Université Paris Descartes
UFR de Mathématiques et Informatique

Classification
Examen partiel du 17/11/2015

Préalables : Ouvrir Rstudio, puis entrer les commandes suivantes :

```
install.packages("ade4")
install.packages("sp")
```

Si une question est posée lors de l'exécution de cette dernière commande, taper les touches "n" puis "entrée".
Puis entrer les commandes suivantes :

```
library(ade4)
library(sp)
```

On écrira toutes les commandes demandées dans l'éditeur en prenant soin d'enregistrer le fichier à la fin de la séance (peu importe le nom et l'emplacement du fichier).

Exercice 1

Calculer (à la main) la dissimilarité de Dice entre les deux vecteurs binaires suivants :

$$X_1 = (1, 0, 0, 1, 1, 0), \qquad X_2 = (0, 0, 1, 1, 1, 0).$$

Exercice 2

Le jeu de données morphosport contient des informations morphologiques sur 153 athlètes pratiquant différents sports.

1. Charger les données avec la commande data("morphosport"), puis observer les données pour comprendre comment elles sont organisées.
2. Effectuer des classifications ascendantes hiérarchiques sur le tableau de données quantitatives en faisant varier les critères de dissimilarité (euclidien, euclidien normalisé, Mahalanobis) et en utilisant le critère d'agrégation de Ward. Afficher les trois dendrogrammes correspondant sur une même fenêtre graphique.
3. Regarder l'aide de la fonction scale et utiliser cette fonction pour recentrer et normaliser les données. Reprendre ensuite les commandes de la question 1 en les exécutant sur les données ainsi transformées. Qu'observez-vous ? Expliquer.
4. On choisit finalement d'utiliser la dissimilarité de Mahalanobis avec le critère de Ward. Quel nombre K de classes vous semble optimal en observant le dendrogramme ? Effectuer la classification pour ce nombre de classes K choisi.
5. A l'aide de la fonction table, comparer la classification obtenue avec la répartition des athlètes suivant les différents sports. Commenter.

6. Effectuer à présent une classification par k-means sur les données en utilisant le même nombre K de classes choisi précédemment. Comparer cette classification avec la répartition en différents sports, ainsi qu'avec la classification obtenue par la CAH.

Il est courant en classification non supervisée d'effectuer une classification mixte qui combine les deux approches (CAH et k-means). La méthode est la suivante :

- On effectue d'abord une classification par k-means en choisissant un nombre assez élevé de classes. On appelle ces classes classes intermédiaires.
- On effectue ensuite une classification hiérarchique sur les données constituées des barycentres des classes obtenues par k-means, puis on choisit un nombre de classes optimal. On appelle ces classes classes finales.
- On récupère la classification finale en affectant à chaque individu la classe obtenue en combinant les deux classifications : chaque individu est affecté à la classe correspondant à la classe finale du barycentre de sa classe intermédiaire.

Cette méthode présente un avantage pour des données de grandes dimension pour lesquelles une classification hiérarchique directe serait trop lente. Même si ce n'est pas le cas ici nous allons tester cette approche.

7. Ecrire le code permettant d'effectuer les trois étapes de la classification mixte. On choisira 50 comme nombre de classes pour la classification par k-means, puis le même nombre K de classes finales choisi auparavant.
8. Comparer les résultats avec les précédents et avec la répartition par sports.