Correction TD Segmentation

Préparé par Sarra Zouari

Exercice 1:

1) Étapes de l'algorithme K-Means :

L'algorithme K-Means est un algorithme de clustering (groupement) qui vise à partitionner un ensemble de données en clusters de telle sorte que les données similaires soient regroupées dans le même cluster.

les étapes:

a/ Initialisation : Sélectionnez le nombre de clusters (groupes) que vous souhaitez créer à partir des données. Sélectionnez également K points comme centres de cluster initiaux.

b. Attribution des points aux clusters : Pour chaque point de données, calculez sa distance par rapport à chaque centre de cluster. Assignez le point au cluster dont le centre est le plus proche, En utilisant l'une des métriques de distance (la distance euclidienne …)

c. Mise à jour des centres des clusters : Une fois que tous les points ont été attribués à des clusters, recalculez les centres de chaque cluster en prenant la moyenne des points attribués à ce cluster.

d. Répétez les étapes b et c jusqu'à ce qu'un critère d'arrêt soit atteint (la stabilisation des centres)

e. Résultat final : Les centres des clusters représentent les "centres" des groupes formés par l'algorithme K-Means, et les points de données sont attribués aux clusters correspondants.

2) Inconvénients de la méthode K-Means :

a. Dépendance à l'initialisation : Une mauvaise initialisation peut entraîner une convergence vers des solutions sous-optimales.

b. Nombre de clusters (K) : Le choix du nombre optimal de clusters (K) n'est pas trivial et peut nécessiter des essais itératifs.

c. Sensible aux valeurs aberrantes : K-Means est sensible aux valeurs aberrantes (outliers) dans les données, car elles peuvent influencer fortement la position des centres de cluster.

e. Convergence vers un minimum local : L'algorithme K-Means peut converger vers un minimum local, ce qui signifie qu'il peut ne pas trouver la meilleure solution globale.

Pour atténuer ce problème, il est courant d'exécuter l'algorithme plusieurs fois avec différentes initialisations et de choisir la meilleure solution parmi celles obtenues.

3) Taux de bonne classification pour Classe 0 = (80*100)/87=91.95%

Taux de bonne classification pour Classe 1 = 52*100/56=92.85%

Taux de Bonne classification Totale = (80 +52)*100/(87+56) = 92.4%

4) Un taux de bonne classification totale de 92,4% est considéré comme élevé et indique en général une bonne performance du modèle.

Exercice 2:

1) les patientes atteintes de fractures de la hanche (FH) sont dans la classe 1 (Groupe 1),
tandis que les patientes atteintes de tassements vertébraux (TV) sont dans la classe 2 (Groupe 2).
Taux de bonne classification totale:(54+72)/(58+79)*100= **91.97%**

2) La CAH n'est pas basée sur un choix aléatoire, contrairement à K-means. Elle permet d'obtenir toutes les classes possibles à partir du dendogramme

3) Taux de bonne class totale:(56+70)/(58+79)*100= **91.97%**
On peut appliquer ces deux algorithmes, K-means et CAH, puisqu'ils donnent le même taux de bonne classification.

Exercice 3:

Personnalisation des services : En identifiant des segments homogènes de clients, une banque peut personnaliser ses offres de produits et de services en fonction des besoins spécifiques de chaque segment.
Marketing ciblé : Une fois que les segments de clients sont identifiés, la banque peut cibler ses campagnes de marketing de manière plus précise. Cela permet d'optimiser les ressources marketing en s'adressant aux segments qui sont les plus susceptibles de répondre favorablement aux offres .
Gestion des risques : En regroupant les clients présentant des caractéristiques similaires, la banque peut mieux évaluer et gérer les risques. Par exemple, elle peut identifier plus facilement les segments de clients présentant un risque de défaut de paiement élevé et prendre des mesures pour atténuer ce risque.

2)La stabilisation de l'inertie totale signifie que l'algorithme a convergé vers une solution où les groupes obtenus ne changent pas de manière significative lors des itérations suivantes. En d'autres termes, les centres des clusters restent relativement inchangés, et les points de données ne sont pas réaffectés fréquemment entre les clusters.

Lorsque l'inertie totale se stabilise, cela signifie que les groupes résultants sont devenus relativement homogènes, car les points de données sont plus proches de leur propre centre de cluster que des centres des autres clusters.

3) les résultats de K-means sont sensibles à l'initialisation aléatoires des centres.
Le nombre de clusters (k) doit être défini à l'avance dans K-means
4) Taux de bonne classification totale …

Exercice 4:

1) Calcul des distances d(S1, S2) et d(S1, S3) :
Pour d(S1, S2) :
nA = 10 (nombre d'espèces sur S1)
nB = 15 (nombre d'espèces sur S2)
nAB = 5 (nombre d'espèces communes à S1 et S2)
d(S1, S2) = (10 + 15 - 2 * 5) / (10 + 15) = (25 - 10) / 25 = 15 / 25 = 3/5 =0,6
Pour d(S1, S3) :
nA = 10 (nombre d'espèces sur S1)
nB = 11 (nombre d'espèces sur S3)
nAB = 8 (nombre d'espèces communes à S1 et S3)
d(S1, S3) = (10 + 11 - 2 * 8) / (10 + 11) = (21 - 16) / 21 = 5 / 21 =0,23

2) La distance entre deux sites ayant exactement les mêmes espèces est de 0.
Cela signifie qu'ils partagent toutes les espèces et sont donc identiques du point de vue de la présence d'espèces communes.
d(A,B) = nA+nA-2nA/2nA = **0**

3) La distance entre deux sites n'ayant aucune espèce en commun est de 1.
Cela signifie qu'ils sont complètement différents en termes de présence d'espèces communes.
d(A,B) = nA+nB/nA+nB = **1**

4) Etape 1: Calculer les distances pour les autres cas.

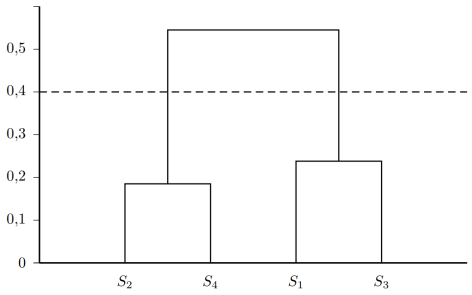
La matrice des distances est calculée en utilisant la formule d(A, B) pour chaque paire de sites (A, B) :

	S1	S2	S3	S4
S1	0	0,6	0,238	0,545
S2		0	0,692	0,185
S3			0	0,739
S4				0

5) Etape 2:Trouver la valeur minimale.

Etape 3:Dessiner le palier dans le dendogramme.

Etape 4:Agréger les points (dans ce cas, on choisit la valeur minimale, puisque nous n'avons pas les coordonnées de chaque point). Dans d'autres cas, on calcule la moyenne.



	S1	S2	S3	**S4**
S1	0	0,6	0,238	0,545
S2		0	0,692	**0,185**
S3			0	0,739
S4				0

	S1	S2S4	**S3**
S1	0	0,545	**0,238**
S2S4		0	0,692
S3			0

	S1S3	**S2S4**
S1S3	0	**0,545**
S2S4		0

6) On coupe l'arbre là où les branches sont longues

7) {S2, S4} et {S1, S3}.