Contrôle écrit - Apprentissage non supervisé - Clustering

Durée : 1h45

Documents non autorisés, Calculettes autorisées

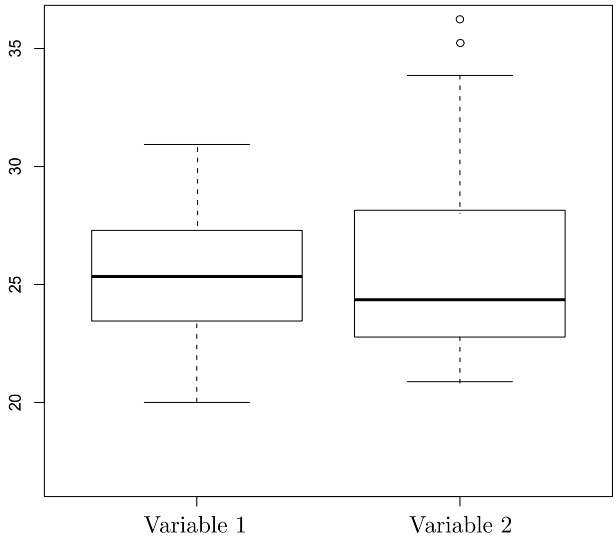
Répondre directement sur les feuilles

Nom :

Prénoms :

Questions de cours (7 points)

1. Décrivez les distributions des variables définies par les diagrammes en boîte suivants :



2. En général, peut-on dire que si deux variables quantitatives ont un coefficient de corrélation linéaire proche de 0 alors elles sont indépendantes ? Si deux variables quantitatives ont un coefficient de corrélation proche de 1 ou de -1, peut-on dire qu'elles dépendent linéairement l'une de l'autre ?

3. Expliquer comment se comportent l'inertie intra-classe I_W des données et l'inertie inter-classes I_B au cours des itérations d'un algorithme classification ascendante hiérarchique.

4. Quel lien existe entre l'algorithme des k-means et l'algorithme de classification ascendante hiérarchique utilisé avec le critère de Ward ?

5. La méthode de classification spectrale et la méthode des k-means sont toutes les deux dédiées à la classification automatique. Citer un point commun à ces deux méthodes ainsi qu'une de leur différence.

6. Décrire le(s) objectif(s) visé(s) par la méthode des cartes auto-organisatrices de Kohonen.

7. Citer les similitudes et les différences qui existent entre l'algorithme SOM (Self Organizing Map) et la version séquentielle des k-means.

8. Si l'objectif visé est la classification, l'algorithme SOM suffit-il à effectuer cette tâche ? Sinon comment ce dernier peut-il être complété pour atteindre cet objectif ?

9. D'un point de vue théorique, est-il possible de trouver la partition réalisant l'optimum global de l'inertie intra-classes ? Justifier votre réponse. Quelle approche pratique est souvent adoptée, notamment lorsqu'on utilise l'algorithme des k-means ?

Exercice 1 (6 points)

Considérons la matrice de données suivante constituée de 5 individus $\Omega = \{x_1, x_2, x_3, x_4, x_5\}$ décrits par 2 variables numériques

$$X = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 4 & 3 \\ 5 & 4 \\ 5 & 1 \end{pmatrix}$$

Le centre de gravité de l'ensemble des 5 individus est

$$g = (3.2, 2.8).$$

Les distances euclidiennes entre les individus et le centre de gravité **g** sont consignées dans le tableau suivant :

$d^2(g, x_1)$	6.28
$d^2(g, x_2)$	5.48
$d^2(g, x_3)$	0.68
$d^2(g, x_4)$	4.68
$d^2(g, x_5)$	6.48

1. Calculer l'inertie I de l'ensemble des données.

2. Soit Ω_1 l'ensemble des points dont la variable 1 est inférieure ou égale à celle du centre de gravité et Ω_2 l'ensemble des points dont la variable 1 est supérieure à celle du centre de gravité. Calculer les inerties I_1 et I_2 de ces deux sous ensembles.
N.B. : les résultats seront donnés avec une précision de deux chiffres après la virgule.

3. Exprimer I en fonction de I_1 et I_2.

4. Les cinq individus sont distants les uns des autres selon le tableau des distances euclidiennes suivant :

	x_1	x_2	x_3	x_4	x_5
x_1	0	....	3,16	....	....
x_2		0	3,16	4,47	4,12
x_3			0	1,41	2,24
x_4				0	....
x_5					0

a. Compléter les valeurs manquantes du tableau.

b. Préciser quels sont les deux points les plus éloignés les uns des autres et dans quelle mesure ils interviennent dans l'inertie I.

c. Compléter les tableaux de distances successifs ci-dessous en utilisant le critère d'agrégation du lien minimum

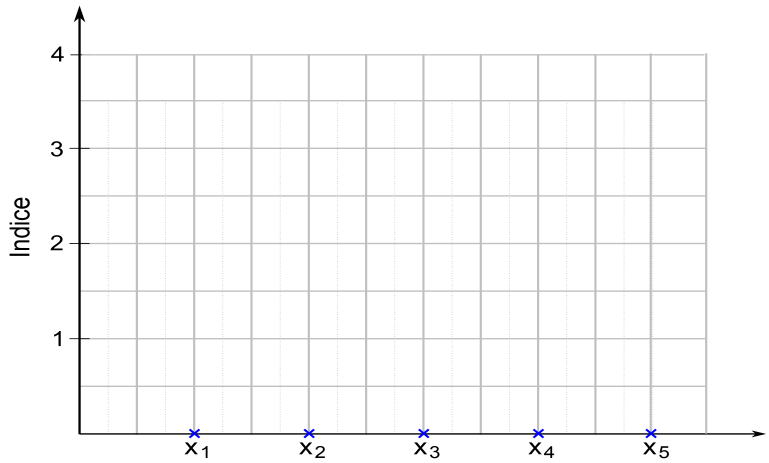
$$D_{min}(A, B) = \min_{x_a \in A\,,\, x_b \in B} d(x_a, x_b).$$

puis construire la hiérarchie (dendrogramme) associée.

	{.......}	{.......}	{.......}	{.......}
{.......}	0	....	....	....
{.......}		0	....	....
{.......}			0	....
{.......}				0

	{............}	{............}	{............}
{............}	0	....	....
{............}		0	....
{............}			0

	{.................}	{.................}
{.................}	0	....
{.................}		0



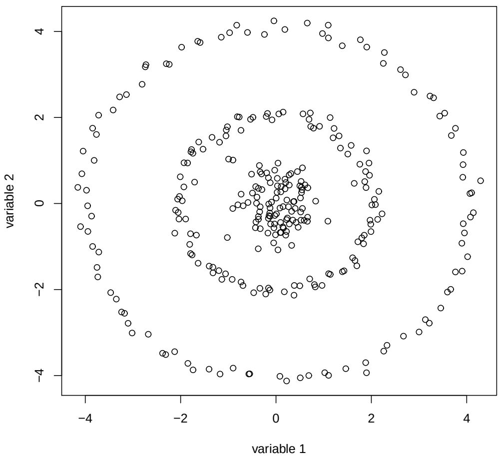
d. En déduire une partition des données en deux classes et une partition en trois classes.

5. Calculer (sous forme matricielle) l'ultramétrique associée à la hiérarchie indicée obtenue en 4.c).

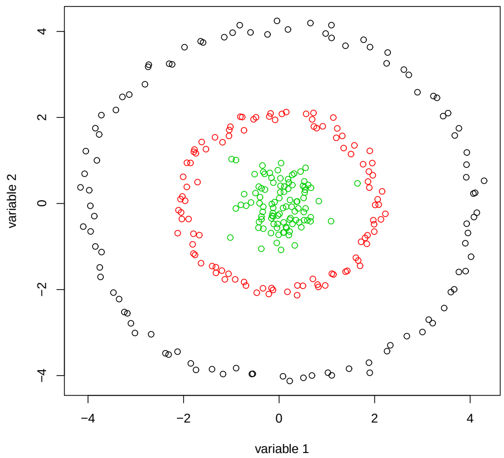
Exercice 2 (4 points)

On considère l'ensemble des 300 observations suivantes, décrites par deux variables numériques et constitué de trois classes.

Données

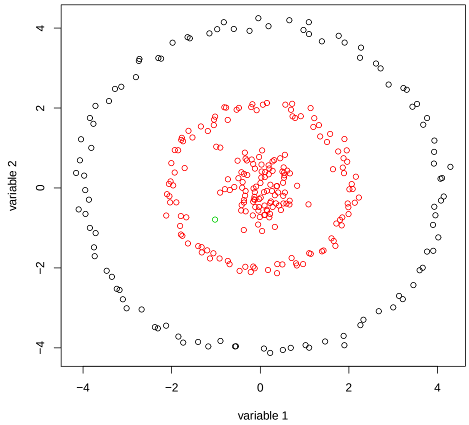


Vraies classes

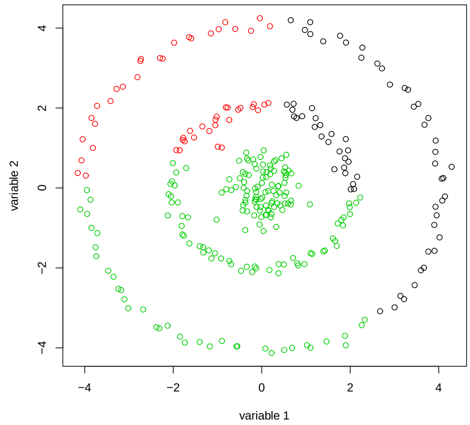


Quatre algorithmes de clustering ont été lancés sur ce jeu de données : l'algorithme de classification ascendante hiérarchique avec le critère du lien minimum (CAH-min) et le critère de Ward (CAH-moyen), l'algorithme de classification spectrale, l'algorithme des k-means. Les résultats obtenus sont donnés dans les figures ci-dessous.

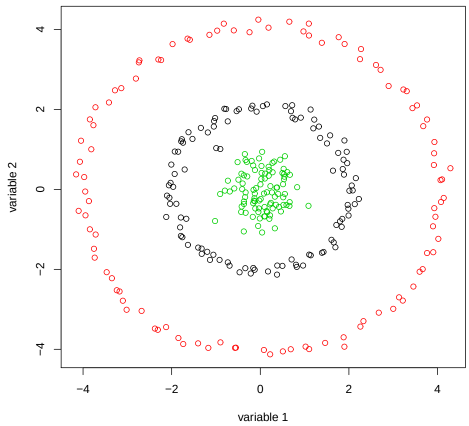
CAH-min



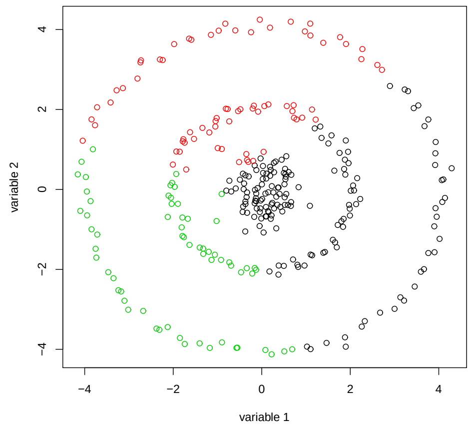
CAH-Ward



Classification spectrale



kmeans



1. Interpreter les partitions obtenues par chacune des quatres méthodes.

Exercice 3 (3 points)

On considère un ensemble de 5 observations $\{x_1, x_2, x_3, x_4, x_5\}$ décrites par le graphe (ou la matrice) de similarité

$$W = \begin{bmatrix} 1.0000 & 0.6065 & 0.2865 & 0.1353 & 0.0439 \\ 0.6065 & 1.0000 & 0.2865 & 0.0821 & 0.1194 \\ 0.2865 & 0.2865 & 1.0000 & 0.7788 & 0.5353 \\ 0.1353 & 0.0821 & 0.7788 & 1.0000 & 0.3247 \\ 0.0439 & 0.1194 & 0.5353 & 0.3247 & 1.0000 \end{bmatrix}.$$

A partir de ce graphe de similarité, on a calculé la matrice des degrés ainsi que la matrice Laplacienne qui sont données par

$$D = \begin{bmatrix} 2.0723 & 0 & 0 & 0 & 0 \\ 0 & 2.0946 & 0 & 0 & 0 \\ 0 & 0 & 2.8871 & 0 & 0 \\ 0 & 0 & 0 & 2.3209 & 0 \\ 0 & 0 & 0 & 0 & 2.0233 \end{bmatrix} \quad L = \begin{bmatrix} 1.0723 & -0.6065 & -0.2865 & -0.1353 & -0.0439 \\ -0.6065 & 1.0946 & -0.2865 & -0.0821 & -0.1194 \\ -0.2865 & -0.2865 & 1.8871 & -0.7788 & -0.5353 \\ -0.1353 & -0.0821 & -0.7788 & 1.3209 & -0.3247 \\ -0.0439 & -0.1194 & -0.5353 & -0.3247 & 1.0233 \end{bmatrix}.$$

La calcul des valeurs propres et des vecteurs propres de la matrice Laplacienne a fourni les résultats suivants :

Valeurs propres de L

0.0000	0.7385	1.4418	1.7058	2.5121

Vecteurs propres de L (rangés par colonne)

-0.4472	0.5519	0.0963	-0.6929	0.0772
-0.4472	0.5197	-0.2454	0.6789	0.0937
-0.4472	-0.2103	0.1975	0.0544	-0.8449
-0.4472	-0.3595	0.6430	0.1458	0.4859
-0.4472	-0.5019	-0.6914	-0.1863	0.1880

A partir de toutes ces informations, déterminer géométriquement une partition des données en $K = 2$ classes par l'algorithme de classification spectrale non normalisé.