Classification non supervisée
k-means et CAH

Charlotte Baey (charlotte.baey@univ-lille.fr)

1 Exercice 1 (R) : données simulées

1. Télécharger le fichier `datasets2.R` qui contient 4 fonctions permettant de générer différents jeux de données.
2. Générer les 4 jeux de données correspondants, en prenant $n = 200$. Tracer les nuages de points obtenus.
3. Sur chaque jeu de données, lancer l'algorithme des k-means à l'aide de la fonction `kmeans` en justifiant le choix du nombre de clusters. Dans un premier temps, laisser les valeurs par défaut pour les options de la fonction.
4. En comparant les résultats obtenus sur les différents jeux de données, pouvez-vous identifier le type de distributions pour lesquelles l'algorithme des k-means n'est pas adapté ?
5. Comparer les résultats précédents avec ceux obtenus par classification ascendante hiérarchique, à l'aide de la fonction `hclust`. Comparer les résultats obtenus avec différents critères d'aggrégation.
6. Les deux algorithmes (k-means et CAH) se comportent-ils de la même façon sur ces jeux de données ?

2 Exercice 2 (R) : classification des villes européennes

Dans cet exercice, on cherche à classer les villes européennes en fonction du climat. Pour cela, on a relevé les températures mensuelles moyennes, la température annuelle moyenne et l'amplitude de variation des températures, ainsi que la latitude et la longitude de 35 villes.

1. Importer la base de données `temperatures.csv` et faire une analyse descriptive de la base de données. Y a t-il des données manquantes ?
2. Proposer une classification de la base de données, à l'aide de l'algorithme des k-means et de l'algorithme de CAH, et sans tenir compte de la variable `Region`. Combien de classes retient-on ? Comparer les classes obtenues avec les deux méthodes.
3. Décrire les classes obtenues, et proposer une interprétation des résultats.

3 Exercice 3 (R) : classification de données médicales

On s'intéresse dans cet exercice à une base de données sur le cancer du sein. Pour 699 femmes, on dispose de plusieurs informations recueillies au cours d'une biopsie de la tumeur et enregistrées dans les variables suivantes :

— `CL.thickness` : l'épaisseur de la membrane plasmique des cellules
— `Cell.size` : uniformité de la taille des cellules
— `Cell.shape` : uniformité de la forme des cellules
— `Marg.adhesion` : expression de la protéine Integrin beta3 au niveau de la surface celullaure
— `Epith.c.size` : taille des cellules épithéliales
— `Bare.nuclei` : score associé à la présence de nuléoles en-dehors du noyau de la cellule
— `Bl.cromatin` : expression de la protéine qui induit l'expression du gène du récepteur d'œstrogènes
— `Normal.nucleoli` : score associé à la présence de l'ADN à l'extérieur de la cellule
— `Mitoses` : score associé au nombre de mitoses
— `Class` : vaut 1 si la tumeur est maligne, et 0 si elle est bénigne.

Toutes ces variables sont comprises entre 1 et 10. On dispose également du numéro d'identification de la patiente.

L'objectif de l'exercice est de classer les données en deux groupes afin de caractériser les tumeurs bégnines et malignes.

1. Importer la base de données `BreastCancerData.csv` et faites une analyse descriptive de la base de données. Y a t-il des données manquantes ?
2. Proposer une classification de la base de données **sans utiliser la variables `Class`**, à l'aide de l'algorithme des k-means et de l'algorithme de CAH. Comparer les classes obtenues avec les deux méthodes.
3. Décrire les classes obtenues, et proposer une interprétation. Quelle classe représente les tumeurs malignes ? A t-on bien identifié ces tumeurs ?