Forage de données

TP02: Travail pratique sur le regroupement

Objectifs

- Approfondissement les connaissances en regroupement
- Acquérir de nouvelles connaissances théoriques et pratiques sur le regroupement

Plan

- Partie théorique
- Partie pratique

Plan

- Partie théorique
- Partie pratique

Exercice 01 (2.5points)

- Le clustering par densité est un type de regroupement où des clusters détectés peuvent avoir différentes formes

- Questions:
  1. Expliquez davantage les principes de base de ce type de clustering en incluant ses principales caractéristiques
  2. Comparez ce type de clustering avec le clustering avec des approches de clustering tel que le K-means

Exercice 02 (2.5points)

- DBSCAN est un des algorithmes développés pour effectuer du clustering basé sur la densité

- Questions:
  1. Expliquez le fonctionnement de cet algorithme
  2. Donnez et expliquez un pseudocode de cet algorithme
  3. En quoi l'algorithme OPTICS qui est une extension de DBSCAN diffère de ce dernier?

Plan

- Partie théorique
- Partie pratique

Exercice 03 (2.5points)

- K-Means et DBSCAN dans Weka:
  1. Effectuez un regroupement des données Iris avec K-Means en utilisant K = 3
  2. Faites la même chose, mais cette fois-ci avec DBSCAN
  3. Comparez les résultats obtenus

Exercice 04 (2.5points)

1. Explorez l'utilisation de K-Means avec Python en utilisant la librairie scikit-learn
    1. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
2. Expliquez la méthode présentée dans la page ci-dessous pour déterminer le K optimal
    1. https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f
3. Appliquez cette méthode à un autre jeu de données de votre choix

Directives

- Travail à réaliser en équipes de deux
- Pour chaque exercice, indiquez le pourcentage de contribution de chaque coéquipier
- Fichiers à remettre:
  - Présentation pptx ou latex
  - Code source
- Éléments qui seront pris en considération lors de l'évaluation:
  - Complétude et exactitude des travaux remis
  - Qualité de la présentation
  - Contribution des coéquipiers
  - Respects des consignes
- Date de remise:
  - 12 avril 23h59
  - Il y aura éventuellement une séance pour la présentation des travaux (date à déterminer): maximum de 10 minutes