Master EID²

Exploration Informatique des Données et Décisionnel
UE Science des Données Numériques
Fiche TP N° : 5

Le but de ce TP est de créer une fonction Python capable de détecter des clusters de données homogènes dans un ensemble de données, puis d'analyser un jeu de données réelles.

A) K-Moyennes :

1. Écrivez en python l'algorithme des K-Moyennes sous la forme d'une fonction. Vous trouverez une description de l'algorithme dans le cours ou sur internet. La fonction prend en entrée deux paramètres : la matrice des données et le nombre de clusters que l'on souhaite. Testez sur les données *Iris* ou sur des données que vous générez. Comparez avec la fonction *Kmeans* de *sklearn.*

2. Expérimenter l'instabilité due à l'initialisation : les centres des clusters étant choisis au hasard lors de l'initialisation, le résultat obtenu peut varier d'une exécution à l'autre. Vérifiez que c'est le cas.

3. Utiliser l'indice de Silhouette (qui est dans le package *sklearn*) pour stabiliser les résultats et sélectionner automatiquement le nombre de groupes. Pour ce faire, créez un script qui applique K-moyenne sur les données pour différents nombres de clusters allant de 2 à 10, 10 fois pour chaque nombre de clusters (soit 90 fois en tout) et qui renvoie la solution ayant le meilleur score de Silhouette.

4. Utiliser un ACP (fonction *PCA* de *sklearn*) pour vérifier visuellement la cohérence des groupes obtenus. Vérifier aussi visuellement la séparabilité et la compacité de ces groupes à l'aide d'une ADL (fonction *LinearDiscriminantAnalysis* de *sklearn).* Quelle est la différence entre les deux méthodes ?

B) Analyse des données « choix projet » :

Les données *choixprojetstab.csv* contiennent des données anonymisées de vœux faits par les étudiants du master 2 pour des projets à réaliser en binômes. Chaque étudiant a fait des vœux sur les différents projets en leur donnant une mention très bien, bien, moyen ou jamais. Les projets non mentionnés peuvent être considérés comme « moyens » et on peut dire que très bien = 3, bien = 2, moyens = 1, jamais = 0. L'objectif est de former des clusters d'étudiants ayant faits des choix similaires.

1. Utilisez le package *csv* (ou l'importation de variable de Spyder) pour lire le fichier et remplir deux variables : la liste des codes « C » représentant les étudiant (première colonne) et la matrice « M » des données (tout sauf la première ligne et la première colonne). La matrice M doit être de type *array* du package *numpy.* Faites attentions à ce que les valeurs dans M soient bien numériques (1, 2, 3) et non textuelle ('1', '2', '3'). Vous pouvez utiliser la méthode *astype* de *numpy* en cas de besoin.

2. Dans *sklearn.cluster* il existe différents algorithmes de clustering. Testez les différents algorithmes du package et proposez le meilleur clustering possible des données selon l'indice Silhouette.