données compte-t-elle de valeurs propres ? Expliquez comment l'on peut construire la base d'un espace à deux dimensions dans lequel projeter les données.

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





3. (1pt) Entourez sur la figure le point correspondant, selon vous, à la projection de l'étudiant.e qui majore la promotion. Notez ce point `A`. **Justifiez votre réponse**.
4. (1pt) Entourez sur la figure un point correspondant à un.e étudiant.e qui serait excellent.e dans les matières informatiques (type `PIM` ou `TOB`), mais qui obtiendrait de mauvaises notes en mathématiques (`INT-PROB`, `AN NUM-STATS`). Notez ce point `B`. **Justifiez votre réponse**.
5. (1pt) D'après vous, que peut-on dire de la note d'`INT-PROB` pour un.e étudiant.e qui a eu une bonne note en `AN NUM-STATS` ?
6. (1pt) D'après vous, que peut-on dire de la note de `RESEAUX` pour un.e étudiant.e qui a eu une mauvaise note en `TELECOM` ?
7. (1pt) D'après vous, que peut-on dire de la note de `SYST - ARCHI` pour un.e étudiant.e qui a eu une mauvaise note en `AN NUM-STATS` ?

Exercice 2 : Moindres carrés

On cherche à ajuster une ellipse à des données expérimentales en utilisant la méthode des moindres carrés. Une ellipse peut être décrite par l'équation générale suivante :

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

En réalité, les coefficients sont définis à une constante près, car pour tout $\alpha \in \mathbb{R}$, on a

$$\alpha ax^2 + \alpha bxy + \alpha cy^2 + \alpha dx + \alpha ey + \alpha f = 0$$

On choisit donc de poser $f = -1$, et on s'intéresse maintenant à l'équation de l'ellipse :

$$ax^2 + bxy + cy^2 + dx + ey = 1$$

On cherche à déterminer les paramètres de cette ellipse. On dispose pour cela de n points P_i de coordonnées (x_i, y_i) dont on sait qu'ils sont au voisinage de l'ellipse.

1. (2 pts) Formulez le problème d'ajustement des paramètres de l'ellipse 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.
2. (1 pt) Expliquez comment estimer les paramètres de l'ellipse à partir des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$.

Exercice 3 : classification Bayésienne pour lois de Rayleigh

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

$$f(x|\omega_i) = \frac{x}{\sigma_i^2} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) I_{\mathbb{R}^+}(x) \quad i = 1, 2 \tag{1}$$

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

1. Montrer que la règle de classification associée à ce problème lorsque les deux classes sont équiprobables consiste à classifier x dans la classe ω_1 si $x > a$, où a est une constante dépendant de σ_1^2 et de σ_2^2.
2. Déterminer la probabilité d'erreur associée à ce classifieur.
3. Que devient la règle de classification de la première question lorsque $P(\omega_1) > P(\omega_2)$? Commenter le résultat obtenu.
4. Si le paramètre σ_1^2 est inconnu, expliquer comment l'estimer à partir de données d'apprentissage de la classe ω_1 en utilisant la méthode du maximum de vraisemblance.