

Data Science et Machine Learning

TP sur la classification automatique non-supervisée (Catégorisation ou Clustering)

A réaliser dans l'environnement Python avec les bibliothèques Scikit-learn, Numpy et Matplotlib

L'objectif est de réaliser une catégorisation (clustering) avec la méthode K-means sur le même jeu de données que le TP précédent.

1. Utilisez la bibliothèque Pandas pour lire le data frame qui se trouve dans le fichier Excel points.xlsx et construisez le tableau Numpy des données X.
2. Construisez une catégorisation des individus en 3 classes avec la fonction KMeans du module cluster de la bibliothèque Scikit-learn en précisant le nombre de classes (n_clusters = 3)

```
from sklearn.cluster import KMeans  
kmeans = KMeans(n_clusters=m, random_state=0).fit(X)
```
3. Affichez la partition finale qui se trouve dans l'attribut labels_.
4. Affichez les centres de gravité qui se trouvent dans l'attribut cluster_centers_.
5. Visualiser la catégorisation obtenue en affectant une couleur à chaque classe et en affichant les centres de gravités avec les individus.
6. Visualiser la catégorisation sur 3 classes en affectant une couleur à chaque classe et en annotant les individus avec leurs numéros.