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

TD 3 : Unsupervised learning

Professeur : Harchli Fidae

Exercice 1. On dispose du jeu de données suivant en deux dimensions, représentant des observations dans un espace à 2 variables :

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 0 \\ 10 & 2 \\ 10 & 4 \\ 10 & 0 \\ 5 & 5 \\ 6 & 6 \\ 7 & 8 \\ 8 & 9 \end{bmatrix}$$

Ce jeu de données contient des points qui semblent appartenir à deux groupes distincts. Cependant, la présence de bruit (les derniers points) pourrait influencer les résultats.

(a) **Application du K-means :**

- Appliquez l'algorithme K-means avec $k = 2$ sur ce jeu de données. Affichez les étapes de l'algorithme, y compris les choix des centroïdes initiaux, l'attribution des points aux clusters, la mise à jour des centroïdes, et les itérations jusqu'à la convergence.

- Représentez graphiquement les points et les clusters obtenus après l'exécution de l'algorithme K-means.

(b) **Clustering Hiérarchique :**

- Appliquez le clustering hiérarchique avec la méthode du lien simple (Single Linkage). Construisez un dendrogramme pour visualiser la hiérarchie des clusters.

- Identifiez les clusters à un seuil de distance donné (choisissez un seuil approprié) et interprétez les résultats.

- Comparez les clusters obtenus avec ceux de l'algorithme K-means. Quelles différences et similarités observez-vous ?

(c) **Analyse en Composantes Principales (ACP) :**

- Appliquez une analyse en composantes principales (ACP) sur ce jeu de données pour réduire sa dimension à 1. Représentez graphiquement les points sur la première composante principale.

- Interprétez les résultats de l'ACP : quel est le rôle de la première composante principale ? En quoi cette réduction de dimension peut-elle aider à mieux comprendre la structure des données ?

- Visualisez les points dans l'espace des deux premières composantes principales. Que remarquez-vous concernant la distribution des points après projection ?

(d) **Comparaison des méthodes :**

- Discutez des avantages et inconvénients de chaque méthode (K-means, Clustering hiérarchique, et ACP) dans le contexte de ce jeu de données. En particulier, comment chaque méthode est-elle influencée par la présence de points bruités ?

- Si vous deviez recommander une méthode pour ce jeu de données en particulier, laquelle choisiriez-vous et pourquoi ?

Exercice 2. Partie 1 : Mesures de Similarité

(a) Calculez la distance euclidienne entre ces deux points.

(b) Calculez la distance de Manhattan entre $x = (3, 4)$ et $y = (1, 2)$.

(c) Calculez la similarité cosinus entre $x = (1, 2, 3)$ et $y = (4, 5, 6)$.

Partie 2 : Clustering

Le clustering est l'une des principales techniques d'apprentissage non supervisé. Il s'agit de regrouper les objets similaires en clusters. Répondez aux questions suivantes en expliquant vos raisonnements.

(a) Expliquer brièvement l'algorithme K-means pour le clustering. Décrivez les étapes de l'algorithme et le rôle de l'initialisation des centroïdes. Que se passe-t-il si les centroïdes sont mal initialisés ?

(b) Expliquer la différence entre le clustering hiérarchique agglomératif et divisif. Quels sont les avantages et inconvénients de chaque méthode ? Dans quel cas choisiriez-vous l'un plutôt que l'autre ?

(c) Proposez une méthode pour évaluer la qualité d'un clustering. En particulier, décrivez le rôle de l'indice de Rand ou de l'indice de Dunn dans cette évaluation.

Partie 3 : Réduction de Dimension

La réduction de dimension est une autre technique importante pour l'apprentissage non supervisé. Répondez aux questions suivantes en expliquant vos raisonnements.

(a) **PCA (Analyse en Composantes Principales) :** Expliquer l'idée de l'Analyse en Composantes Principales (PCA). Comment PCA permet-elle de réduire la dimensionnalité d'un jeu de données tout en préservant sa variance ?

(b) **Interprétation des Composantes Principales :** Après avoir appliqué une PCA sur un jeu de données à deux dimensions, on obtient les composantes principales suivantes :

$$\mathbf{v}_1 = (0.7, 0.7), \quad \mathbf{v}_2 = (-0.7, 0.7)$$

Interprétez ces composantes principales dans le contexte de réduction de dimension et expliquez ce qu'elles signifient en termes de structure des données.

(c) **t-SNE (t-Distributed Stochastic Neighbor Embedding) :** Le t-SNE est une autre méthode de réduction de dimension souvent utilisée pour la visualisation. Expliquez brièvement comment t-SNE fonctionne et en quoi il diffère de PCA. Dans quel cas t-SNE serait-il préférable à PCA ?

Partie 4 : Application Pratique

On vous donne le jeu de données suivant, qui représente des points dans un espace à 2 dimensions :

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

(a) Appliquez l'algorithme K-means avec $k = 3$ sur ce jeu de données. Décrivez les étapes de l'algorithme et interprétez les clusters obtenus.

(b) Appliquez le clustering hiérarchique sur le même jeu de données. Construisez un dendrogramme et identifiez les clusters pour un seuil donné. Comparez les résultats obtenus avec ceux du K-means.

(c) Appliquez l'ACP sur ce jeu de données et réduisez la dimension à 1. Représentez graphiquement les points projetés sur la première composante principale. Interprétez les résultats.

Exercice 3. Pour un jeu de données multidimensionnel fourni (ou généré aléatoirement) :

(a) Réduire la dimension avec l'ACP (conserver 90% de la variance).

(b) Appliquer le clustering K-means sur les données projetées.

(c) Interpréter les résultats obtenus.

(d) Comparer les performances du clustering K-means avec et sans réduction de dimension.

Exercice 4. Étude de cas : segmentation d'images

(a) Charger une image en niveaux de gris et représenter chaque pixel par ses coordonnées (x, y) et son intensité.

(b) Appliquer le clustering K-means pour segmenter l'image en $k = 3$ régions.

(c) Expérimenter avec d'autres valeurs de k et comparer visuellement les résultats.

(d) Discuter des avantages et des limites de cette méthode pour la segmentation d'images.