SY09 Printemps 2016
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 principales (ACP) étudiée dans le TD précédent, ainsi que l'analyse factorielle d'un tableau de distances (AFTD), présentée en cours. On considérera trois jeux de données : les `Iris` de Fisher, les données `Crabs`, et un jeu de données `Mutations` de dissimilarités entre espèces (voir la figure 1 pour un descriptif des données).

On rappelle que l'AFTD peut être vue comme un équivalent de l'ACP pour des données se présentant sous la forme d'un tableau $n \times n$ de dissimilarités δ_{ij} entre n individus ($i, j \in \{1, \ldots, n\}$) : elle calcule une représentation multidimensionnelle de ces individus (dont le tableau de dissimilarités ne donne qu'une description implicite) dans un espace euclidien de dimension $p \leq n$. Cette représentation est exacte lorsque les dissimilarités sont des distances euclidiennes.

Après sélection d'un certain nombre de variables, la qualité de la représentation peut être évaluée numériquement par un critère similaire au pourcentage d'inertie de l'ACP, ou graphiquement au moyen d'un *diagramme de Shepard* : sur ce graphique, la distance $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ entre les représentations de $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ déterminées par l'AFTD est représentée en fonction de la dissimilarité initiale δ_{ij}, pour chaque couple d'individus $(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Sous R, l'AFTD peut être effectuée au moyen de la commande `cmdscale`, et les distances nécessaires au tracé du diagramme de Shepard sont obtenues par la fonction `Shepard` (bibliothèque `MASS`).

1. Charger les données `Iris` :
   `> data(iris)`
   Afficher les données dans le premier plan factoriel, tout d'abord sans tenir compte de l'espèce. Que constatez-vous ? Combien de groupes de points le jeu de données semble-t-il comporter ?
   Afficher ensuite les données en ajoutant l'information d'espèce (par exemple via la couleur ou le symbole des points affichés). Que constatez-vous ? À quoi peut-on s'attendre si l'on recherche une partition des données ?
2. Charger le jeu de données `Crabs` à partir du fichier `crabs2.csv` disponible sur le site de l'UV.
   `> crabs2 <- read.csv("crabs2.csv", header=T)`
   Afficher les données dans le premier plan factoriel, tout d'abord sans tenir compte de l'espèce ou du sexe des crabes. Combien de groupes de points apparaissent ?
   Afficher ensuite les données en tenant compte des informations d'espèce et de sexe. Que constatez-vous ?
3. Charger les données de `Mutations` à partir du fichier `mutations2.csv` et déclarer les données comme tableau de dissimilarités ; on utilisera la commande suivante :
   `> mut <- read.csv("mutations2.csv", header=T, row.names=1)`
   `> mut <- as.dist(mut, diag=T, upper=T)`
   Calculer une représentation euclidienne des données en $d = 2$ variables par AFTD, et l'afficher ; calculer la qualité de la représentation, et afficher le diagramme de Shepard. Que peut-on dire ?
   Recommencer avec un nombre de variables d de représentation allant de 3 à 5. Interpréter les résultats.

Remarques : on pourra utiliser la fonction `text` *pour afficher des chaînes de caractère à des coordonnées spécifiées dans une matrice. La fonction* `interaction` *permettra de fusionner deux variables qualitatives (*`factors`*) pour n'en former plus qu'une.*

```
+--------------------------------------------------------------------+
|     MUTATION DISTANCES AMONG 20 SPECIES (FITCH AND MARGOLIASH)     |
|                                                                    |
|     The source of this data is a paper by Fitch and Margoliash     |
| in Science(1967).   For a more recent reference see Scientific     |
| American (1972?).                                                  |
| Every species has a protein  molecule, Cytochrome c, which varies  |
| from species to species but has a similar function for all. It     |
| consists of a long chain of amino acids. There are only a few      |
| acids,  but different  molecules are  obtained  by varying  the    |
| acids in each  position in the chain. The  number of positions     |
| with different  acids measures  distance between  two species.     |
| these distances are given in the data below.                       |
| For example, the amino acids in Cytochrome c for two species look  |
| like this:                                                         |
|  Moth            XXYVPLY .........SEXI                             |
|  Screwworm fly   XXYVPLY .........LSEI                             |
| where the whole chain is 110 in length, and the letters represent  |
| particular amino acids. Each difference contributes to mutation    |
| distance according to the minimum number of nucleotides that would |
| need to be changed to convert one into the other.                  |
| Fitch & Margoliash used these data to construct a phylogenetic     |
| tree.                                                              |
| Ref: Science, v. 155, 279-284.                                     |
+--------------------------------------------------------------------+

Man                 0
Monkey             01  0
Dog                13 12  0
Horse              17 16 10  0
Donkey             16 15 08 01  0
Pig                13 12 04 05 04  0
Rabbit             12 11 06 11 10 06  0
Kangaroo           12 13 07 11 12 07 07  0
Pekin Duck         17 16 12 16 15 13 10 14  0
Pigeon             16 15 12 16 15 13 08 14 03  0
Chicken            18 17 14 16 15 13 11 15 03 04  0
King Penguin       18 17 14 17 16 14 11 13 03 04 02  0
Snapping Turtle    19 18 13 16 15 13 11 14 07 08 08 08  0
Rattlesnake        20 21 30 32 31 30 25 30 24 24 28 28 30  0
Tuna               31 32 29 27 26 25 26 27 27 27 26 27 27 38  0
Screwworm Fly      33 32 24 24 25 26 23 26 26 26 26 28 30 40 34  0
Moth               36 35 28 33 32 31 29 31 30 30 31 30 33 41 41 16  0
Bakers Mould       63 62 64 64 64 64 62 66 59 59 61 62 65 61 72 58 59  0
Bread Yeast        56 57 61 60 59 59 59 58 62 62 62 61 64 61 66 63 60 57  0
Skin Fungus        66 65 66 68 67 67 67 68 66 66 66 65 67 69 69 65 61 61 41 0
```

FIGURE 1 – Description du jeu de données de Mutations

Exercice 2. Classification hiérarchique

1. En utilisant la fonction `hclust`, effectuer la classification hiérarchique ascendante (avec les différents critères d'agrégation disponibles) des données de `Mutations`. Commenter et comparer les résultats obtenus, en vous appuyant sur la représentation obtenue par AFTD.
2. Effectuer la classification hiérarchique ascendante des données `Iris`, après calcul des distances associées (on utilisera la fonction `dist` pour ce faire). Commenter les résultats obtenus, en vous appuyant sur votre connaissance de ce jeu de données.
3. Effectuer la classification hiérarchique descendante des données `Iris`, au moyen de la fonction `diana` (bibliothèque `cluster`). Comparer aux résultats obtenus au moyen de la CAH.

Remarque importante : dans les anciennes versions de R, il faut élever les distances au carré avant d'effectuer une CAH via la fonction `hclust` *avec le critère de Ward (lorsque celui-ci a un sens : tableau de distances euclidiennes). Dans les versions les plus récentes, il existe deux critères :* `ward.D` *et* `ward.D2` *; on choisira le second (*`ward.D2`*), qui seul implémente le critère de Ward.*

Exercice 3. Méthode des centres mobiles

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

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 des données en $K = 3$ classes. Observer les résultats, en termes de partition 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 à nouveau $N = 100$ classifications en $K = 3$ classes, $K = 4$ classes, ... jusqu'à $K = 10$ classes. On constitue ainsi neuf échantillons iid $I_2, \ldots, I_{10}$ contenant chacun $N = 100$ valeurs d'inertie intra-classe, que l'on pourra stocker dans un tableau de taille 100×9.
   (b) Pour chaque valeur de K, calculer l'inertie intra-classe minimale $\widehat{I}_K = \min_{i=1,\ldots,100} I_{K\,i}$. Représenter la variation d'inertie minimale en fonction de K (on inclura l'inertie totale, que l'on peut assimiler à l'inertie intra-classe pour $K = 1$).

   Proposer un nombre de classes à partir de ces informations, en utilisant la méthode du coude.
4. Comparer les résultats de la partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes.

Données Crabs

1. Effectuer plusieurs classifications en $K = 2$ classes des données `Crabs` chargées précédemment. Les résultats obtenus sont-ils toujours les mêmes ? À quoi correspondent-ils ?
2. Effectuer une classification en $K = 4$ classes des données. Comparer à la partition réelle suivant l'espèce et le sexe. Que peut-on conclure ?

Données Mutations

On calculera tout d'abord une représentation des données `mutations` dans un espace de dimension $d = 5$, sur laquelle on pourra utiliser la fonction `kmeans`.

1. Effectuer plusieurs classifications de cette représentation en $K = 3$ classes au moyen de l'algorithme des centres mobiles. On pourra représenter les résultats obtenus dans le premier plan factoriel de l'AFTD.
2. Étudier la stabilité du résultat de la partition. Commenter et interpréter.