Université Paris Cité
UFR de Mathématiques et Informatique

Classification Master 1 IMB
Examen partiel du 9 mars 2023 - durée 1h30

Partie 1 (épreuve sans documents ni outils informatiques, à rédiger sur copie)

Exercice 1

1. Rappeler les formules permettant de définir les dissimilarités suivantes : distance euclidienne, distance euclidienne normalisée et distance de Mahalanobis. On précisera bien les notations utilisées dans les formules.
2. Pour quel type de données les dissimilarités précédentes peuvent être utilisées ? Citer d'autres dissimilarités pouvant être utilisées lorsque les données ne sont pas du même type.

Exercice 2

On considère les données suivantes :

Observation	X^1	X^2	X^3	Y
e_1	3	4	0	1
e_2	2	2	3	0
e_3	5	4	1	1
e_4	1	2	2	0
e_5	4	4	1	0
e_6	4	5	1	

Effectuer une classification aux k plus proches voisins à partir des observations $e_1, \ldots, e_5$, afin de classer e_6 pour la variable cible Y, en utilisant la distance euclidienne, pour $k = 1$, $k = 3$ et $k = 5$.

Partie 2 (à réaliser avec Rstudio, tous documents autorisés)

On écrira tous les codes R dans un document `Nom_Prenom.R` ou `Nom_Prenom.Rmd`. Les packages suivants seront utilisés : `ade4`, `cluster.datasets`. Utiliser la commande `library` pour charger ces packages. Si un ou plusieurs de ces packages ne sont pas disponibles, utiliser la commande `install.packages` pour les installer.

Exercice 3

Dans cet exercice on va travailler avec les données `life.expectancy.age.sex.1971` du package `cluster.datasets` On va chercher à comparer différentes méthodes de partitionnement à partir de ces données : la classification acendante hiérarchique (CAH) et les k-moyennes.

1. Charger les données et consulter l'aide associée. Expliquer ce que représente ce tableau de données et les différentes variables.
2. Créer un sous-tableau de données `tab` contenant uniquement les variables d'intérêt pour la classification (justifier le choix des variables conservées).
3. Dans la suite on va uniquement travailler avec la distance euclidienne pour ces données. Est-ce que ce choix de dissimilarité vous semble justifié ?
4. Calculer la matrice des distances euclidiennes pour les données `tab` , puis effectuer des classifications ascendantes hiérarchiques en testant plusieurs stratégies d'agrégation. Afficher le dendrogramme obtenu dans chaque cas.
5. Expliquer pourquoi il ne serait pas intéressant de choisir la partition obtenue par le critère du saut maximal. Dans la suite, on décide plutôt d'effectuer toujours des partitionnements en 4 classes.
6. Choisir une des stratégies testées précédemment, puis calculer la classification obtenue avec un partitionnement 4 classes pour ce choix de CAH.
7. Effectuer également une classification à l'aide des k-moyennes pour ces mêmes données, en prenant $k = 4$, et avec les paramètres par défaut.
8. Afficher une table de contingence pour les deux partitionnements obtenus par CAH et k-moyennes, et commenter le résultat.

Une méthode possible pour comparer quantitativement deux partitions P_1, P_2 de $\{e_1, \ldots, e_n\}$ consiste à calculer l'*indice de Rand*. On définit d'abord les quantités suivantes :

— a = nombre de couples d'observations (e_i, e_j), $i < j$ tels que e_i et e_j appartiennent à la même classe dans la partition P_1, et également dans P_2,

— b = nombre de couples d'observations (e_i, e_j), $i < j$ tels que e_i et e_j appartiennent à la même classe dans la partition P_1, mais pas dans P_2,

— c = nombre de couples d'observations (e_i, e_j), $i < j$ tels que e_i et e_j appartiennent à la même classe dans la partition P_2, mais pas dans P_1,

— d = nombre de couples d'observations (e_i, e_j), $i < j$ tels que e_i et e_j sont dans des classes différentes pour P_1, et aussi pour P_2.

L'indice de Rand est alors calculé par la formule $RI(P_1, P2) = \frac{a+d}{a+b+c+d}$.

9. Que vaut $a + b + c + d$?
10. Des partitions similaires correspondent-elles à un indice de Rand élevé ou bas ?
11. Ecrire une fonction `IndiceRand = function(cl1, cl2)` qui prend en entrée deux vecteurs de classification (donnant la classe de chaque observation), et renvoie l'indice de Rand associé, calculé à partir de la définition précédente. Tester cette fonction sur les deux partitions obtenues précédemment.