Exercice 1: Clustering K-means sur un petit jeu de données

Objectifs :

- Il illustre la manipulation de données multidimensionnelles.
- Il montre le fonctionnement simple d'un algorithme clé en Big Data.
- Il peut être étendu à des données plus volumineuses avec outils comme Spark, Hadoop.

Contexte

Une entreprise analyse les habitudes d'achat de ses clients. Elle dispose de données sur deux variables :

- Montant moyen dépensé par client (en euros)
- Nombre moyen d'achats par mois

On veut segmenter les clients en 2 groupes homogènes pour adapter la stratégie marketing.

Données (9 clients) :

Client	Montant (€)	Achats/mois
C1	50	5
C2	52	6
C3	48	5
C4	200	20
C5	190	22
C6	210	21
C7	55	6
C8	205	23
C9	53	5

Questions

1. Choisir les centres initiaux comme C1 et C4.
2. Appliquer une itération de l'algorithme K-means :
    - Assigner chaque client au centre le plus proche (distance euclidienne)
    - Recalculer les centres des clusters
3. Donner les clusters obtenus après cette itération.

Solution

Données

Client	Montant (€)	Achats/mois
C1	50	5
C2	52	6
C3	48	5
C4	200	20
C5	190	22
C6	210	21
C7	55	6
C8	205	23

Client	Montant (€)	Achats/mois
C9	53	5

Étape 1 : Centres initiaux

- Centre 1 : C1 → (50, 5)
- Centre 2 : C4 → (200, 20)

Étape 2 : Calcul des distances euclidiennes et assignation

La distance euclidienne entre deux points (x1,y1) et (x2,y2) est :

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Calcul de la distance de chaque client aux deux centres :

Client	Dist. à C1(50,5)	Dist. à C4(200,20)	Cluster assigné
C1(50,5)	0	$\sqrt{(50-200)^2+(5-20)^2} = \sqrt{150^2+15^2} = \sqrt{22500+225} = \sqrt{22725} \approx 150.75$	C1
C2(52,6)	$\sqrt{(52-50)^2+(6-5)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.24$	$\sqrt{(52-200)^2+(6-20)^2} = \sqrt{148^2+14^2} = \sqrt{21904+196} = \sqrt{22100} \approx 148.66$	C1
C3(48,5)	$\sqrt{(48-50)^2+(5-5)^2} = \sqrt{4+0} = 2$	≈ 152.49	C1
C4(200,20)	150.75	0	C4

Étape 3 : Clusters après assignation

- Cluster 1 (centre C1) : C1, C2, C3, C7, C9
- Cluster 2 (centre C4) : C4, C5, C6, C8

Étape 4 : Recalcul des centres (moyenne des points dans chaque cluster)

- Pour Cluster 1 :

Montant	Achats
50	5
52	6
48	5
55	6
53	5

Moyennes :

$$\text{Montant} = \frac{50 + 52 + 48 + 55 + 53}{5} = \frac{258}{5} = 51.6$$

$$\text{Achats} = \frac{5 + 6 + 5 + 6 + 5}{5} = \frac{27}{5} = 5.4$$

Nouveau centre 1 = (51.6, 5.4)

- Pour Cluster 2 :

Montant	Achats
200	20
190	22
210	21
205	23

Moyennes :

$$\text{Montant} = \frac{200 + 190 + 210 + 205}{4} = \frac{805}{4} = 201.25$$

$$\text{Achats} = \frac{20 + 22 + 21 + 23}{4} = \frac{86}{4} = 21.5$$

Nouveau centre 2 = (201.25, 21.5)

Résumé après 1 itération

- Centres :
  - C1 = (51.6, 5.4)
  - C2 = (201.25, 21.5)
- Clusters :
  - Cluster 1 : C1, C2, C3, C7, C9
  - Cluster 2 : C4, C5, C6, C8

Étape 3 : 2ème itération — réassignation des points aux nouveaux centres

On calcule les distances euclidiennes de chaque client aux nouveaux centres et on réassigne.

Calcul des distances

Client	Coordonnées	Dist. à C1(51.6,5.4)	Dist. à C2(201.25,21.5)	Cluster assigné
C1	(50, 5)	$\sqrt{(50-51.6)^2+(5-5.4)^2} = \sqrt{2.56+0.16} = \sqrt{2.72} \approx 1.65$	≈ 150.54	C1
C2	(52, 6)	$\sqrt{(52-51.6)^2+(6-5.4)^2} = \sqrt{0.16+0.36} = \sqrt{0.52} \approx 0.72$	≈ 148.59	C1
C3	(48, 5)	$\sqrt{(48-51.6)^2+(5-5.4)^2} = \sqrt{12.96+0.16} = \sqrt{13.12} \approx 3.62$	≈ 153.21	C1
C4	(200, 20)	≈ 150.54	$\sqrt{(200-201.25)^2+(20-21.5)^2} = \sqrt{1.56+2.25} = \sqrt{3.81} \approx 1.95$	C2
C5	(190, 22)	$\sqrt{(190-51.6)^2+(22-5.4)^2} =$	$\sqrt{(190-201.25)^2+(22-21.5)^2} =$	C2

C6	(210, 21)	≈ 158.73	$\sqrt{(210-201.25)^2+(21-21.5)^2} = \sqrt{76.56+0.25} = \sqrt{76.81} \approx 8.76$	C2
C7	(55, 6)	$\sqrt{(55-51.6)^2+(6-5.4)^2} = \sqrt{11.56+0.36} = \sqrt{11.92} \approx 3.45$	≈ 146.28	C1
C8	(205, 23)	≈ 157.17	$\sqrt{(205-201.25)^2+(23-21.5)^2} = \sqrt{14.06+2.25} = \sqrt{16.31} \approx 4.04$	C2
C9	(53, 5)	$\sqrt{(53-51.6)^2+(5-5.4)^2} = \sqrt{1.96+0.16} = \sqrt{2.12} \approx 1.46$	≈ 148.45	C1

Nouvelle assignation des clusters

- Cluster 1 : C1, C2, C3, C7, C9
- Cluster 2 : C4, C5, C6, C8

Remarque : les clusters n'ont pas changé par rapport à la première itération.

Étape 4 : Recalcul des centres

Calcul des moyennes des points dans chaque cluster :

- Cluster 1:

$$\text{Montant} = \frac{50+52+48+55+53}{5} = 51.6$$

$$\text{Achats} = \frac{5+6+5+6+5}{5} = 5.4$$

Cluster 2:

$$\text{Montant} = \frac{200+190+210+205}{4} = 201.25$$

$$\text{Achats} = \frac{20+22+21+23}{4} = 21.5$$

Les centres restent donc identiques à ceux de l'itération précédente.

Conclusion

- **Les centres ne changent plus, donc l'algorithme converge.**
- Clusters finaux :

Cluster 1 (C1 = (51.6, 5.4))	**Cluster 2 (C2 = (201.25, 21.5))**
C1 (50, 5)	C4 (200, 20)
C2 (52, 6)	C5 (190, 22)
C3 (48, 5)	C6 (210, 21)

Cluster 1 (C1 = (51.6, 5.4))	Cluster 2 (C2 = (201.25, 21.5))
C7 (55, 6)	C8 (205, 23)
C9 (53, 5)	

Résumé final

- L'algorithme a convergé en **2 itérations**.
- Il a identifié deux groupes homogènes de clients :
  - **Cluster 1** : clients avec faible montant et faible fréquence d'achat
  - **Cluster 2** : clients avec montant élevé et fréquence élevée

Exercice 2 : Algorithme clé en Big Data : Filtrage collaboratif par similarité cosinus

Résumé

- Filtrage collaboratif est une méthode clé en Big Data pour systèmes de recommandation.
- La similarité cosinus est une mesure simple pour comparer profils utilisateurs.
- Le système recommande des items que des utilisateurs similaires ont appréciés.

Contexte

Une plateforme de streaming veut recommander des films à ses utilisateurs en se basant sur les notes données par d'autres utilisateurs.

Exemple simple

Données (notes sur 5 films par 4 utilisateurs)

Utilisateur	Film 1	Film 2	Film 3	Film 4	Film 5
U1	5	3	0	1	0
U2	4	0	0	1	0
U3	1	1	0	5	4
U4	0	0	5	4	5

0 signifie que l'utilisateur n'a pas noté ce film.

Objectif

Pour recommander un film à l'utilisateur U1, on va :

1. Calculer la similarité entre U1 et les autres utilisateurs (U2, U3, U4) avec la **similarité cosinus**.
2. Trouver l'utilisateur le plus similaire.

3. Recommander à U1 un film que cet utilisateur similaire a bien noté et que U1 n'a pas encore vu.

Étape 1 : Similarité cosinus

La similarité cosinus entre deux vecteurs A et B est :

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Étape 2 : Calcul des similarités entre U1 et les autres

On considère les vecteurs des notes de chaque utilisateur :

- U1 = [5, 3, 0, 1, 0]
- U2 = [4, 0, 0, 1, 0]
- U3 = [1, 1, 0, 5, 4]
- U4 = [0, 0, 5, 4, 5]

Calculons la similarité entre U1 et U2 :

- Produit scalaire $U1 \cdot U2 = 5*4+3*0+0*0+1*1+0*0 = 20+0+0+1+0 = 21$
- Normes :

- $\|U1\| = \sqrt{5^2 + 3^2 + 0 + 1^2 + 0} = \sqrt{25 + 9 + 0 + 1 + 0} = \sqrt{35} \approx 5.92$
- $\|U2\| = \sqrt{4^2 + 0 + 0 + 1^2 + 0} = \sqrt{16 + 0 + 0 + 1 + 0} = \sqrt{17} \approx 4.12$

- Similarité :

$$\frac{21}{5.92 \times 4.12} = \frac{21}{24.39} \approx 0.86$$

Calculons la similarité entre U1 et U3 :

- Produit scalaire $U1 \cdot U3 = 5*1+3*1+0*0+1*5+0*4 = 5+3+0+5+0 = 13$
- Normes :

- $\|U3\| = \sqrt{1^2 + 1^2 + 0 + 5^2 + 4^2} = \sqrt{1 + 1 + 0 + 25 + 16} = \sqrt{43} \approx 6.56$

- Similarité :

$$\frac{13}{5.92 \times 6.56} = \frac{13}{38.83} \approx 0.33$$

Calculons la similarité entre U1 et U4 :

- Produit scalaire $U1 \cdot U4 = 5*0+3*0+0*5+1*4+0*5 = 0+0+0+4+0 = 4$

- Normes :
- $\|U4\| = \sqrt{0^2 + 0^2 + 5^2 + 4^2 + 5^2} = \sqrt{0 + 0 + 25 + 16 + 25} = \sqrt{66} \approx 8.12$
- Similarité :

$$\frac{4}{5.92 \times 8.12} = \frac{4}{48.06} \approx 0.08$$

Étape 3 : Choix de l'utilisateur le plus similaire

- U2 : 0.86
- U3 : 0.33
- U4 : 0.08

L'utilisateur le plus similaire à U1 est donc **U2**.

Étape 4 : Recommandation

- Films que U2 a notés et U1 n'a pas regardés (note 0) :
  - Film 2 : U1 a noté 3 (donc vu)
  - Film 5 : U1 a noté 0 (pas vu), U2 a noté 0 (pas vu) → pas utile
  - Film 3 : U1=0, U2=0 → pas vu par les deux
- Films non vus par U1 mais bien notés par U2 : aucun.

On regarde aussi les notes de U2 pour films non vus par U1 : les films 3 et 5 ont 0, donc pas recommandables.

On peut élargir la recommandation au 2e utilisateur similaire (U3)

- Films non vus par U1 mais vus par U3 :
  - Film 5 : U1=0, U3=4 (bien noté)

Conclusion

Recommander **Film 5** à U1 (grâce à similarité avec U3).

Exercice 3: Algorithme clé en Big Data : PageRank

Résumé

- PageRank est un algorithme de ranking très utilisé en Big Data pour mesurer l’importance relative dans un graphe.
- Il est calculé itérativement jusqu’à convergence.
- Utilisé dans moteurs de recherche, analyse de réseaux sociaux, etc.

Contexte

PageRank est un algorithme d'analyse de graphes qui attribue un score d'importance à chaque nœud (page web) dans un graphe orienté (liens entre pages). Ce score reflète la "popularité" ou l'influence d'une page.

Principe de base

- Chaque page envoie son "score" à toutes les pages qu'elle référence (ses liens sortants).
- Le score d'une page est la somme des parts reçues des pages qui pointent vers elle.
- Un facteur d'amortissement (damping factor) modélise le comportement aléatoire d'un utilisateur qui saute sur une page au hasard.

Formule (simplifiée)

Pour une page Pi

$$PR(P_i) = \frac{1-d}{N} + d \times \sum_{P_j \in M_i} \frac{PR(P_j)}{L(P_j)}$$

- $PR(P_i)$: score PageRank de P_i
- d : facteur d'amortissement (classiquement 0.85)
- N : nombre total de pages
- M_i : ensemble des pages qui pointent vers P_i
- $L(P_j)$: nombre de liens sortants de la page P_j

Exemple simple

Supposons un graphe de 4 pages : A, B, C, D
Les liens sont :

- A → B, C
- B → C
- C → A
- D → C

Étape 1 : Initialisation

- Nombre total de pages N=4
- Initialiser PR de chaque page à 1/N=0.25

- Choisir d=0.85

Étape 2 : Calcul de PR pour chaque page

Par exemple, pour PR(A) :

- Pages pointant vers A : C
- L(C)=1 (C → A)

$$PR(A) = \frac{1 - 0.85}{4} + 0.85 \times \frac{PR(C)}{1} = 0.0375 + 0.85 \times PR(C)$$

De même pour B :

- Pages pointant vers B : A
- L(A)=2 (A → B, C)

$$PR(B) = 0.0375 + 0.85 \times \frac{PR(A)}{2}$$

Pour C :

- Pages pointant vers C : A, B, D
- L(A)=2, L(B)=1, L(D)=1

$$PR(C) = 0.0375 + 0.85 \times \quad \frac{PR(A)}{2} + PR(B) + PR(D)$$

Pour D :

- Pages pointant vers D : aucune

$$PR(D) = 0.0375 + 0$$

Étape 3 : Itération 1 (avec PR=0.25 initial)

Calculons PR valeurs à l'itération 1 :

- $PR(A) = 0.0375 + 0.85 \times 0.25 = 0.0375 + 0.2125 = 0.25$
- $PR(B) = 0.0375 + 0.85 \times \frac{0.25}{2} = 0.0375 + 0.10625 = 0.14375$
- $PR(C) = 0.0375 + 0.85 \times \quad \frac{0.25}{2} + 0.25 + 0.25$

  $= 0.0375 + 0.85 \times (0.125 + 0.25 + 0.25) = 0.0375 + 0.85 \times 0.625 = 0.0375 + 0.53125 = 0.56875$

- $PR(D) = 0.0375$

Étape 4 : Itération 2

Nouvelle itération avec ces valeurs :

- $PR(A) = 0.0375 + 0.85 \times PR(C) = 0.0375 + 0.85 \times 0.56875 = 0.0375 + 0.48344 = 0.52094$
- $PR(B) = 0.0375 + 0.85 \times \frac{PR(A)}{2} = 0.0375 + 0.85 \times 0.26047 = 0.0375 + 0.22140 = 0.2589$
- $PR(C) = 0.0375 + 0.85 \times \frac{PR(A)}{2} + PR(B) + PR(D)$

$= 0.0375 + 0.85 \times (0.26047 + 0.14375 + 0.0375) = 0.0375 + 0.85 \times 0.44172 = 0.0375 + 0.37546 = 0.41296$

- $PR(D) = 0.0375$

Étape 5 : Itération 3

Répétez jusqu'à convergence (les valeurs se stabilisent).

Page A : PageRank ≈ 0.38
Page B : PageRank ≈ 0.17
Page C : PageRank ≈ 0.38
Page D : PageRank ≈ 0.07