TP6 : Clustering

Durée : 6h

L'objectif de ce TP est l'étude des deux algorithmes de clustering k-means et CAH. Les exercices à faire « à la main » permettent de mieux comprendre l'algorithme. Le logiciel R permet d'étudier le comportement des algorithmes suivant différents cas de figures illustrés par des jeux de données simulées.

Exercice 1

On considère 5 points $x_1=1$, $x_2=2$, $x_3= 9$, $x_4=12$ et $x_5=20$.

1) Appliquer l'algorithme des k-means avec les valeurs de k et les points de départ suivants. Calculer le pourcentage d'inertie expliquée par la partition obtenue.

a) k=2, $g_1=1$ et $g_2=20$

b) k=2, $g_1=2$ et $g_2=9$

c) k=3, $g_1=1$, $g_2=9$ et $g_3=12$

d) Quel est le meilleur regroupement des trois ?

2) Appliquer une méthode de classification hiérarchique ascendante en utilisant la distance minimum comme critère de dissimilarité entre classes. Tracer le dendrogramme. Quel regroupement vous parait correct ?

Exercice 2 : Kmeans et CAH

L'objectif de cet exercice est de tester les algorithmes k-means et CAH sur des jeux de données simulés,

Test_Clusters_Distincts.txt

Test_Clusters_Random.txt

Test_Clusters_Melanges.txt

Test_Clusters_Atypiques.txt

Pour cela, on utilisera le langage R avec ses fonctions ***kmeans*** et ***hclust***.

Partie 1 : Algorithme des Kmeans

a) Tester l'algorithme des kmeans sur les données simulées Test_Clusters_Distincts.txt. Essayer plusieurs nombres de classes et choisir le meilleur.

b) Tester l'algorithme des kmeans sur les données simulées Test_Clusters_Distincts.txt, Test_Clusters_Melanges.txt et Test_Clusters_Random.txt. Constater l'évolution de l'inertie expliquée.

c) Tester l'algorithme des kmeans sur les données simulées Test_Clusters_Corr.txt. Que pourrait-on faire pour améliorer le résultat.

d) Tester l'algorithme des kmeans sur les données Test_Clusters_Atypique.txt avec les individus n°1 et n°1499 pour initialisation.

Partie 2 : Classification Ascendante Hiérarchique (CAH)

a) Tester l'algorithme CAH sur les données Test_Clusters_Distincts.txt, Test_Clusters_Melanges.txt et Test_Clusters_Random.txt.

b) Tester l'algorithme CAH sur les données Test_Clusters_Atypique.txt avec la méthode « ward.D2 ».

c) Comparer les résultats obtenus entre CAH et Kmeans sur les données Test_Clusters_Distincts.txt, Test_Clusters_Melanges.txt et Test_Clusters_Random.txt.

Exercice 3 : jeu de données

Déterminer des clusters dans le jeu de données « iris ». Est-ce que les clusters correspondent aux trois types de fleurs ?