Chapitre 4 : Clustering

1. Introduction au Clustering

Le clustering, également connu sous le nom de regroupement ou partitionnement de données, est une technique d'apprentissage non supervisé largement utilisée dans le domaine de l'exploration de données. Contrairement à l'apprentissage supervisé, où les données d'entraînement sont étiquetées, le clustering vise à regrouper les données similaires en groupes ou clusters sans aucune étiquette préalable. C'est une méthode puissante pour découvrir des structures impliquées dans les données et pour révéler des informations cachées.

2. Importance et applications du clustering

Le clustering joue un rôle essentiel dans l'analyse de données, la recherche d'informations, la segmentation de la clientèle, la biologie, la détection d'anomalies, et bien plus encore. Voici quelques-unes de ses applications clés :

- Segmentation de Marché : Les entreprises utilisent le clustering pour identifier des segments de clients similaires, ce qui permet une personnalisation efficace du marketing et des offres.

- Biologie : Dans la génomique et la protéomique, le clustering est utilisé pour regrouper des séquences d'ADN similaires ou des protéines ayant des fonctions similaires.

- Traitement de l'image : Le clustering peut être utilisé pour la segmentation d'images, où les régions d'intérêt similaires sont identifiées.

- Détection d'Anomalies : En identifiant des groupes cohérents de données, il est plus facile de repérer les anomalies qui ne suivent pas les motifs attendus.

- Recherche d'Information : Le clustering peut être utilisé pour organiser des documents en groupes similaires, ce qui facilite la recherche et l'exploration.

3. Évaluation des Clusters

L'évaluation des clusters est essentielle pour déterminer la qualité des regroupements obtenus à partir des méthodes de clustering. Les mesures d'évaluation peuvent être classées en indices internes et externes.

Indices Internes : Cohérence Intra-cluster, Séparation Inter-cluster

Cohérence Intra-cluster : Cette mesure évalue la similarité des points au sein d'un même cluster. Des valeurs plus faibles indiquant que les points dans un cluster sont plus similaires les uns aux autres.

Séparation Inter-cluster : Cette mesure évalue la distance entre les clusters. Des valeurs plus élevées indiquent que les clusters sont bien séparés.

Autres mesures pour évaluer la qualité d'un cluster : Indice de Silhouette, Méthodes de validation interne, Visualisation, Connaissance de domaine, Stabilité des clusters, etc.

Il est à noter qu'il n'y a pas de mesure unique de la qualité des clusters qui fonctionnent pour tous les cas. Selon le contexte, une combinaison de plusieurs de ces mesures peut être nécessaire pour obtenir une évaluation complète et précise de la qualité d'un cluster.

4. Méthodes de Clustering

4.1. Clustering K-means

Principe de Base : Assignation Itérative et Optimisation

Le K-means est l'une des méthodes de clustering les plus utilisées. Son principe fondamental est de regrouper les données en calculant les centroïdes (moyennes) des clusters et en affectant chaque point de données au cluster dont le centroïde est le plus proche. L'algorithme vise à minimiser la somme des carrés des distances entre les points de données et les centroïdes de leurs clusters respectifs.

4.1.1. Algorithme K-means : Initialisation, Affectation, Mise à Jour des Centroïdes

- Initialisation : Choisissez aléatoirement K centroïdes, généralement en utilisant des points de données du jeu de données.

- Affectation : Pour chaque point de données, calculez la distance par rapport à tous les centroïdes et attribuez le point au cluster du centroïde le plus proche.

- Mise à Jour des Centroïdes : Recalculez les nouveaux centroïdes en prenant la moyenne des points de données dans chaque cluster.

- Répétez les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de changement d'affectation ou jusqu'à atteindre un nombre maximal d'itérations.

4.1.2. Choix du Nombre Optimal de Clusters

Le choix du nombre optimal de clusters (K) est crucial pour le succès du K-means. Plusieurs méthodes peuvent être utilisées, notamment :

- Méthode du Coude : Tracez la somme des carrés des distances intra-cluster en fonction du nombre de clusters. L'endroit où cette courbe commence à s'aplatir est le nombre optimal de clusters.

- Méthode de Silhouette : Calculez le coefficient de silhouette pour différents nombres de clusters et choisissez celui qui donne la valeur maximale. Le coefficient de silhouette mesure la similarité entre les points d'un cluster par rapport à d'autres clusters.

4.1.3. Avantages et Inconvénients du K-means

Avantages :

- Efficace pour les ensembles de données de grande taille.
- Facile à comprendre et à mettre en œuvre.
- Donne des clusters de forme convexe.

Inconvénients :

- Sensible aux valeurs aberrantes.
- Nécessité de connaître le nombre de clusters à l'avance.
- Donne des clusters de forme sphérique, ce qui peut ne pas être adapté à tous les types de données.

4.2. Algorithme K-medoid : Initialisation, Affectation, Mise à Jour des Medoids

4.2.1. Définition

Le K-medoid est une variante du K-means qui diffère dans la manière dont les centres de cluster sont choisis. Contrairement au K-means qui utilise la moyenne des points de données, le K-medoid choisit un point de données réel comme représentant, généralement le point qui minimise la somme des distances aux autres points du cluster.

4.2.2. les étapes de l'algorithme K-medoid :

Initialisation : Choisissez aléatoirement K medoids, généralement en utilisant des points de données du jeu de données.

Affectation : Pour chaque point de données, calculez la distance par rapport à tous les medoids et attribuez le point au cluster du medoid le plus proche.

Mise à Jour des Medoids : Recalculez les nouveaux medoids en choisissant le point qui minimise la somme des distances aux autres points de données dans le cluster.

Répétez les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de changement d'affectation ou jusqu'à atteindre un nombre maximal d'itérations.

Exercice : Clustering avec l'algorithme K-means

Supposons que nous ayons 8 points bidimensionnels (2D) que nous voulons regrouper en 3 clusters à l'aide de l'algorithme K-Means. Les points sont les suivants :

	A	B	C	D	E	F	G	H
X	3	4	6	8	9	1	1	5
Y	5	7	2	3	6	4	3	9

Utilisez l'algorithme K-Means pour initialiser les centres de cluster aléatoirement, attribuer les points aux clusters et répéter jusqu'à convergence (quand les affectations de cluster ne changent plus).

Solution :

- Choisissons aléatoirement 3 centres de cluster initiaux. Par exemple, nous pouvons choisir C1 = A (3, 5), C2 = C (6, 2) et C3 = G (1, 3) comme centres initiaux.

- Calculons la distance euclidienne entre les points et les centres initiaux, puis attributions des points aux clusters en fonction de la distance minimale.

	A (3, 5)	B (4, 7)	C (6, 2)	D (8, 3)	E (9, 6)	F (1, 4)	G (1, 3)	H (5, 9)
C1 = A	0,00	2,24	4,24	5,39	6,08	2,24	2,83	4,47
C2 = C	4,24	5,39	0,00	2,24	5,00	5,39	5,10	7,07
C3 = G	2,83	5,00	5,10	7,00	8,54	1,00	0,00	7,21

Grappe 1 : A, B, H.

Grappe 2 : C, D, E.

Grappe 3 : F, G.

- Mise à jour des centroïdes en utilisant les coordonnées moyennes des points appartenant à chaque cluster :

C1 : x = (3 + 4 + 5) / 3, y = (5 + 7 + 9)/3 → C1 : (4, 7)

C2 : x = (6 + 8 + 9) / 3, y = (2 + 3 + 6)/3 → C1 : (7,7, 3,7)

C3 : x = (1 + 1) / 2, y = (4 + 3)/2 → C1 : (1, 3,5)

- Répétition des étapes 2 et 3 jusqu'à convergence (aucun changement d'affectation).

	A (3, 5)	B (4, 7)	C (6, 2)	D (8, 3)	E (9, 6)	F (1, 4)	G (1, 3)	H (5, 9)
C1=(4, 7)	2,24	0,00	5,39	5,66	5,1	4,24	5	2,24
C2=(7.7, 3.7)	4,85	4,96	2,36	0,75	2,69	6,68	6,7	5,96
C3=(1, 3.5)	2,5	4,61	5,22	7,08	8,38	0,5	0,5	6,80

Grappe 1 : A, B, H.

Grappe 2 : C, D, E.

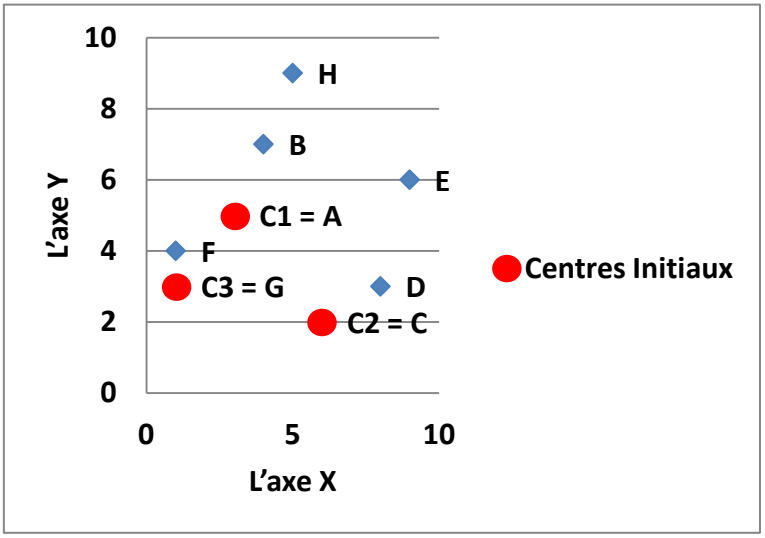
Grappe 3 : F, G.

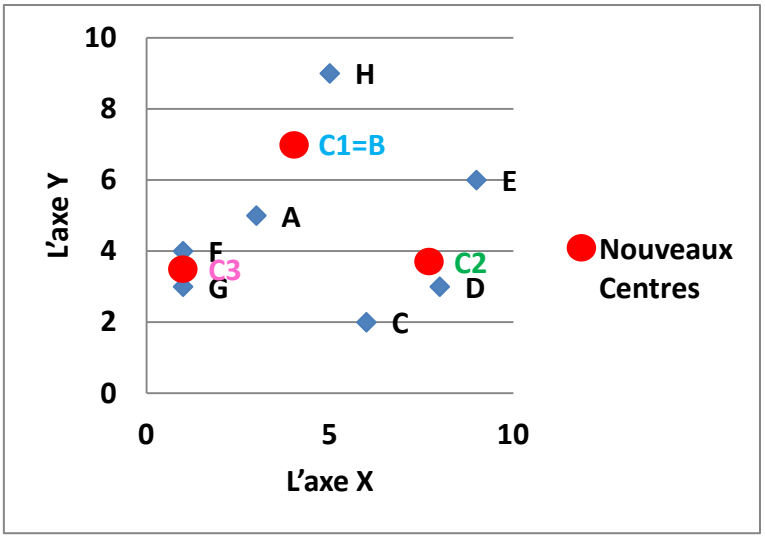
Nous remarquons qu'il n'y a aucun changement d'affectation des clusters, et les points sont regroupés en trois clusters C1, C2 et C3.

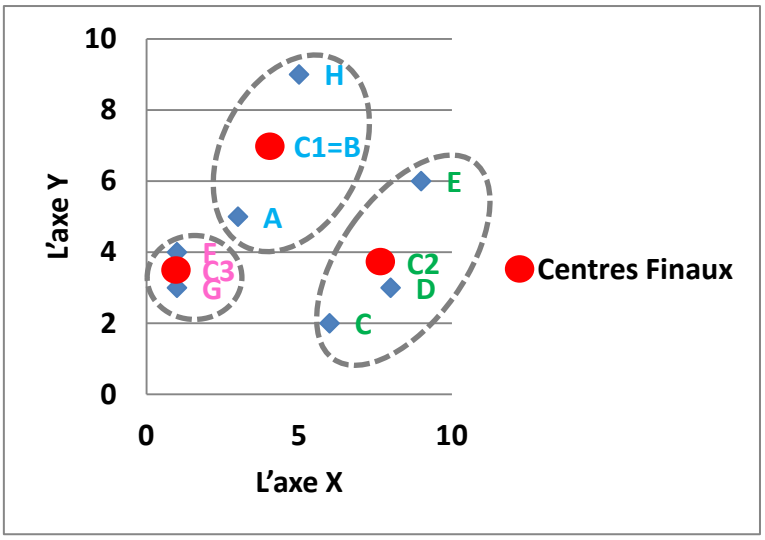
NB : Les valeurs initiales des centroïdes peuvent influencer les résultats du clustering. Dans cet exemple, les centroïdes initiaux ont été choisis de manière aléatoire pour démarrer le processus. En pratique, des méthodes d'initialisation peuvent être utilisées pour obtenir de meilleures solutions (exemple K-means++).

NB : La méthode du coude et la méthode de la silhouette sont utilisées pour déterminer le nombre optimal de clusters, tandis que la méthode K-means++ améliore l'initialisation des centroïdes pour l'algorithme K-means

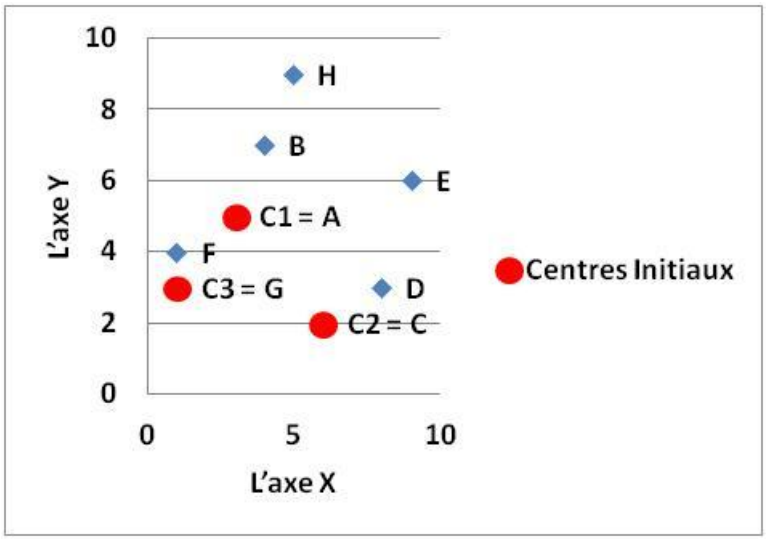
Illustrations :

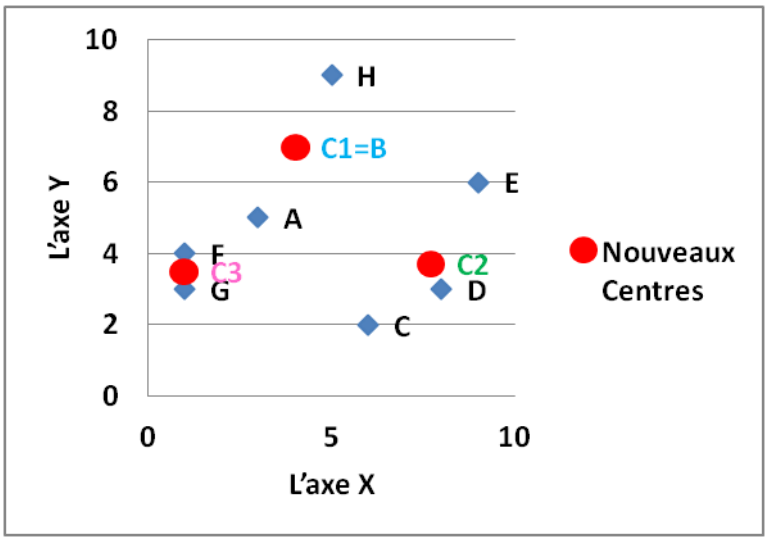


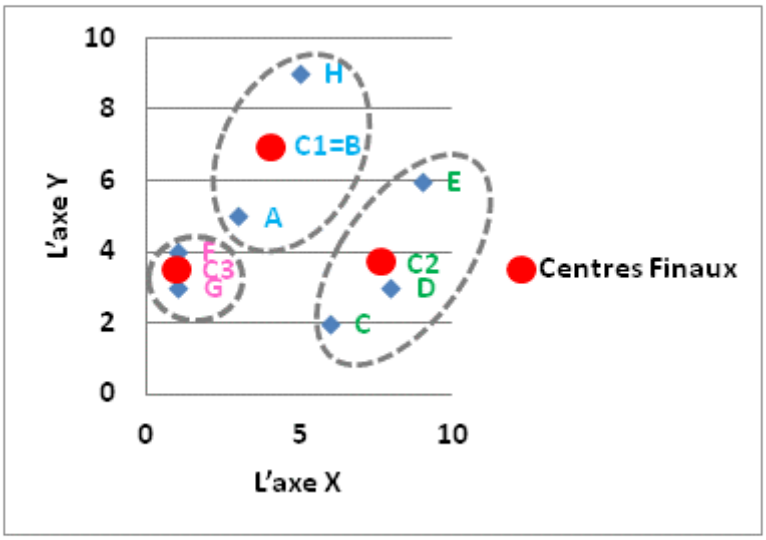




Illustrations :







4.3. Classification Hiérarchique

La classification hiérarchique est une approche de clustering qui construit une structure arborescente de clusters, appelée dendrogramme. Contrairement à la méthode K-means, la classification hiérarchique ne nécessite pas de préciser le nombre de clusters à l'avance. Elle utilise une matrice de distance comme critère de clustering. Elle peut être divisée en deux approches :

A. Classification ascendante hierarchique – CAH (Agglomerative): Cette approche commence par considérer chaque point de données comme un cluster individuel et fusionne progressivement les clusters similaires en un seul cluster à chaque étape.

B. Classification descendante hierarchique - CDH (Divisive) : Cette approche commence avec tous les points dans un seul cluster et les divise en progressivement sous-clusters plus petits.

4.3.1. Méthodes d'Agrégation

Lors de l'agglomération, des méthodes d'agrégation sont utilisées pour calculer la similarité entre les clusters. Voici quelques méthodes courantes :

A. Simple (Single Link) : Utilisez la distance la plus courte entre les paires de points des clusters.

$$D(i,j) = min_{x \in Ci, y \in Cj}\{d(x,y)\}$$

B. Complète (Complete Link) : Utiliser la distance la plus grande entre les paires de points des clusters.

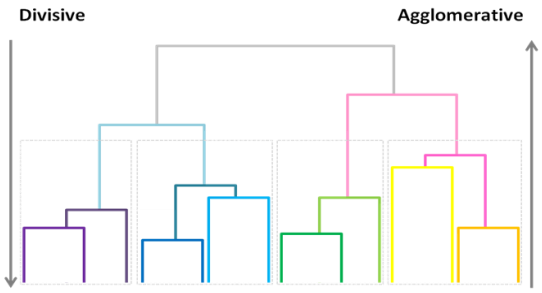
$$D(i,j) = max_{x \in Ci, y \in Cj}\{d(x,y)\}$$

C. Moyenne (Average Link) : Utiliser la moyenne des distances entre les paires de points des clusters.

$$D(i,j) = avg_{x \in Ci, y \in Cj}\{d(x,y)\}$$

4.3.2. Représentation du Dendrogramme

Un dendrogramme est un diagramme arborescent qui représente les étapes d'agglomération ou de division des clusters. Chaque nœud interne du dendrogramme représente un cluster formé par la fusion de deux sous-clusters.



4.3.3. Découpage du Dendrogramme en Clusters

Pour obtenir des clusters spécifiques à partir du dendrogramme, il suffit de couper les branches à une certaine hauteur. Plus la hauteur de coupe est élevée, moins de grappes seront formées.

4.3.4. Méthodes hiérarchiques

A. Méthode Hiérarchique Agglomérative (Ascendante) :

- Initialisation : Chaque point est considéré comme un cluster individuel. Au départ, le nombre de clusters est égal au nombre de points.

- Calcul des distances : Calculez les distances entre tous les clusters. Les distances peuvent être calculées de différentes manières, telles que la distance euclidienne.

- Fusion des clusters : Fusionnez les deux clusters les plus proches en un seul cluster.

- Mise à jour des distances : Les distances entre le nouveau cluster fusionné et les autres clusters sont recalculées, en utilisant la méthode spécifiée, comme la distance min, max, etc.

- Répétition : Répétez les étapes 3 et 4 jusqu'à ce qu'il ne reste qu'un seul cluster contenant tous les points, ce qui forme l'arbre hiérarchique complet.

B. Méthode Hiérarchique Divisive (Descendante) :

- Initialisation : Tous les points sont prévus comme faisant partie d'un seul grand cluster.

- Choix du cluster à diviser : Choisissez un cluster à diviser en sous-clusters. Cela peut se faire en utilisant différentes techniques, telles que la méthode de Ward, qui cherche à minimiser l'augmentation de la somme des carrés des distances à l'intérieur des sous-clusters.

- Division du cluster : Divisez le cluster sélectionné en deux sous-clusters en utilisant un algorithme de regroupement approprié.

- Mise à jour des distances : Recalculer les distances entre les sous-clusters nouvellement créés et les autres clusters.

- Répétition : Répétez les étapes 2 à 4 pour chaque cluster nouvellement créé jusqu'à ce que chaque point soit son propre cluster, ce qui forme l'arbre hiérarchique complet.

4.3.5. Avantages et Inconvénients de la Classification Hiérarchique

Avantages :

- Ne nécessite pas de préciser le nombre de clusters à l'avance.

- Donne une vue d'ensemble des relations entre les clusters.

- Convient aux données où la structure hiérarchique est pertinente.

Inconvénients :

- Peut être informatiquement coûteux pour de grandes données.

- Le choix de la méthode d'agrégation peut influencer les résultats.

- Une mauvaise compréhension du dendrogramme peut conduire à des interprétations erronées.

Exercice : Classification Hiérarchique agglomérative

Supposons que nous disposions des données suivantes représentant la taille (en centimètres) et le poids (en kilogrammes) de huit objets :

	A	B	C	D	E	F	G	H
Taille (Cm)	160	165	152	170	155	180	168	162
Poids (Kg)	50	58	43	68	46	80	70	55

Utilisez la méthode de classification hiérarchique agglomérative avec la distance euclidienne pour regrouper ces objets en trois clusters. Calculez les distances entre les clusters à chaque étape et dessinez le dendrogramme correspondant.

Solution

Nous calculons les distances euclidiennes entre tous les objets et commençons avec chaque objet comme un cluster individuel.

	A	B	C	D	E	F	G	H
A	0							
B	9,43	0						
C	10,63	19,85	0					
D	20,59	11,18	30,81	0				
E	6,40	15,62	4,24	26,63	0			
F	36,06	26,63	46,40	15,62	42,20	0		
G	21,54	12,37	31,39	**2,83**	27,30	15,62	0	
H	5,39	4,24	15,62	15,26	11,40	30,81	16,15	0

Fusionnons les deux clusters les plus proches (D et G)

Répétons les calcules des distances a chaque fois entre le nouveau cluster fusionné et les autres clusters jusqu'à ce qu'il ne reste qu'un seul cluster contenant tous les points.

NB : La définition de "proche" dépend de la mesure de distance utilisée (ex. : distance minimale (single link), maximale (complete link), moyenne (average link), etc.).

NB : Prenons l'exemple de la méthode Single Link où la distance entre deux clusters est déterminée par la distance la plus courte entre les points des deux clusters.

Supposons que nous avons fusionné les clusters (D) et (G) dans un nouveau cluster (DG). Maintenant, pour calculer la distance entre ce nouveau cluster (DG) et un autre point ou

cluster, nous calculons la distance la plus courte entre n'importe quel point de (DG) et n'importe quel point du point ou du cluster en question.

Par exemple, si nous calculons la distance entre le nouveau cluster (DG) et le point (A), nous devons calculer la distance la plus courte entre tous les points de (DG) et le point (A). Ensuite, nous pouvons utiliser cette distance comme résultat de la distance entre (DG) et (A).

	A	B	C	D , G	E	F	H
A	0						
B	9,43	0					
C	10,63	19,85	0				
D, G	20,59	11,18	30,81	0			
E	6,40	15,62	**4,24**	26,63	0		
F	36,06	26,63	46,40	15,62	42,20	0	
H	5,39	**4,24**	15,62	15,26	11,40	30,81	0

Fusionnons les deux clusters les plus proches, soit (B et H) ou (C et E)

	A	B, H	C	D, G	E	F
A	0					
B, H	5,39	0				
C	10,63	15,62	0			
D, G	20,59	11,18	30,81	0		
E	6,40	11,40	**4,24**	26,63	0	
F	36,06	26,63	46,40	15,62	42,20	0

Fusionnons les deux clusters les plus proches (C et E)

	A	B, H	C, E	D, G	F
A	0				
B, H	**5,39**	0			
C, E	6,40	15,62	0		
D, G	20,59	11,18	26,63	0	
F	36,06	26,63	42,20	15,62	0

Fusionnons les deux clusters les plus proches (A et (BH))

	A, B, H	C, E	D, G	F
A, B, H	0			
C, E	**6,40**	0		
D, G	11,18	26,63	0	
F	26,63	42,20	15,62	0

Fusionnons les deux clusters les plus proches ((A, B, H) et (C, E))

	A, B, H, C, E	D, G	F
A, B, H, C, E	0		

D, G	**11,18**	0	
F	26,63	15,62	0

Fusionnons les deux clusters les plus proches ((A, B, H, C, E) et (D, G))

	A, B, H, C, E, D, G	F
A, B, H, C, E, D, G	0	
F	**15,62**	0

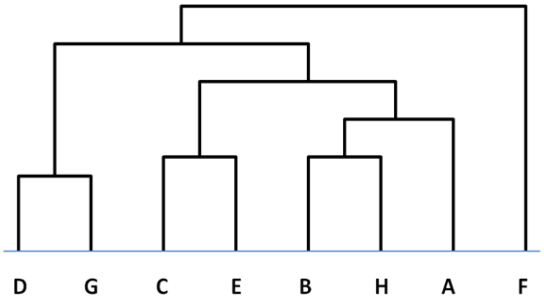
Fusionnons les deux clusters les plus proches ((A, B, H, C, E, D, G) et F)

Le dendrogramme (arbre hiérarchique) correspondant serait un diagramme qui montre les étapes de fusion des clusters et les distances associées.

Cluster 1 : {C, E, B, H, A}.

Cluster 2 : {D, G}.

Cluster 3 : {F}.



4.4. DBSCAN (Regroupement spatial basé sur la densité d'applications avec bruit)

DBSCAN est un algorithme de regroupement spatial qui se base sur la densité des points de données dans l'espace. Il peut identifier des groupes de points qui sont densément connectés, tout en étiquetant les points isolés comme du bruit.

Soient les définitions importantes suivantes :

- **Epsilon (ε) :** C'est le rayon défini autour de chaque point. Si au moins "MinPts" points sont présents dans ce rayon (compris le point central), alors ces points sont prévus comme étant dans la même région dense.

- **MinPts :** C'est le nombre minimum de points requis pour former une région dense. Si un point à au moins "MinPts" voisins dans son rayon ε, il est considéré comme un point de cœur.

- **Points de cœur :** Ce sont les points qui ont au moins "MinPts" points dans leur rayon ε.

- **Points frontières :** Ce sont les points qui ne sont pas des points de cœur-mêmes, mais qui sont situés eux dans le rayon ε d'un point de cœur.

- **Points de bruit :** Ce sont les points qui ne sont ni des points de cœur ni des points frontières.

4.4.1. Les étapes du DBSCAN :

- Choisissez un point de données non visité aléatoirement.

- Si ce point à au moins "MinPts" voisins dans son rayon ε, il forme un nouveau groupe (cluster).

- Ajoute tous les points accessibles (dans le rayon ε) à ce groupe. Cela signifie que si un point est un point de cœur, tous les points dans son rayon ε font partie du même groupe.

- Répétez le processus pour chaque point ajouté au groupe.

- Continuez jusqu'à ce que tous les points soient visités.

4.4.2. Avantages et Inconvénients de DBSCAN

Avantages de DBSCAN :

- Capable de détecter des formes de clusters complexes et non linéaires.

- Résistant au bruit et capable de gérer les points aberrants.

- N'exige pas de préciser le nombre de clusters à l'avance.

- Indépendant de l'ordre des données.

- Peut gérer des clusters de densités variables.

Inconvénients de DBSCAN :

- Sensible aux paramètres ε et MinPts.

- Peut-être avoir du mal à détecter des clusters dans des régions de densité uniforme.

- Dépend des métriques de distance.

- Complexité algorithmique supérieure à d'autres méthodes plus simples.

- Peut-être avoir du mal à gérer les clusters de tailles très différentes.

Exercice : DBScan

Données disponibles : A (2, 3), B (3, 5), C (4, 8), D (6, 9), E (8, 7), F (9, 6). Utilisez les paramètres ε = 3 et MinPts = 3, puis appliquez l'algorithme DBSCAN sur ces données.

Solution

Nous calculons d'abord la matrice des distances

Distance euclidienne :

	A (2, 3)	B (3, 5)	C (4, 8)	D (6, 9)	E (8, 7)	F (9, 6)
A (2, 3)	0					
B (3, 5)	2,24	0				
C (4, 8)	5,39	3,16	0			
D (6, 9)	7,21	5	2,24	0		
E (8, 7)	7,21	5,39	4,12	2,83	0	
F (9, 6)	7,62	6,08	5,39	4,24	1,41	0

- Commençons par choisir un point non visité, par exemple le point A.

- Le rayon ε contient le point B c'est tout, et la même chose pour le point B, le rayon epsilon contient seulement le point A.

- Nous remarquons que les deux points A et B sont proches l'un de l'autre mais ils ne satisfont pas les critères pour être des points de cœur ou des points frontières, ils seront toujours définis comme des points de bruit dans le contexte de DBSCAN.

- Passons au point C, il a un seul voisin D, ce qui ne satisfait pas MinPts.

- Le point D a 2 voisins (C et E), ce qui a satisfait MinPts, donc, C'est un point de cœur, et le point C est un point frontiére, puisque il est situé dans le rayon du point cœur D.

- Le point E a 2 voisins (D et F), ce qui a satisfait MinPts, donc, C'est un point de cœur.

- Le point F a un seul voisin, ce qui ne satisfait pas MinPts, mais il est situé dans le rayon du point cœur E, donc c'est un point frontiére.

Résultats :

Point	A	B	C	D	E	F
Voisinage	B	A	D	C, E	D, F	E

Groupe 1 : {C, D, E, F}

Points de cœur : {D, E}

Points frontières : {C, F}

Points de bruit : {A, B}

Conclusion

Dans ce chapitre, nous avons exploré diverses méthodes de clustering, y compris les K-means, la classification hiérarchique et DBSCAN. Chaque méthode a ses avantages et ses inconvénients, et le choix de la méthode dépend du problème et des données.

Le succès du clustering dépend en grande partie de la qualité des données et de la connaissance du domaine. L'évaluation des clusters est nécessaire pour valider les résultats et choisir la meilleure méthode.