



Module 6 - Regroupement

Exercices - Consignes

Exercice 1*

L'objectif de cet exercice est de créer des partitions sur un ensemble de données en utilisant la technique C -moyennes .

Soit l'ensemble de données suivantes :

X	2	3	10	13	21
---	---	---	----	----	----

- Appliquez l'algorithme C -moyennes dans le cas des valeurs de c et des moyennes initiales suivantes :
 - $c = 2$, $\mu_1 = 2$ et $\mu_2 = 21$
 - $c = 3$, $\mu_1 = 2$, $\mu_2 = 13$ et $\mu_3 = 21$
- Calculer l'inertie intra-groupes dans les deux cas précédents.
- Comparez les résultats.

Exercice 2*

La matrice de distance d'un jeu de données est décrite par le tableau suivant :

- Nous désirons regrouper ces données en utilisant le regroupement agglomératif à lien simple.
 - Expliquez le principe du regroupement à lien simple.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

- Illustrez pas à pas et expliquez le regroupement agglomératif à lien simple sur ce jeu de données.
 - Représentez le dendrogramme correspondant.
- Nous désirons regrouper ces données en utilisant le regroupement agglomératif à lien complet.
 - Expliquez le principe du regroupement à lien complet.
 - Illustrez pas à pas et expliquez le regroupement agglomératif à lien complet sur ce jeu de données.
 - Représentez le dendrogramme correspondant.

Exercice 3*

Soit la base de données DataSetFromage qui décrit les fromage selon les 9 attributs suivant : calories, sodium, calcium, lipides, retinol, folates, proteines, cholesterol et magnesium.

- Écrivez les lignes de code qui permettent de lire le fichier Excel.
- Écrivez les lignes de code qui permettent d'afficher le contenu du fichier.
- Calculez les statistiques descriptives de chacune des variables (minimum, maximum, moyenne, mediane).
- Représentez le croisement deux à deux des variables calories, sodium, calcium, lipides.
- Nous désirons regrouper ces données en utilisant le regroupement agglomératif.
 - Écrivez les lignes de code qui permettent d'effectuer une regroupement hiérarchique en utilisant le regroupement à lien simple.
 - Représentez le dendrogramme correspondant.
 - Écrivez les lignes de code qui permettent de représenter le dendrogramme en 4 groupes.

(d) Représentez le dendrogramme en 4 groupes.

Exercice 4

Soit la base de données IRIS qui contient un jeu de données comprenant un total de 150 observations, réparties de manière égale entre les trois espèces de fleurs d'iris (Setosa, Virginica et Versicolor). Ces espèces sont caractérisées par quatre caractéristiques mesurées en centimètres à savoir la longueur et la largeur du sépale et la longueur et la largeur du pétale.

1. Affichez le contenu de la base de données `iris` contenant l'ensemble des données à utiliser.
2. Représentez graphiquement la variation de la longueur du sépale en fonction de la largeur du sépale pour les trois classes. Considérez une couleur différente pour chacune des classes Setosa, Virginica et Versicolor.
3. Représentez graphiquement la variation de la longueur des pétales en fonction de la largeur des pétales pour les trois classes. Considérez une couleur différentes pour chacune des classes Setosa, Virginica et Versicolor.
4. Écrivez les lignes de code qui permettent d'effectuer un regroupement hiérarchique en utilisant :
 - (a) le regroupement à lien simple ;
 - (b) le regroupement à lien complet ;
 - (c) le regroupement à lien moyen ;
5. Représentez le dendrogramme correspondant à chacun des regroupements obtenus précédemment.