Faculté des Nouvelles Technologies de l'Information et de la Communication

Département : Informatique Fondamentale et ses Applications

Année Universitaire : 2023/2024

Module : DAMI (Master 1 SDIA)

TD N° 4.1 : Clustering – Algorithme K-Means

❖ Exercice 1 :

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).

❖ Exercice 2 :

Considérez le jeu de données suivant dans un espace tridimensionnel :

A(1,2,3), B(2,3,4), C(5,8,6), D(8,3,2), E(10,6,8), F(12,8,5)

- Initialisez les centres des clusters en choisissant les trois points de données suivant :

C1 = (-1, 2, 3), C2 = (8, -3, 2) et C3 = (12, 8, -5)

- Appliquez deux itérations de l'algorithme K-means pour attribuer chaque point de données au cluster le plus proche et recalculez les centres des clusters.
- Répétez l'étape 2 jusqu'à convergence.
- Présentez les clusters finaux et les centres finaux.

Solution exercice 1 :

- 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 → C2 : (7,7, 3,7)

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

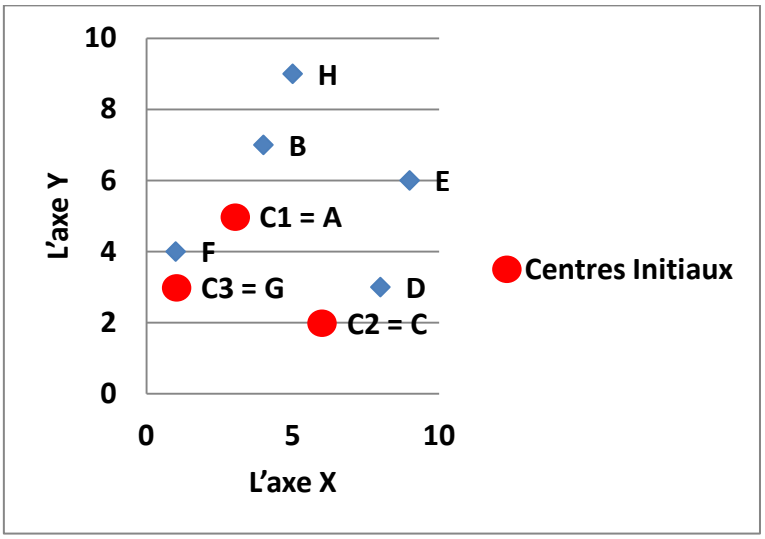
Grappe 1 : A, B, H.

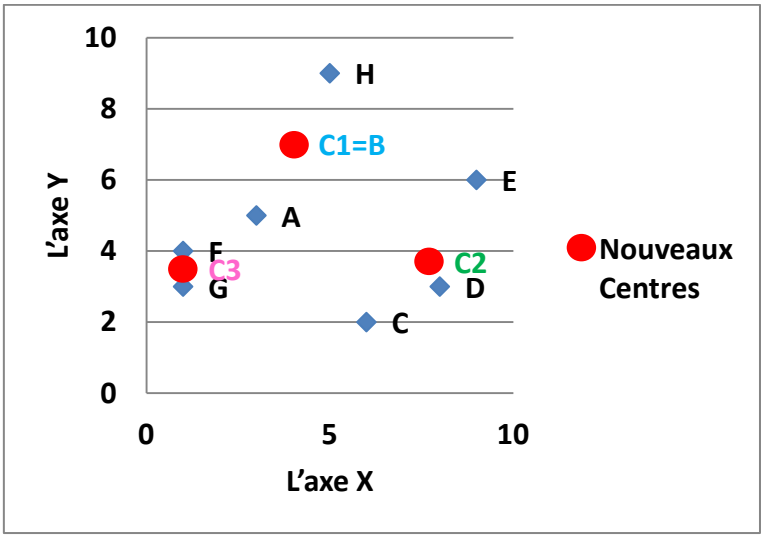
Grappe 2 : C, D, E.

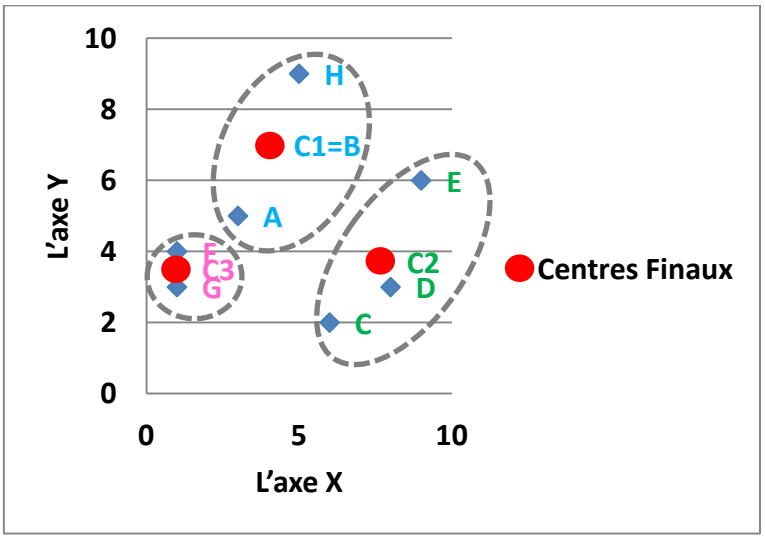
Grappe 3 : F, G.

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.

Illustrations :







Solution exercice 2 :

- Choisissons aléatoirement 3 centres de cluster initiaux. Par exemple, nous pouvons choisir C1 = (3, 5, 7), C2 = (6, 2, 0) et C3 = (1, 3, 5) 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(1,2,3)	B(2,3,4)	C(5,8,6)	D(8,3,2)	E(10,6,8)	F(12,8,5)
C1 = (-1, 2, 3)	2	3,31	9	9,11	12,72	14,45
C2 = (8, -3, 2)	8,66	8,71	12,08	6	11	12,08
C3 = (12, 8, -5)	14,86	14,35	13,03	9,48	13,30	10

Grappe 1 : A, B, C.

Grappe 2 : D, E.

Grappe 3 : F.

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

C1 : x = (1 + 2 + 5) / 3, y = (2 + 3 + 8) / 3, z = (3 + 4 + 6) / 3 → C1 : (2,66, 4,33, 4,33)

C2 : x = (8 + 10) / 2, y = (3 + 6) / 2, z = (2 + 8) / 2 → C2 : (9, 4,5, 5)

C3 : (12, 8, 5)

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

	A(1,2,3)	B(2,3,4)	C(5,8,6)	D(8,3,2)	E(10,6,8)	F(12,8,5)
C1 = (2,66, 4,33, 4,33)	3,15	1,52	4,66	5,97	8,37	10,05
C2 = (9, 4,5, 5)	8,61	7,22	5,40	3,50	3,50	4,60
C3 = (12, 8, 5)	12,68	11,22	7,07	7,07	4,12	0

Grappe 1 : A, B, C.

Grappe 2 : D, E.

Grappe 3 : F.

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.

- Les clusters finaux et les centres finaux :

Grappe 1 : A, B, C.

Grappe 2 : D, E.

Grappe 3 : F.

Pour :

C1 : (2,66, 4,33, 4,33).

C2 : (9, 4,5, 5).

C3 : (12, 8, 5).

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