ROYAUME DU MAROC		المملكة المغربية
Université Abdelmalek Essaâdi		جامعة عبد المالك السعدي
Faculté des Sciences de Tétouan		كلية العلوم بتطوان
Tétouan		تطوان

TD 1 : K-means, Clustering hiérarchique

Professeur : Harchli Fidae

Exercice 1. Soit l'ensemble D des entiers suivants : $D = 2, 5, 8, 10, 11, 18, 20$
On veut répartir les données de D en trois clusters, en utilisant l'algorithme Kmeans. La distance d entre deux nombres a et b est calculée ainsi : $d(a, b) = |a - b|$

(a) Appliquez Kmeans en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul.

(b) Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires.

(c) Peut-on avoir un nombre d'itérations inférieur pour ce problème ? Discutez.

Exercice 2. On considère le jeu de données suivant (en deux dimensions) :

$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 0 \\ 10 & 2 \\ 10 & 4 \\ 10 & 0 \end{bmatrix}$$

(a) Appliquer l'algorithme K-means avec $k = 2$. Montrer les étapes de calcul (centroïdes initiaux, regroupements, mises à jour).

(b) Représenter graphiquement les points et les clusters obtenus.

(c) Expérimenter avec $k = 3$ et comparer les résultats. Qu'observe-t-on ?

(d) Ajouter des données bruitées (par exemple, $[5, 5]$, $[6, 6]$) et analyser leur impact sur le clustering.

Exercice 3. Pour le même jeu de données :

(a) Construire un dendrogramme en utilisant la méthode du lien simple.

(b) Identifier les clusters pour un seuil donné sur la distance.

(c) Comparer les résultats obtenus avec ceux du clustering K-means.

(d) Proposer une méthode pour évaluer la qualité des clusters obtenus.

Exercice 4. On considère les points suivants dans un espace bidimensionnel :

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \\ 8 & 8 \\ 9 & 9 \\ 10 & 10 \\ 20 & 25 \\ 21 & 26 \\ 22 & 27 \end{bmatrix}$$

Après application d'un algorithme de clustering avec $k = 3$, on obtient les clusters suivants :

- Cluster 1 : $(1, 2), (2, 3), (3, 4)$
- Cluster 2 : $(8, 8), (9, 9), (10, 10)$
- Cluster 3 : $(20, 25), (21, 26), (22, 27)$

Critères d'Évaluation

Pour évaluer la qualité des clusters, on utilise les critères suivants :

(1) **Compacité Intra-cluster SSW** : Reflète l'homogénéité des clusters. Plus elle est faible, mieux les points sont regroupés autour de leurs centroïdes. Elle est définie par :

$$W(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - \mu_i\|^2$$

où C_i est un cluster, $|C_i|$ est le nombre de points dans C_i, et μ_i est le centroïde de C_i.

(2) **Séparation Inter-cluster SSB** : Indique la distance entre les clusters. Une valeur élevée montre des clusters bien distincts. Elle est définie par :

$$S(C_i, C_j) = \|\mu_i - \mu_j\|$$

où μ_i et μ_j sont les centroïdes des clusters C_i et C_j.

(3) **Indice Silhouette d'un Point** : Évalue la pertinence de l'assignation des points à leurs clusters, idéalement proche de 1.

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

où :

- $a(x)$ est la distance moyenne entre x et les points de son propre cluster.
- $b(x)$ est la distance moyenne entre x et les points du cluster le plus proche.

Questions

(a) **Application des Critères :**

- Calculez la compacité intra-cluster pour chacun des clusters (1, 2, et 3).

- Calculez la séparation inter-cluster entre les clusters 1 et 2, puis entre les clusters 2 et 3.

- Calculez l'indice silhouette pour le point $(9, 9)$.

(b) **Analyse des Résultats :**

- Comparez les valeurs de compacité obtenues pour les trois clusters. Que remarquez-vous ? Quel cluster semble le plus compact ?

- Analysez la séparation entre les clusters. Quels clusters semblent les plus éloignés ?

- Interprétez l'indice silhouette calculé pour le point $(9, 9)$. Que reflète-t-il sur la qualité de l'assignation de ce point à son cluster ?

Exercice 5. Vous disposez des données suivantes sous forme de points (x_i, y_i) dans un espace à deux dimensions. Exemple de données :

$$\{(1.5, 2.2), (2.0, 1.8), (3.1, 3.4), (6.5, 7.3), (7.0, 6.9), (8.1, 8.2), (3.5, 3.6), (6.2, 7.1)\}$$

(1) Appliquez l'algorithme K-means pour partitionner ces points en $K = 3$ clusters. Après avoir obtenu les clusters, représentez graphiquement les clusters dans un graphique 2D, en indiquant les centres des clusters.

(2) Utilisez les mesures de similarité suivantes pour évaluer la qualité des clusters :

- **Distance intra-cluster** : Calculez la distance moyenne entre chaque point d'un cluster et son centre. Soit C_k un cluster et c_k son centre, la distance intra-cluster est donnée par :
$$D_{\text{intra}}(C_k) = \frac{1}{|C_k|} \sum_{i \in C_k} \|x_i - c_k\|$$
où $|C_k|$ est le nombre d'éléments dans le cluster C_k et x_i sont les points du cluster.
- **Distance inter-cluster** : Calculez la distance entre les centres de deux clusters différents C_k et C_j :
$$D_{\text{inter}}(C_k, C_j) = \|c_k - c_j\|$$

(3) Évaluez la qualité de votre partition en utilisant les critères suivants :

- **Indice de Rand (RI)** : Cet indice mesure la similarité entre deux partitions en comparant les paires d'objets. Plus la valeur de RI est proche de 1, plus la partition est bonne.
- **Indice de Davies-Bouldin (DBI)** : Mesure de la compacité et de la séparation des clusters. Un indice plus bas indique une meilleure partition. Il est défini par :
$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right)$$
où s_i est la compacité du cluster C_i, $d(c_i, c_j)$ est la distance entre les centres des clusters C_i et C_j, et K est le nombre total de clusters.