Fouille de données
Classification ascendante hiérarchique

L'objectif de ce TP est de se familiariser avec la pratique de la classification ascendante hiérarchique (CAH) dans R pour données quantitatives. Un compte-rendu de TP vous est demandé. Ce compte rendu doit contenir le code `R` utilisé et d'abondants commentaires.

1 Les données USArrests

1. Charger les données `USArrests` en tapant `data(USArrests)`.
2. Créer les données centrées-réduites correspondantes en utilisant la fonction `scale`. Ces dernières seront stockées dans la matrice `d`.
3. Afin de réaliser une CAH utilisant le saut de Ward (`method = "ward.D"`), créer une matrice de distances (au carré) en tapant :
   ```
   > distances <- dist(d)^2/2
   ```
   Que calcule `dist` exactement ? Expliquez la présence du carré ainsi que la division par 2.
4. Effectuer alors la CAH avec saut de Ward en utilisant la fonction `hclust` et représenter le dendrogramme. Essayer aussi `plot(h,hang=-1)`.
5. Représenter la distance entre classes agrégées ("hauteur") en fonction des regroupements. Expliquer pourquoi les partitions en 2 et 4 classes semblent intéressantes.
6. En utilisant la fonction `cutree`, récupérer la partition en 2 classes en tapant :
   ```
   > classe <- cutree(h,2)
   ```
7. Rajouter la variable `classe` à la matrix `d`. Les individus de la première classe peuvent être récupérés en tapant :
   ```
   > classe1 <- subset(d,classe==1)
   ```
8. Calculer le cardinal et le centre de gravité de la classe 1 en tapant :
   ```
   > n1 <- nrow(classe1[,1:4])
   > cg1 <- colMeans(classe1[,1:4])
   ```
   et son inertie :
   ```
   > nvar <- ncol(USArrests)
   > inertie1 <- mean(rowSums(classe1[,1:4] - matrix(cg1,n1,nvar,byrow=T))^2)
   ```
9. Calculer également le cardinal, le centre de gravité et l'inertie de la classe 2. Laquelle des deux classes est la plus dispersée ?
10. Interpréter les classes.

2 Visualisation des résultats dans le premier plan factoriel

1. Visualiser la classication dans le plan engendré par les deux premières composantes principales des données centrées-réduites. Interpréter les composantes principales et la partition en deux classes retenue.
2. Répéter toutes les étapes précédentes avec la partition en quatre classes.

3 Influence du lien

Effectuez la CAH des données `USArrests` centrées-réduites en utilisant la distance Euclidienne et, successivement, le lien simple, le lien complet et le lien moyen. Comparez visuellement les trois dendrogrammes obtenus. Que pouvez-vous dire ?