Université Paris Descartes
UFR de Mathématiques et Informatique

Classification Master 1 Ingénierie Mathématique
Examen partiel du 21 mars 2019

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). On pourra aussi enregistrer les résultats graphiques dans des fichiers pdf (Menu "Export" puis "Save as pdf").

Exercice 1 (à faire sur feuille)
Soit le tableau de données suivant (5 individus, deux variables quantitatives)

	e_1	e_2	e_3	e_4	e_5
X^1	1.5	4	0	0	4
X^2	0	1	0	1	1.5

1. Représenter les observations sur un graphique avec X^1 et abscisse et X^2 en ordonnée.
2. Donner les différentes étapes de la classification ascendante hiérarchique de ces données pour la distance euclidienne et avec la stratégie d'agrégation du minimum.

Exercice 2
Le jeu de données `ecomor` de la librairie `ade4` regroupe plusieurs tableaux de données donnant des informations écologiques et morphologiques sur 129 espèces d'oiseaux. On va ici travailler principalement avec les tableaux `forsub`, `diet`, `habitat` et `morpho`.

1. Charger la librairie et le jeu de données. Consulter l'aide sur ces données et expliquer ce que contiennent ces quatre tableaux. Parmi ces quatre tableaux, dire lesquels contiennent des variables quantitatives et lesquels contiennent des variables qualitatives.

On s'intéresse tout d'abord aux données morphologiques : `morpho`.

2. Rappeler la définition des distances euclidiennes, euclidiennes normalisées et de Mahalanobis pour des observations $X_i \in \mathbb{R}^p$ (on écrira la formule de $d(X_i, X_j)$ dans chaque cas).
3. A l'aide de la fonction `dist.quant`, calculer les distances euclidiennes, euclidiennes normalisées et Mahalanobis entre les espèces d'oiseaux sur ces données.
4. Effectuer des classifications ascendantes hiérarchiques (CAH) sur la matrice des dissimilarités, en comparant plusieurs méthodes d'agrégation. Afficher les différents dendrogrammes obtenus sur une même figure.
5. On décide d'effectuer un partitionnement avec la stratégie d'agrégation de Ward et en utilisant les distances de Mahalanobis. Quel nombre de classes correspond alors à un saut maximal du critère d'agrégation ? Effectuer la classification pour ce nombre de classes.
6. Normaliser le tableau de données avec la fonction `scale`, puis effectuer une classification sur ces données normalisées par la méthode des k-moyennes en choisissant le même nombre de classes que précédemment, et avec les options par défaut. Quelle est précisément l'opération réalisée par cette fonction `scale`, et quel est l'intérêt de l'utiliser avant d'effectuer les k-moyennes ?

7. Comparer les classes obtenues avec les deux méthodes (CAH et k-moyennes). Obtient-on des partitionnements similaires ?
8. Comparer également ces deux partitionnements avec le partitionnement des espèces en différents ordres (information contenue dans le tableau `taxo`). Commenter le résultat.

On s'intéresse à présent aux données écologiques : `forsub`, `diet`, et `habitat`.

9. A l'aide de la fonction `data.frame`, regrouper ces trois tableaux en un seul.
10. Quels types de dissimilarités peut-on utiliser pour de telles données ?
11. Effectuer des CAH en comparant plusieurs dissimilarités et plusieurs méthodes d'agrégation. Afficher les différents dendrogrammes obtenus sur une même figure.
12. Choisir une dissimilarité ainsi qu'une méthode d'agrégation, et effectuer le partitionnement avec le même nombre de classes que précédemment. Comparer le partitionnement obtenu avec les partitionnements obtenus avec les données morphologiques, et commenter le résultat.

Enfin nous allons chercher à réaliser un partitionnement global sur l'ensemble des variables morphologiques et écologiques.

13. Pour définir une dissimilarité sur toutes les variables, on va simplement faire la somme des dissimilarités sur les données morphologiques et écologiques. Montrer que cette somme vérifie bien les propriétés d'une dissimilarité telle que définie en cours.
14. Calculer cette somme, puis, comme précédemment, effectuer à nouveau des CAH en comparant plusieurs méthodes, puis enfin choisir une méthode et un nombre de classes qui semble optimal.
15. Plutôt que faire une simple somme, on pourrait effectuer une somme pondérée des dissimilarités. Quel serait l'intérêt ? Comment pourrait-on choisir les poids ?