

Classification Automatique

TP 2 : K-means sur données continues

1. Charger les données mtcars à partir de la librairie de R et effectuer une analyse sommaire des données (description des données, analyse univariée et bivariée,.). description du fichier :

mpg : Miles/(US) gallon

cyl : Number of cylinders

disp : Displacement (cu.in.)

hp : Gross horsepower

drat : Rear axle ratio

wt : Weight (1000 lbs)

qsec : 1/4 mile time

vs : Engine (0 = V-shaped, 1 = straight)

am : Transmission (0 = automatic, 1 = manual)

gear : Number of forward gears

carb : Number of carburetors

2. Visualiser le jeu de données à partir des attributs des variables wt et qsec, à partir variables wt et mpg. Combien de sous-ensembles d'individus proposeriez vous sur la base de ces représentations ?
3. Les méthodes de segmentation de type k-means sont implémentées dans la fonction **k-means** de R. En utilisant la méthode de **mac Queen**, faire une segmentation des mtcars à partir uniquement des variables wt et qsec.
 - Donner les coordonnées des centres des classes ainsi que les variances inter classes et intra-classe obtenus.
 - Calculer le R^2 .
 - Visualiser le nuage de points, en affectant à chaque individu une couleur propre à sa classe d'affectation, les centres des classes devront se distinguer des autres points.
 - Donner la matrice de confusion et commenter les résultats obtenus.
 - Décrire et analyser les classes obtenues.
4. Reprendre la question précédente en utilisant les attributs wt et mpg, Commenter les résultats.
5. On suppose le nombre de classe inconnu on considère les attributs wt et mpg.

- Calculer et représenter la courbe de R^2 , déterminer le nombre de classes optimal
- Faire une représentation graphique de la segmentation pour le nombre de classe retenue

Discuter des résultats obtenus.

6. Reprendre la question 3 en utilisant l'option **nstart** de la fonction k-means, décrire la fonctionnalité de cette option, segmenter les données en l'utilisant. comparer la segmentation obtenue avec celle de la question 5.
7. Reprendre la question 3 en appliquant respectivement les méthodes de **FORGY** et des **nuées dynamiques**. Comparer à chaque fois les résultats avec ceux de la méthode de **Mac-Queen**.