Université Abdelmalek Essaâdi
Faculté des Sciences et Techniques d'Al-Hoceima
Département de Mathématiques et Informatique

Filière : LST- IDLL
Module : Intelligence artificielle
A .U : 2023/2024
Pr : ZANNOU Abderrahim

TP N°5

Énoncé (Clustering)

Version 1

L'objectif de ce TP est d'appliquer :

- ✓ K-means
- ✓ DBSCAN

Tâche 1 : Clustering en utilisant l'algorithme de K-means

- ✓ **Jeu de données (Dataset) :** points (*label, x,y*)
- ✓ **Le fichier associé :** TP4_1_Clustering_K-Means_Points
- ✓ **Les étapes essentielles :**
  - Charger les données du Dataset
  - Visualiser les données du Dataset
  - Répartir les données pour le clustering
  - Créer le modèle de K-Means

```
from sklearn.cluster import KMeans
Model=KMeans(n_clusters=3)
```

  - Ajuster le modèle
  - Déterminer les clusters

```
cluster_labels=model.labels_
```

  - Visualiser les données après la phase de clustering
  - Evaluer le modèle

Tâche 2 : Clustering en utilisant l'algorithme de DBSCAN

- ✓ *Les deux paramètres de DBSCAN :*
  - Minimum de points ("MinPts") : Le nombre minimum de points requis pour former un cluster
  - ε (epsilon ou "eps") : La distance maximale entre deux points pour qu'ils appartiennent au même cluster.
- ✓ Minimum de points ("MinPts") :

- Il n'existe pas de méthode automatique pour déterminer la valeur de MinPts pour DBSCAN. La valeur de MinPts doit être définie en utilisant la connaissance du domaine et la familiarité avec l'ensemble de données.
- Voici quelques règles empiriques pour choisir la valeur de MinPts:
    - Taille de l'ensemble de données : Plus l'ensemble de données est volumineux, plus la valeur de MinPts doit être élevée.
    - Bruit dans l'ensemble de données : Si l'ensemble de données est plus bruité, choisissez une valeur de MinPts plus élevée.
    - Dimensionnalité des données : Généralement, MinPts doit être supérieur ou égal à la dimensionnalité de l'ensemble de données.
    - Données à 2 dimensions : Utilisez la valeur par défaut de DBSCAN, MinPts = 4 (Ester et al., 1996).
    - Données à plus de 2 dimensions : Choisissez MinPts = 2*dim, où dim représente la dimensionnalité de votre ensemble de données (Sander et al., 1998).

✓ Epsilon (ε) : Une fois que vous avez choisi votre valeur de MinPts, vous pouvez passer à la détermination de ε.

- On peut estimer automatiquement la valeur optimale de ε. Cette technique calcule la distance moyenne entre chaque point et ses k plus proches voisins, où k correspond à la valeur de MinPts que vous avez sélectionnée.
- Les distances moyennes pour les k plus proches voisins sont ensuite représentées graphiquement par ordre croissant sur un graphique k-distance. La valeur optimale pour ε se situe au point de courbure maximale (c'est-à-dire là où le graphique présente la plus forte pente).
- Le code suivant permet d'estimer la valeur de Epsilon (ε) :

```
from sklearn.neighbors import NearestNeighbors
neighbors = NearestNeighbors(n_neighbors=5) # On prend 4 points voisins
neighbors_fit = neighbors.fit(X)
distances, indices = neighbors_fit.kneighbors(X)
distances = distances[:,1:].mean(axis=1)
distances = np.sort(distances, axis=0)
```

plt.plot(distances)

plt.grid(True)

plt.show()

- ✓ **Jeu de données (Dataset) :** points (*label, x,y*)
- ✓ **Le fichier associé :** TP4_2_Clustering_DBSCAN_Sans_Normalisation_Points
- ✓ **Les étapes essentielles :**
  - Charger les données du Dataset
  - Visualiser les données du Dataset
  - Répartir les données pour le clustering
  - Déterminer les valeurs de MinPts et Epsilon (ε)
  - Créer le modèle de DBSCAN

    from sklearn.cluster import DBSCAN

    model=DBSCAN(eps= ?,min_samples= ?)
  - Ajuster le modèle
  - Déterminer les clusters

    cluster_labels=model.labels_
  - Visualiser les données après la phase de clustering
  - Evaluer le modèle

Tâche 3: Refaire la tâche 2 en normalisant les données du Dataset

- ➢ Comparer les résultats obtenus par les deux tâches ?
- ➢ **Le fichier associé :** TP4_3_Clustering_DBSCAN_Avec_Normalisation_Points