

Examen Final du Data Mining

Master: I2A & STIC – Année: 2025/2026 – Durée: 2 h – R.M: Ali LALOUCI

Exercice 1 (6 points)

Répondez brièvement aux questions suivantes :

- 1) Que ce que le nettoyage de données ?
- 2) Citez quatre méthodes pour la normalisation de données ?
- 3) C'est quoi la Visualisation de données (Data Visualization) ?
- 4) Citez les causes de données manquantes, et comment les gérer ?
- 5) Expliquez le coefficient de corrélation ?
- 6) Citez et expliquez brièvement les méthodes de traitement et d'analyse de données dans le cas multidimensionnel ?

Exercice 2 (5 points)

On examine l'évaluation d'une variable y en fonction de deux variables exogènes x et z , on dispose de n observations de ces variables. Si on note $X = (1, x, z)$, où 1 est le vecteur constant et x et z sont les vecteurs des variables explicatives. Nous avons obtenu les résultats suivants :

$$X^t X = \begin{bmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{bmatrix}$$

- 1) Dans le cas général, donnez la forme matricielle de : X , X^t , et $X^t X$?
- 2) Déduire les valeurs manquantes. Que vaut n ?
- 3) Sachant que la régression de Y sur $(1, x, z)$ donne : $y = -1.6 + 0.61x + 0.46z + \mathcal{E}$. Déterminez la moyenne de y ?

Exercice 3 (5 points)

Soit l'ensemble D des éléments entiers suivants :

$D = \{ 2, 5, 8, 10, 11, 18, 20 \}$

On veut répartir les données de D en trois (3) clusters, en utilisant l'algorithme K-means. La distance d'entre deux nombres a et b est calculée ainsi : $d(a, b) = |a - b|$.

Questions :

- 1) Appliquez K-means en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul.
- 2) Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires.
- 3) Peut-on avoir un nombre d'itérations inférieur pour ce problème ? Discutez.

Exercice 4 (interrogation) (4 points)

On considère le petit jeu de données suivant :

Étudiant	Note (/20)	Âge (ans)	Filière	Note Normalisée (Min–Max)	Age normalisé (Score Z)	Filière Encodée (Label Encoder)
1	10	18	Informatique			
2	14	20	Math			
3	8	19	Informatique			
4	16	22	Physique			
5	12	21	Math			

Questions :

- 1) Appliquer la normalisation Min–Max aux notes (complétez le tableau).
- 2) Calculer la moyenne et l'écart-type des âges.
- 3) Calculer le score Z pour chaque âge (complétez le tableau).
- 4) Attribuer un code numérique (Label Encoder) à chaque catégorie de filière (complétez le tableau).

Bon courage

Solution

Exercice 1(7 points)

Répondez brièvement aux questions suivantes :

1. Que ce que le nettoyage de données ? **(1 points)**
 - Gestion des valeurs manquantes (imputation par la moyenne, la médiane, etc., ou suppression),
 - détection et correction des erreurs ou des valeurs aberrantes (outliers) qui pourraient biaiser l'analyse
2. Citez quatre méthodes pour la normalisation de données ? **(1 points)**
Normalisation Min-Max, Normalisation par Score Z, Mise à l'échelle Robuste (Robust Scaling), Label Encoder.
3. C'est quoi la Visualisation de données(Data Visualization) ? **(1 points)**
Utilisation des techniques (de graphiques, tableaux de bord et autres représentations visuelles) pour simplifier et communiquer clairement les informations complexes extraites.
4. Les causes de données manquantes : **(0.5 points)**
 - dysfonctionnement de l'équipement
 - incohérentes avec d'autres données enregistrées et donc supprimées
 - données non saisies en raison d'un malentendu
 - certaines données peuvent ne pas être considérées comme importantes

Comment gérer les données manquantes ? **(0.5 points)**

- a) Utiliser l'attribut moyen
 - b) Utiliser l'attribut moyen pour tous les échantillons appartenant à la même classe.
 - c) Utiliser la valeur la plus probable : *basée sur l'inférence, telle que la formule bayésienne ou l'arbre de décision.*
5. Expliquez le coefficient de corrélation ? **(1 points)**
 $r = \text{cov}(x,y)/\text{sd}(x)*\text{sd}(y)$ avec: $-1 \leq r \leq 1$
 - a) Si $r = 1$: Il existe une relation linéaire
 - b) Si $r = 3/4$: Il existe une liaison forte
 - c) r proche de 0 : moins les deux variables sont corrélées
 - d) Si $r = 0$: Les deux variables X et Y ne sont pas corrélées
6. Citez les méthodes de traitement de données dans le cas multidimensionnel?**(1 points)**
 - a) Analyse en composantes principales (ACP) : réduire p variables corrélées en q variables non corrélées.
 - b) Analyse factorielle des correspondances (AFC): trouver des liens ou correspondances entre deux variables qualitatives nominales
 - c) Analyse des correspondances multiples (ACM : L'ACM est l'équivalent de l'ACP pour les variables qualitatives et elle se réduit à l'AFC lorsque le nombre de variables qualitatives est égal à 2

Exercice 2(6 points)

1. Donnez la forme matricielle de : X, X^t , et X^tX ? **(1.5 points)**

$$X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_n & z_n \end{pmatrix} \quad X^t = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ z_1 & z_2 & \dots & z_n \end{pmatrix} \quad X^tX = \begin{pmatrix} n & \sum x_i & \sum z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i \\ \sum z_i & \sum x_i z_i & \sum z_i^2 \end{pmatrix}$$

2. Déduire les valeurs manquantes. Que vaut n ? **(2 points)**

n=25

$$X^tX = \begin{pmatrix} 25 & 0 & 0 \\ 0 & 9.3 & 5.4 \\ 0 & 5.4 & 12.7 \end{pmatrix}$$

3. Déterminez la moyenne de y ? **(1.5 points)**

$$Moy(y) = -1.6 + 0.61 * Moy(x) + 0.46 Moy(z)$$

$$Moy(x) = 1/n \sum x_i = 1/25 * 0 = 0$$

$$Moy(z) = 1/n \sum z_i = 1/25 * 0 = 0$$

$$\text{Donc : } Moy(y) = -1.6$$

Exercice 3 (5 points)

1) Appliquez K-means en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul. **(3 points)**

Itération 1: (1 point)

x	d(x,8)	d(x,10)	d(x,11)	Cluster
2	6	8	9	Cl ₁
5	3	5	6	Cl ₁
8	0	2	3	Cl ₁
10	2	0	1	Cl ₂
11	3	1	0	Cl ₃
18	10	8	7	Cl ₃
20	12	10	9	Cl ₃

(0.5 point)

Clusters obtenus : Cl₁={2,5,8}, Cl₂={10}, Cl₃={11,18,20}. **(0.25 point)**

Les nouveaux centres : **(0.25 point)**

- C₁ = (2+5+8)/3 = 5
- C₂ = 10
- C₃ = (11+18+20)/3=49/3≈16,

Itération 2

x	d(x,5)	d(x,10)	d(x,16,33)	Cluster
2	3	8	14,33	Cl ₁
5	0	5	11,33	Cl ₁
8	3	2	8,33	Cl ₂
10	5	0	6,33	Cl ₂
11	6	1	5,33	Cl ₂
18	13	8	1,67	Cl ₃
20	15	10	3,67	Cl ₃

Clusters obtenus: Cl₁={2,5}, Cl₂={8,10,11}, Cl₃={18,20}

Les nouveaux centres :

- C₁(2)=2+5/2=3,5
- C₂(2)=8+10+11/3=9,67
- C₃(2)= 18+20/2=19

Itération 3

x	d(x,3,5)	d(x,9,67)	d(x,19)	Cluster
2	1,5	7,67	17	C ₁
5	1,5	4,67	14	C ₁
8	4,5	1,67	11	C ₂
10	6,5	0,33	9	C ₂
11	7,5	1,33	8	C ₂
18	14,5	8,33	1	C ₃
20	16,5	10,33	1	C ₃

Les clusters ne changent plus, donc : arrêt

2) Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires. **1.point)**

3 itérations, avec Clusters finaux

- Cluster 1 : {2, 5} → centre = 3,5
- Cluster 2 : {8, 10, 11} → centre ≈ 9,67
- Cluster 3 : {18, 20} → centre = 19

4. Peut-on avoir un nombre d'itérations inférieur pour ce problème ? Discutez. **(1 point)**

Oui, c'est possible car le nombre d'itérations dépend fortement :

- Du choix des centres initiaux
- De la répartition des données
- Si les centres initiaux avaient été plus proches des centres finaux (par exemple 3, 10 et 19), l'algorithme aurait convergé en 1 ou 2 itérations.

Exercice 4

Étudiant	Note	Âge	Filière	Note Min-Max	Âge Score Z	Filière encodée
A	10	18	Informatique	0,25	-1,42	0
B	14	20	Math	0,75	0,00	1
C	8	19	Informatique	0,00	-0,71	0
D	16	22	Physique	1,00	+1,42	2
E	12	21	Math	0,50	+0,71	1

1) Appliquer la normalisation Min-Max aux notes (complétez le tableau). **(1 point)**

$$x' = (x - \min) / (\max - \min)$$

les calculs: $(10-8)/8$, ..., $(12-8)/8$.

2) Calculer la moyenne et l'écart-type des âges. **(1 point)**

$$\mu = 20 \quad \sigma = \sqrt{2} \approx 1,41$$

3) Calculer le score Z pour chaque âge (complétez le tableau). **(1 point)**

$$Z = (x - \mu) / \sigma$$

les calculs: $(18-20)/1,41$, ..., $(21-20)/1,41$.

4) Attribuer un code numérique (Label Encoder) à chaque catégorie de filière (complétez le tableau). **(1 point)**

Catégories distinctes : Informatique, Math, Physique

Encodage choisi : 0, 1, 2