Département IA & SD le 27/10/2025

Section: 4 ING Software

Module : FD1 TP N° 2

La fouille de données ou "data mining" est un processus qui consiste à explorer de grandes quantités de données pour y découvrir des **informations utiles et inconnues au préalable**.

La fouille de données est un outil puissant pour transformer des données brutes en connaissances exploitables.

Ce TP N°2 portes sur le clustering de données à partir d'un benchmark. Voici les étapes à suivre :

1. **Prétraitement des données :** Avant de commencer le clustering, il est essentiel de préparer les données. Cette étape peut inclure le nettoyage des données (gestion des valeurs manquantes, suppression des doublons), la normalisation ou la standardisation des données pour éviter que certaines variables n'influencent trop les résultats, et éventuellement la réduction de dimension pour simplifier le problème.
2. **Détermination du nombre de clusters (k) :**
    - **Courbe d'Elbow :** Tracer la courbe d'Elbow (coude) permet de visualiser l'inertie intra-cluster en fonction du nombre de clusters. L'inertie représente la somme des distances au carré entre chaque point et le centre de son cluster. On recherche le point où la diminution de l'inertie ralentit (le "coude"), ce qui indique un nombre de clusters optimal.
3. **Algorithme K-Means :**
    - Appliquer l'algorithme K-Means avec le nombre de clusters (k) déterminé à l'étape précédente. K-Means est un algorithme de clustering qui divise les données en k clusters en minimisant l'inertie intra-cluster.
    - **Mesures de performance :** Évaluer les résultats du clustering à l'aide de mesures de performance telles que le coefficient de silhouette. Les indices permettent de quantifier la qualité du clustering en mesurant la compacité des clusters et leur séparation.
4. **Algorithme K-Medoids :**
    - Appliquer l'algorithme K-Medoids avec le même nombre de clusters (k). K-Medoids est similaire à K-Means, mais au lieu d'utiliser le centre de gravité (moyenne) des points pour représenter un cluster, il utilise le point le plus proche des autres points du cluster (médoïde).
    - **Mesures de performance :** Calculer les mêmes mesures de performance que pour K-Means afin de pouvoir comparer les deux algorithmes.

5. **Comparaison des performances :**
    - **Histogramme des inerties :** Créer un histogramme comparant les inerties obtenues avec les deux méthodes (K-Means et K-Medoids). Cela permettra de visualiser et de comparer la dispersion des inerties pour chaque méthode.
    - **Analyse :** Analyser les résultats obtenus avec les deux méthodes en comparant les mesures de performance et les histogrammes d'inertie. Discuter des avantages et des inconvénients de chaque méthode en fonction des résultats observés.

Rapport à remettre le 10 /11/2025