

TP – Clustering et classification

novembre 2024

Il y a quatre sujets, choisissez trois et traitez les.

1. Programmez dans le langage de votre choix la méthode de k-moyennes, ensuite testez le code sur des données "jouet" avec des clusters bien séparés et, dans un deuxième temps, sur des données plus compliquées. Comparez vos résultats avec ceux issus d'une fonction existante.
2. En utilisant R appliquez les fonctions de clustering (classification non-supervisée) sur les données `iris` qui existent dans la base R. Il y a déjà une classification faite (en trois classes), cette classification nous permettra de juger de la qualité des méthodes appliquées.
 - Représenter graphiquement les résultats de chaque méthode de clustering.
 - Interpréter les résultats obtenus.
 - Utilisez maintenant la fonction de classification par la méthode K-NN. Dressez la matrice de confusion.
3. Travaillez les données `mammal.dentition` qui existent dans la base R dans la librairie `cluster.datasets`. Il faut, naturellement, ignorer la première colonne.
 - Appliquez les fonctions de clustering (classification non-supervisée) sur les données
 - Indiquez quelle méthode produit les meilleurs résultats.

- Classifiez manuellement les animaux ou prenez la classification fournie sur eCampus, éliminez (s'il y a) les classes avec une seule observation, puis utilisez la fonction de classification par la méthode k-NN. Dressez la matrice de confusion.
 - Quelle est le meilleur choix pour la valeur k ?
4. Travaillez les données **dataset diabtes** qui sont sur eCampus.
- Examinez attentivement les données.
 - Appliquez des fonctions de clustering (classification non-supervisée) sur les données afin de voir si on retrouve la classification initiale.
 - Analysez la qualité des résultats avec les données brutes et celles scalées et centrées.
 - Tentez ensuite la classification avec k-NN.
 - Commentez la qualité des résultats obtenus en essayant de donner de explications.

*Le compte-rendu sous la forme d'un unique fichier (code R commenté ou archive **.zip** ou **.rar** avec code **.r** et compte-rendu en **.pdf**) à rendre sur eCampus.*

*Si vous n'êtes absolument pas à l'aise avec R, vous pouvez utiliser Python et les librairies de Python, y compris **scikit-learn***