Révision contrôle : Modèles Statistiques

Exercice 1 : Régression Linéaire

Une entreprise lance une campagne publicitaire dans plusieurs villes et souhaite analyser l'effet de la dépense publicitaire sur le nombre de produits vendus. On mesure :

- X : le budget publicitaire investi dans chaque ville (en milliers de dirhams),
- Y : le nombre de produits vendus dans cette ville.

Les données pour 10 villes sont présentées ci-dessous :

Ville	**Budget Publicité (X)**	**Produits Vendus (Y)**
1	2	18
2	3	24
3	5	32
4	4	27
5	6	36
6	8	45

Table 1: Données de budget publicitaire et ventes

1. Calculer les moyennes $\bar{X}$ et $\bar{Y}$, ainsi que les variances $\mathrm{Var}(X)$ et $\mathrm{Var}(Y)$.
2. Calculer la covariance $\mathrm{Cov}(X, Y)$ entre les deux variables.
3. Déterminer le coefficient de corrélation linéaire entre X et Y. Interprétez ce coefficient.
4. Déterminer la droite de régression $Y = a + bX$. Tracez également cette droite sur le nuage de points (X, Y).
5. Évaluer la performance du modèle (par exemple à l'aide du coefficient de détermination R^2).
6. À l'aide du modèle obtenu, estimer combien de produits seront vendus si le budget publicitaire est de 4,5 milliers de dirhams.

1 Exercice 2 : Classification des fruits avec K-means

On considère le tableau suivant ayant comme individus des fruits avec 2 caractéristiques :

- Feature 1 représente le poids du fruit (en grammes).
- Feature 2 représente la teneur en sucre (en grammes).

Fruit	**Poids (g)**	**Teneur en sucre (g)**
Fruit1	150	12
Fruit2	160	14
Fruit3	180	15
Fruit4	40	5
Fruit5	50	7
Fruit6	60	6

1. Représenter le nuage des points en identifiant les fruits par F1, F2, etc.
2. Appliquer l'algorithme K-means pour $K = 2$ en affectant les centroïdes aux fruits suivants :
   - Centroïde 1 = Fruit1
   - Centroïde 2 = Fruit4

   Représenter les deux clusters en nuage de points.

Exercice 3: Forêt aléatoire (Crédit Bancaire)

1.1 Données Initiales

On s'intéresse cette fois à un mini-jeu de données de classification illustrant si une **demande de crédit** est acceptée (**Oui**) ou refusée (**Non**) en fonction de plusieurs facteurs : le **revenu du client**, la **durée du contrat de travail**, et l'**âge**.

ID	Revenu	Contrat	Âge	Crédit Accepté ?
1	Bas	Court	Moyen	Non
2	Bas	Long	Jeune	Oui
3	Moyen	Court	Moyen	Non
4	Élevé	Long	Jeune	Oui
5	Élevé	Long	Âgé	Oui
6	Bas	Long	Âgé	Non
7	Moyen	Long	Âgé	Non
8	Élevé	Court	Moyen	Oui
9	Moyen	Long	Jeune	Oui

Table 2: Jeu de données initial (crédit bancaire)

1.2 Premier Bootstrap (Échantillon n°1)

Supposons avoir obtenu l'échantillon suivant :

ID tiré	Revenu	Contrat	Âge	Crédit Accepté ?
3	Moyen	Court	Moyen	Non
1	Bas	Court	Moyen	Non
4	Élevé	Long	Jeune	Oui
5	Élevé	Long	Âgé	Oui
6	Bas	Long	Âgé	Non
3	Moyen	Court	Moyen	Non
7	Moyen	Long	Âgé	Non
2	Bas	Long	Jeune	Oui
8	Élevé	Court	Moyen	Oui

Table 3: Premier échantillon Bootstrap

1. **Premier Arbre** : Construisez un arbre en se basant sur cet échantillon : Sélectionnez pour la première séparation les deux attributs (**Revenu** et **Âge**) et calculez de l'indice de Gini pour chaque séparation, puis Choisissez celui qui minimise l'impureté. Continuez à splitter jusqu'à obtenir des feuilles pures ou presque pures (toutes **Oui** ou toutes **Non**).

1.3 Autres échantillons et résultats d'arbres

Les arbres de décision obtenus en se basant sur deux autres échantillon sont :

- **Arbre 2** : *Si Contrat = Long, alors Oui ; sinon Non.*
- **Arbre 3** : *Si Revenu = Élevé, alors Oui ; sinon Non.*

1.4 Prédiction avec la Forêt Aléatoire

Supposons qu'un client se présente avec les caractéristiques suivantes :

Revenu = Moyen, Contrat = Court, Âge = Âgé

Question : La demande de crédit est-elle acceptée ou refusée selon **la forêt** ?