SY09 Printemps 2011
TP 2
Classification automatique

Exercice 1. 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. Les données peuvent être chargées à partir du fichier mutations.txt au moyen de la commande suivante :

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

   Remarque importante : utiliser la fonction hclust pour effectuer la classification hiérarchique ascendante avec le critère de Ward nécessite d'élever les distances fournies au carré.
2. Effectuer la classification hiérarchique ascendante des données Iris. Pour les charger et sélectionner les variables quantitatives, vous pourrez utiliser le code R suivant :

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

   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.

Exercice 2. Les 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

Charger les données Crabs, puis les pré-traiter 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]
```

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