SY09 Printemps 2014
TP 2
Classification automatique

Exercice 1. Visualisation des données

L'objectif de cet exercice est de visualiser les données qui seront étudiées dans la suite de ce TD. Pour ce faire, on pourra utiliser l'analyse en composantes (ACP) principales vue au TD précédent, ainsi que l'analyse factorielle d'un tableau de distances (AFTD). Cette méthode, que nous ne détaillerons pas (elle est étudiée plus précisément en SY19), peut être vue comme l'équivalent de l'ACP lorsque les données disponibles se présentent sous la forme d'une matrice de dissimilarités.

L'AFTD permet ainsi d'obtenir une représentation multidimensionnelle des données exacte lorsque les dissimilarités correspondent à une distance Euclidienne. Dans le cas contraire, la représentation des données obtenue est plus ou moins fidèle aux dissimilarités présentes dans la matrice. Dans tous les cas, après sélection d'un certain nombre de variables, la qualité de la représentation peut être évaluée au moyen d'un diagramme de Shepard.

Sous R, l'AFTD peut être effectuée au moyen de la commande cmdscale, le diagramme de Shepard par la fonction Shepard.

1. Charger les données Iris et sélectionner les variables quantitatives, au moyen du code R suivant :

   ```
   > library(MASS)
   > data(iris)
   > donnees <- NULL
   > donnees$num <- iris[,c(1:4)]
   > donnees$cls <- iris[,5]
   ```

   Afficher les données dans le premier plan factoriel. Que constatez-vous ?

2. Effectuer l'ACP des données Crabs, préalablement traitées de manière à supprimer l'effet taille :

   ```
   > library(MASS)
   > data(crabs)
   > crabsquant <- crabs[,4:8]
   > crabsquant <- crabsquant/matrix(rep(crabsquant[,4],dim(crabsquant)[2]),
   > nrow=dim(crabsquant)[1],byrow=F)
   > crabsquant <- crabsquant[,-4]
   ```

   Comparer à la représentation des Crabes suivant les variables RW et BD. Que remarquez-vous ?

3. Effectuer l'AFTD des données Mutations. Les données peuvent être chargées à partir du fichier mutations2.txt au moyen de la commande suivante :

   ```
   > read.table("mutations2.txt", header=F, row.names=1)
   ```

   Afficher et analyser la représentation ainsi obtenue. Que pouvez-vous en dire ?

Exercice 2. Classification hiérarchique

1. En utilisant la fonction hclust, effectuer la classification hiérarchique ascendante (avec les différents critères d'agrégation disponible) des données de mutations. Commenter et comparer les résultats obtenus.
2. Effectuer la classification hiérarchique ascendante des données Iris. Commenter les résultats obtenus, en vous appuyant sur votre connaissance de ce jeu de données, et sur leur représentation dans le premier plan factoriel.
3. Effectuer la classification hiérarchique descendante des données Iris, au moyen de la fonction diana (module cluster). Comparer aux résultats obtenus au moyen de la CAH.

Remarque importante : l'utilisation de la fonction hclust pour effectuer la classification hiérarchique ascendante avec le critère de Ward, lorsque celui-ci a un sens (tableau de distances euclidiennes), nécessite d'élever les distances fournies au carré.

Exercice 3. Centres mobiles

Le but de cet exercice est de tester les performances de l'algorithme des centres mobiles sur deux jeux de données réelles : Iris et Crabs.

Données Iris

1. Tenter une partition en $K \in \{2, 3, 4\}$ classes avec la fonction `kmeans` ; visualiser et commenter.
2. On cherche à présent à étudier la stabilité du résultat de la partition. Effectuer plusieurs classifications en $K = 3$ classes du jeu de données. Observer les résultats, en termes de classification obtenue et d'inertie intra-classes. Ces résultats sont-ils toujours les mêmes ? Commenter et interpréter.
3. On cherche à déterminer le nombre de classes optimal.
   (a) Effectuer $n = 100$ classifications en prenant $K = 2$ classes, puis $K = 3$ classes, $K = 4$ classes, ..., jusqu'à $K = 10$ classes. On constitue ainsi neuf échantillons iid $\{I_{K\,1}, \ldots, I_{K\,100}\}$ contenant 100 valeurs d'inertie intra-classe chacun.
   (b) Pour chaque valeur de K, calculer l'inertie intra-classe minimale $\widehat{I_K}$. Représenter la variation de l'inertie minimale en fonction de K. Proposer un nombre de classes en se basant sur ces informations.
4. Comparer les résultats de la partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes.

Données Crabs

Effectuer la classification des données Crabs au moyen de l'algorithme des centres mobiles. Comparer à la partition réelle des crabes suivant l'espèce et le sexe.