TP : Non-supervisée

Vous pouvez faire les exercices dans le langage de votre choix.

1 Exercice 1 : Arrestation aux USA

Dans cet exercice, on se propose de faire un clustering des états américains par rapport à certaines statistiques d'arrestation.

1. Télécharger le jeu de données suivant : lien.
2. Faire un clustering des états en utilisant la méthode des plus proches voisins et la distance euclidienne. Tracer le dendrogramme.
3. Couper le dendrogramme à une hauteur qui résulte en trois groupes distincts. Quels états appartiennent à quel groupe ? Interpréter les groupes obtenus.
4. Faire le même clustering, mais cette fois après avoir standardisé les variables. Les variables auront donc un écart-type de 1. Tracer le dendrogramme et le couper pour avoir trois groupes.
5. Quel effet à la standardisation des variables sur le clustering obtenu ? À votre avis, doit-on standardiser les variables avant que les distances entre les observations soient calculées ? Justifier.

2 Exercice 2 : De la musique.

Dans cet exercice, on se propose de comparer la classification ascendante hiérarchique et la classification descendante hiérarchique.

1. Télécharger le jeu de données suivant : lien.
2. Faire un clustering en 6 groupes des variables `m_1` à `m_9` à l'aide d'une classification ascendante hiérarchique en utilisant la méthode de la moyenne et une distance euclidienne. Interpréter le dendrogramme.

3. Faire un clustering en 6 groupes des variables `m_1` à `m_9` à l'aide d'une classification descendante hiérarchique en utilisant la méthode de la moyenne et une distance euclidienne. Interpréter le dendrogramme.
4. Calculer le critère de Silhouette des deux clustering obtenus. Interpréter les résultats.
5. Comparer les résultats des méthodes ascendante et descendante.

3 Exercice 3 : De la musique (suite)

Dans cet exercice, on se propose de tester l'algorithme des k-means.

1. Télécharger le jeu de données suivant : lien.
2. Utiliser le k-means pour faire un regroupement des observations en $K = 4$ groupes.
3. Décrire brièvement les quatre groupes avec les statistiques descriptives classiques.
4. Donner la taille des groupes. Identifier les musiques appréciées ou non dans chaque groupe.