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 de la Classification Ascendante Hiérarchique en utilisant 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. Réaliser un regroupement progressif des individus avec la méthode de classification ascendante hiérarchique qui est implémentée dans la fonction AgglomerativeClustering du module cluster de la bibliothèque Scikit-learn. Remarquez que la méthode réalise par défaut une catégorisation en deux classes. Il faut donc préciser le nombre m voulu de classes s'il est différent de 2 avec le paramètre (n_ clusters = m).
   ```
   from sklearn.cluster import AgglomerativeClustering
   AC = AgglomerativeClustering(n_clusters=3, linkage='ward', compute_distances=True).fit(X)
   ```

3. Affichez la partition finale qui se trouve dans l'attribut labels_.

4. Affichez le regroupement hiérarchique qui se trouve dans l'attribut children_. C'est le lien Ward qui est utilisé par défaut. Pour les liens, minimal, moyen ou maximal, il faut préciser, respectivement, single, average ou complete, comme valeur du paramètre linkage.

5. Affichez les distances correspondantes au regroupement hiérarchique qui se trouvent dans l'attribut distances_.

6. Visualiser la catégorisation sur 3 classes en affectant une couleur à chaque classe et en annotant les individus avec leurs numéros. Vérifiez sur le dessin que c'est bien logique que le premier groupement soit celui des deux singletons {15} et {28}.