Université Abbas Laghrour de Khenchela
Département d'Informatique

Matière: Data mining **Master 2 IA**

Examen final

Date : 20/01/2026 **Duré : 1.5 h**

Questions : (6 points)

1) Donner une définition des concepts: Données transactionnelles et données de séries chronologiques
2) Pourquoi est-il important de réduire la dimensionnalité d'un jeu de données ? Citer deux approches de réduction de la dimensionnalité des données.
3) Quelles est la différence entre la normalisation et la standardistaion des données ?
4) Que représentent les valeurs propres et les vecteurs propres dans la PCA ?

Exercice 1 (7 points)

On considère le jeu de données suivant utilisé pour la prédiction du diabète. Chaque ligne représente un patient.
Jeu de données initial :

ID_patient	Âge	Glycémie	IMC	Pression	Sexe	Code_hôpital	Diagnostic
1	45	180	28.5	80	F	H01	1
2	50	100	31.2	85	M	H02	1
3	37	95	22.1	70	F	H01	0
4	45	180	28.5	80	F	H01	1
5	29	400	19.0	65	M	H03	0
6	NaN	110	27.0	75	F	H02	0
7	60	130	NaN	90	M	H01	1

1) Identifier quatre problèmes dans ce jeu de données.
2) Proposer un pipeline de prétraitement (étapes successives).
3) Identifier les valeurs aberrantes dans la colonne *Glycémie* utilisant la méthode Z-Score, prenant le seuil : $|z| > 3$.

Exercice 2 (7 points)

On souhaite prédire la consommation d'électricité Y (en kWh) en fonction de deux variables :

- X1: température extérieure moyenne (en °C)
- X2: nombre d'appareils électriques utilisés

On dispose du jeu de données suivant :

Instance	X1 (°C)	X2 (nb appareils)	Y (kWh)
1	10	2	50
2	20	3	65
3	30	5	90

1) Écrire le modèle de régression linéaire multiple associé à ce problème.
2) Représenter ce jeu de données par la forme matricielle.
3) Calculer les coefficients de régression de ce modèle utilisant la méthode des moindres carrés
4) Donner une interprétation au coefficient de régression trouvées

Bon courage

Corriger type

Questions

1. Données transactionnelles vs données de séries chronologiques (1.5)

- **Données transactionnelles** :
  Données organisées sous forme de transactions indépendantes, chacune contenant un ensemble d'éléments (ex. achats d'un client).
- **Données de séries chronologiques** :
  Données ordonnées dans le temps, où chaque observation dépend de l'instant temporel (ex. température quotidienne).

2. Importance de la réduction de dimensionnalité (1.5)

La réduction de dimensionnalité permet de :

- Réduire la complexité et le coût de calcul
- Limiter le surapprentissage
- Améliorer la visualisation et l'interprétation

Deux approches :

- **Sélection de variables** (ex. sélection par corrélation)
- **Extraction de caractéristiques** (ex. PCA)

3. Différence entre normalisation et standardisation (1.5)

- **Normalisation** :
  Met les données dans un intervalle fixe (ex. [0,1]).
- **Standardisation** :
  Centre les données autour de 0 avec un écart-type égal à 1.

4. Valeurs propres et vecteurs propres en PCA (1.5)

- **Vecteurs propres** :
  Directions des composantes principales.
- **Valeurs propres** :
  Quantité de variance expliquée par chaque composante principale.

Exercice 1

1. **Identification des problèmes dans les données (02)**

On peut identifier les problèmes suivants :

1. Valeurs manquantes
- **Âge** du patient 6 : NaN
- **IMC** du patient 7 : NaN

2. Données dupliquées
- Les lignes **1 et 4** sont strictement identiques

3. Valeurs aberrantes (outliers)
- La **glycémie = 400** (patient 5) est anormalement élevée par rapport aux autres valeurs

4. Variables non pertinentes
- **ID_patient** : identifiant sans valeur prédictive

2. **Proposition d'un pipeline de prétraitement (02)**

Un pipeline de prétraitement possible est :

Étape 1 : Suppression des colonnes non utiles
- Supprimer ID_patient

Étape 2 : Suppression des doublons
- Éliminer les lignes dupliquées (patients 1 et 4)

Étape 3 : Traitement des valeurs manquantes
- Remplacer :
  - l'âge manquant par la **moyenne** ou la **médiane**
  - l'IMC manquant par la **moyenne**

Étape 4 : Détection et traitement des valeurs aberrantes
- Méthode Z-Score ou IQR
- Suppression ou capage des valeurs extrêmes

Étape 5 : Encodage des variables catégorielles
- Sexe → encodage binaire (0/1)
- Code_hôpital → One-Hot Encoding (si conservé)

Étape 6 : Normalisation / standardisation
- Appliquée aux variables numériques (Âge, Glycémie, IMC, Pression)

3. **Détection des valeurs aberrantes dans la Glycémie (méthode Z-Score) (03)**

Formule du Z-Score

$$Z = \frac{x - \mu}{\sigma}$$

Valeurs de la glycémie : [180, 100, 95, 180, 400, 110, 130]

Moyenne : $\mu = \frac{1195}{7} \approx 170.7$

Écart-type (≈) σ≈107.1

Calcul du Z-Score pour la valeur suspecte (400)

$$Z_{400} = \frac{400 - 170.7}{107.1} \approx 2.14$$

Interprétation
- Le seuil est **|Z| > 3** → aucune valeur aberrante détectée

Exercice 2

1. Modèle de régression linéaire multiple (01)

Le modèle de régression linéaire multiple s'écrit : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Où :

- β_0: constante (consommation de base),
- β_1: effet de la température extérieure,
- β_2: effet du nombre d'appareils électriques.

2. La forme matricielle (01)

$$X = \begin{bmatrix} 1 & 10 & 2 \\ 1 & 20 & 3 \\ 1 & 30 & 5 \end{bmatrix} \quad Y = \begin{bmatrix} 50 \\ 65 \\ 90 \end{bmatrix}$$

3. Calcul des coefficients de régression (03)

Étape 1 : Calcul de X^T

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 10 & 20 & 30 \\ 2 & 3 & 5 \end{bmatrix}$$

Étape 2 : Calcul de $X^T X$

$$X^T X = \begin{bmatrix} 3 & 60 & 10 \\ 60 & 1400 & 230 \\ 10 & 230 & 38 \end{bmatrix}$$

Étape 3 : Calcul de $X^T Y$

$$X^T Y = \begin{bmatrix} 205 \\ 4550 \\ 765 \end{bmatrix}$$

Étape 4 : Résolution du système

On résout :

$$(X^T X)\boldsymbol{\beta} = X^T Y$$

$$\begin{bmatrix} 3 & 60 & 10 \\ 60 & 1400 & 230 \\ 10 & 230 & 38 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 205 \\ 4550 \\ 765 \end{bmatrix}$$

La résolution de ce système linéaire donne :

$$\boldsymbol{\beta} = \begin{bmatrix} 25 \\ 0.5 \\ 10 \end{bmatrix}$$

Modèle final obtenu

$$\boxed{Y = 25 + 0.5X_1 + 10X_2}$$

4. Interprétation des coefficients de régression (02)

1. Constante :25

La valeur **25** représente la **valeur de y** lorsque **x1=0** et **x2=0**.

Interprétation :
Même en absence de x1 et x2, la valeur attendue de y est **25**.

2. Coefficient de x1 :0.5

- Le coefficient **0.5** signifie que'une augmentation d'une unité de x1 entraîne une **augmentation moyenne de 0.5 unités de y**.

Interprétation :
L'effet de x1 sur y est **positif mais modéré**.

3. Coefficient de x2 :10

- Le coefficient **10** indique que, **à x1 constant**, une augmentation d'une unité de x2 entraîne une **augmentation moyenne de 10 unités de y**.

Interprétation :
x2 a un **impact fort et significatif** sur y par rapport à x1.