TD K-Means Clustering

Exemple Résolu : K-Means Clustering

Énoncé :

- La tâche de data mining consiste à regrouper des points en trois clusters.
- Les points sont :

$$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9).$$

- La fonction de distance utilisée est la **distance Euclidienne**.
- Initialement, nous assignons A_1, B_1 et C_1 comme centres de chaque cluster respectivement.

Solution

Centres Initiaux :

- $A_1 : (2,10)$
- $B_1 : (5,8)$
- $C_1 : (1,2)$

Calcul des Distances :

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Les résultats sont résumés dans le tableau suivant :

Points de Données	Distance à (2,10)	Distance à (5,8)	Distance à (1,2)	Cluster
$A_1(2,10)$	0.00	3.61	8.06	1
$A_2(2,5)$	5.00	4.24	3.16	3
$A_3(8,4)$	8.49	5.00	7.28	2
$B_1(5,8)$	3.61	0.00	7.21	2
$B_2(7,5)$	7.07	3.61	6.71	2
$B_3(6,4)$	7.21	4.12	5.39	2
$C_1(1,2)$	8.06	7.21	0.00	3
$C_2(4,9)$	2.24	1.41	7.62	1

Nouveaux Centres :

- A_1 : $(2, 10)$
- B_1 : $(6, 6)$
- C_1 : $(1.5, 3.5)$

Questions :

1. Recalculez les distances entre chaque point et les centres initiaux.
2. Attribuez chaque point au cluster correspondant.
3. Déterminez les nouveaux centres après une itération.

Exercice 2 : Ajout d'un nouveau point

Ajoutez le point de données $D_1(3, 6)$ aux points suivants :

- $A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$,
- $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$,
- $C_1(1, 2)$, $C_2(4, 9)$.

Questions :

1. Calculez les distances entre chaque point (y compris D_1) et les centres initiaux : $A_1(2, 10)$, $B_1(5, 8)$, $C_1(1, 2)$.
2. Attribuez chaque point au cluster correspondant.
3. Recalculez les nouveaux centres après une itération.

Exercice 3 : Distance de Manhattan

Reprenez les points suivants :

- $A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$,
- $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$,
- $C_1(1, 2)$, $C_2(4, 9)$.

Utilisez la distance de Manhattan :

$$d(p_1, p_2) = |x_2 - x_1| + |y_2 - y_1|.$$

Questions :

1. Calculez les distances entre chaque point et les centres initiaux : $A_1(2, 10)$, $B_1(5, 8)$, $C_1(1, 2)$.
2. Attribuez chaque point au cluster correspondant.
3. Déterminez les nouveaux centres après une itération.

Exercice 4 : Analyse de données médicales

Les points représentent des patients avec leurs caractéristiques mesurées :

- $P_1(60, 120)$: âge (en années) et pression artérielle systolique (en mmHg),
- $P_2(45, 130)$: âge (en années) et pression artérielle systolique (en mmHg),
- $P_3(50, 110)$: âge (en années) et pression artérielle systolique (en mmHg),
- $P_4(30, 100)$: âge (en années) et pression artérielle systolique (en mmHg),
- $P_5(25, 90)$: âge (en années) et pression artérielle systolique (en mmHg),
- $P_6(65, 140)$: âge (en années) et pression artérielle systolique (en mmHg).

Les centres initiaux sont :

- $C_1(60, 120)$,
- $C_2(30, 100)$.

Questions :

1. Calculez les distances euclidiennes entre chaque patient et les centres initiaux.
2. Assignez chaque patient à un cluster.
3. Déterminez les nouveaux centres après une itération.
4. Interprétez les clusters en termes médicaux (ex. : groupe à haut risque, groupe à faible risque).

Exercice 5 : Analyse des ventes de produits

Les points représentent des produits avec leurs ventes mensuelles (en milliers d'euros) et leur taux de retour client (en pourcentage) :

- $P_1(50, 5)$: Produit 1,
- $P_2(60, 10)$: Produit 2,
- $P_3(40, 8)$: Produit 3,
- $P_4(70, 12)$: Produit 4,
- $P_5(30, 6)$: Produit 5,
- $P_6(80, 15)$: Produit 6.

Les centres initiaux sont :

- $C_1(50, 5)$,
- $C_2(70, 12)$.

Questions :

1. Calculez les distances euclidiennes entre chaque produit et les centres initiaux.

2. Assignez chaque produit au cluster le plus proche.
3. Recalculez les centres après une itération.
4. Interprétez les clusters obtenus : quels produits nécessitent des ajustements dans leur stratégie de vente?

Exercice 6 : Classification des clients

Les points représentent des clients avec leur montant total dépensé (en milliers d'euros) et leur fréquence d'achat (en visites par mois) :

- $C_1(10, 2)$,
- $C_2(15, 5)$,
- $C_3(8, 3)$,
- $C_4(20, 7)$,
- $C_5(12, 4)$,
- $C_6(25, 6)$.

Les centres initiaux sont :

- $M_1(10, 2)$,
- $M_2(20, 7)$.

Questions :

1. Calculez les distances euclidiennes entre chaque client et les centres initiaux.
2. Attribuez chaque client à un cluster.
3. Déterminez les nouveaux centres après une itération.
4. Expliquez ce que représentent les clusters : quels groupes de clients nécessitent des offres promotionnelles ou des programmes de fidélité?

Exercice 7 : Optimisation des prix dans un magasin

Les points représentent des produits avec leur prix unitaire (en euros) et leur popularité (nombre de ventes par mois) :

- $P_1(20, 50)$,
- $P_2(15, 40)$,
- $P_3(30, 60)$,
- $P_4(25, 70)$,
- $P_5(10, 30)$,
- $P_6(35, 80)$.

Les centres initiaux sont :

- $C_1(20, 50)$,
- $C_2(30, 60)$.

Questions :

1. Calculez les distances entre chaque produit et les centres initiaux.
2. Attribuez chaque produit à un cluster.
3. Recalculez les centres après une itération.
4. Analysez les clusters : quels produits nécessitent des ajustements de prix pour maximiser les ventes?

Exercice 7 : Analyse des performances d'équipes sportives avec K-means

Les données suivantes représentent les performances de 10 équipes sportives sur une saison. Chaque équipe est caractérisée par :

- x_1 : Nombre de victoires,
- x_2 : Nombre de défaites,
- x_3 : Nombre de matchs nuls,
- x_4 : Nombre total de buts marqués,
- x_5 : Nombre total de buts encaissés.

Les données des équipes sont les suivantes :

$$\begin{array}{ll} \text{Équipe 1} : (20, 5, 3, 50, 20) & \text{Équipe 2} : (18, 6, 4, 45, 25) \\ \text{Équipe 3} : (22, 3, 3, 55, 18) & \text{Équipe 4} : (10, 12, 6, 30, 40) \\ \text{Équipe 5} : (15, 8, 5, 40, 30) & \text{Équipe 6} : (5, 18, 5, 20, 50) \\ \text{Équipe 7} : (8, 15, 5, 25, 45) & \text{Équipe 8} : (25, 3, 0, 60, 15) \\ \text{Équipe 9} : (12, 10, 6, 35, 35) & \text{Équipe 10} : (16, 7, 5, 42, 28) \end{array}$$

Les centres initiaux sont choisis arbitrairement comme suit :

$$C_1(20, 5, 3, 50, 20) \quad C_2(10, 12, 6, 30, 40) \quad C_3(5, 18, 5, 20, 50)$$

Questions

1. **Distances et affectation initiale :**
   (a) Calculez les distances euclidiennes entre chaque équipe et les centres initiaux.
   (b) Assignez chaque équipe au cluster le plus proche.
2. **Recalcul des centres :**

(a) Déterminez les nouveaux centres des clusters après la première itération.

(b) Répétez les calculs pour une deuxième itération.

3. **Visualisation en 2D :**

   - Utilisez une projection 2D (par exemple, sur les dimensions x_4 et x_5) pour représenter graphiquement les clusters obtenus.

4. **Analyse des clusters :**

   - Expliquez les caractéristiques des clusters. Quels types d'équipes (fortes, moyennes, faibles) se retrouvent dans chaque cluster?

5. **Programmation Python :**

   (a) Implémentez l'algorithme K-means en Python pour effectuer le clustering.

   (b) Ajoutez une visualisation des clusters en 2D avec des couleurs différentes pour chaque cluster.

   (c) Calculez la somme des distances intra-cluster (somme des distances des points à leur centre respectif).

   (d) Testez votre algorithme avec des centres initiaux différents pour observer l'impact sur les résultats.

6. **Exploration avancée :**

   - Intégrez la méthode du coude (Elbow Method) pour déterminer le nombre optimal de clusters.
   - Utilisez la bibliothèque `scikit-learn` pour comparer les résultats avec votre implémentation.