Contrôle écrit - Apprentissage non supervisé - Clustering

Durée : 2h00
Documents non autorisés, Calculettes autorisées
Répondre directement sur les feuilles

Nom :
Prénoms :

Questions de cours (6 points)

1. Expliquer comment se comportent l'inertie intra-classes I_W et l'inertie inter-classes I_B au cours des itérations de l'algorithme de classification ascendante hiérarchique.

2. Quel point commun existe t-il entre l'algorithme des k-means et l'algorithme de classification ascendante hiérarchique utilisé avec le critère de Ward ?

3. Préciser de quelle façon la classification spectrale diffère des autres approches itératives de classification automatique.

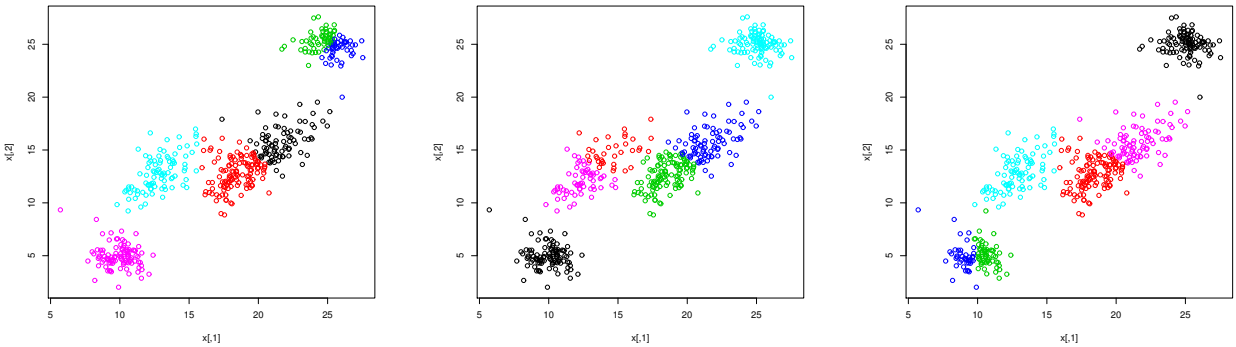
4. Quelle heuristique pourrait être associée aux algorithmes de classification spectrale pour choisir le nombre de classes ?

5. Décrire le(s) objectif(s) visé(s) par la méthode des cartes auto-organisatrices de Kohonen. Préciser comment les résultats fournis par cet algorithme peuvent être complétés pour répondre effectivement à un objectif de classification.

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

7. Sur un jeu de données bidimensionnel, trois lancers consécutifs de l'algorithme des k-means avec 6 classes ont conduit aux partitions suivantes :

lancer 1 lancer 2 lancer 3



Expliquer pourquoi les partitions obtenues diffèrent d'un lancer à l'autre de l'algorithme. Quelle solution est généralement adoptée pour palier ce problème ?

Exercice 1 (6 points)

Considérons le tableau de données suivant constitué de 7 individus $\Omega = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ décrits par 2 variables quantitatives

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 3 & 2 \\ 3 & 5 \\ 2 & 7 \\ 5 & 7 \\ 8 & 3 \end{pmatrix}$$

Le centre de gravité de l'ensemble des 7 individus est donné par $\mathbf{g} = (3.29, 3.86)$.
On donne également les deux tableaux de distance suivants :

Distances euclidiennes (au carré) par rapport au centre de gravité $\mathbf{g}$

$d^2(\mathbf{g}, x_1)$	13.42
$d^2(\mathbf{g}, x_2)$	8.70
$d^2(\mathbf{g}, x_3)$	3.54
$d^2(\mathbf{g}, x_4)$	1.38
$d^2(\mathbf{g}, x_5)$	11.52
$d^2(\mathbf{g}, x_6)$	12.78
$d^2(\mathbf{g}, x_7)$	22.92

Distances euclidiennes croisées

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0	1	2.24	4.47	6.08	7.21	7.28
x_2	1	0	2	3.61	5.10	6.40	7.07
x_3	2.24	2	0	3	5.10	5.39	5.10
x_4	4.47	3.61	3	0	2.24	2.83	5.39
x_5	6.08	5.10	5.10	2.24	0	3	7.21
x_6	7.21	6.40	5.39	2.83	3	0	5
x_7	7.28	7.07	5.10	5.39	7.21	5	0

1. Calculer l'inertie I de l'ensemble Ω.

2. En vous appuyant sur le critère d'agrégation du lien minimum, construire la hiérarchie indicée (dendrogramme) associée aux données.

3. En déduire une partition des données en trois classes P_1, P_2, P_3.

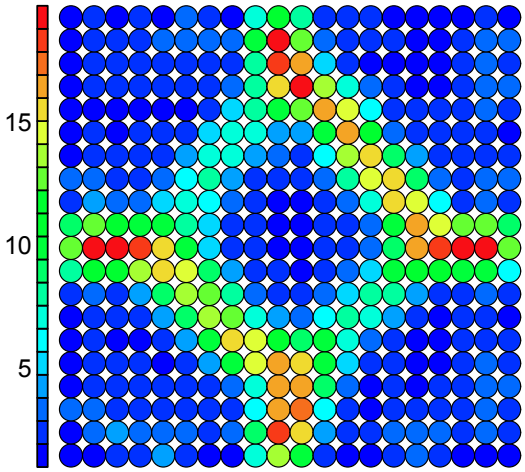
4. Exprimer sans faire de calcul numérique l'inertie globale I en fonction des inerties I_1, I_2 et I_3 des classes P_1, P_2, P_3.

5. Calculer (sous forme matricielle) l'ultramétrique associée à la hiérarchie indicée obtenue en 2.).

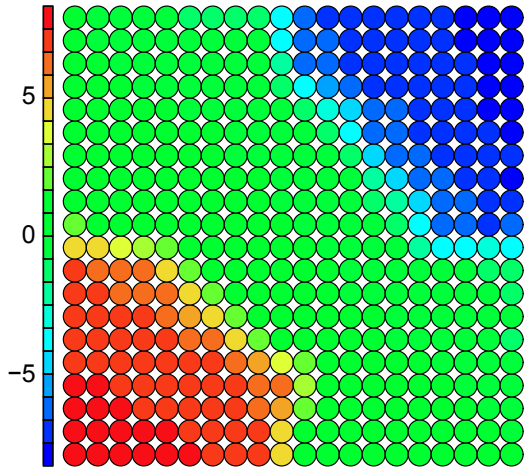
Exercice 2 (5 points)

Sur un jeu de données décrites par deux variables quantitatives et constitué de **classes sphériques de même proportion**, on a lancé l'algorithme SOM (Self Organizing Map) avec une grille 20×20. Les résultats obtenus sont donnés par les trois cartes ci-dessous (les couleurs indiquent les valeurs associées aux nœuds de la grille).

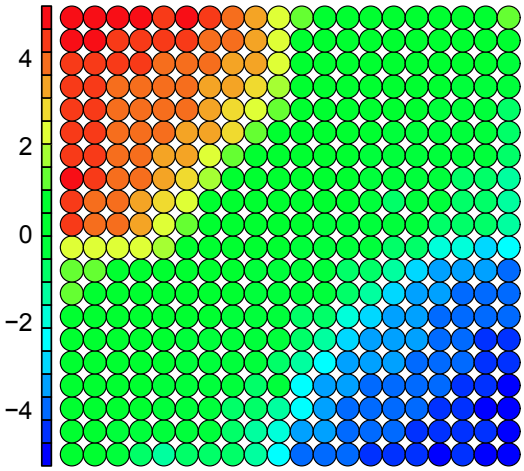
U-matrix



Component planes : variable 1



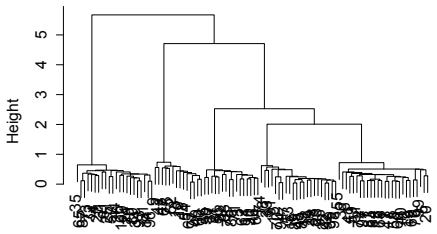
Component planes : variable 2



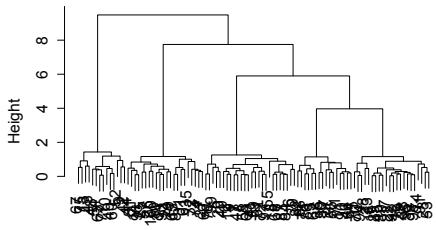
1. Que nous renseigne ces cartes sur la topologie des données ?

Sur le même jeu de données, l'algorithme de classification ascendante hiérarchique (CAH) a été lancé avec les critères d'agrégation du lien minimum et du lien moyen. Ensuite, l'algorithme des k-means a été lancé en faisant varier le nombre de classes de 2 à 10. Le dendrogramme et la courbe de variation de l'inertie intra-classes sont donnés ci-dessous.

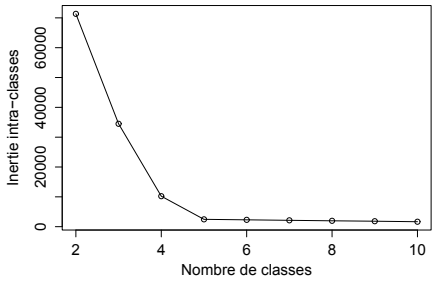
CAH - Lien minimum



CAH - Lien moyen



k-means



2. Quel(s) nombre(s) de classes sont suggérés par ces trois méthodes. Commentez.

3. Proposer, en justifiant vos choix, une configuration géométrique de ce jeu de données dans $\mathbb{R}^2$.

Exercice 3 (3 points)

On considère un ensemble de 8 observations $\Omega = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ décrites par le graphe (ou la matrice) de similarité suivants :

$$W = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.14 & 0 & 0.37 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.37 & 0 & 0.61 \\ 0 & 0 & 1 & 0.61 & 0.01 & 0.02 & 0 & 0 \\ 0 & 0 & 0.61 & 1 & 0 & 0 & 0 & 0 \\ 0.14 & 0 & 0.01 & 0 & 1 & 0.01 & 0.37 & 0.01 \\ 0 & 0.37 & 0.02 & 0 & 0.01 & 1 & 0 & 0.61 \\ 0.37 & 0 & 0 & 0 & 0.37 & 0 & 1 & 0 \\ 0 & 0.61 & 0 & 0 & 0.01 & 0.61 & 0 & 1 \end{bmatrix}.$$

A partir de ce graphe de similarité, on a obtenu la matrice des degrés ainsi que la matrice Laplacienne suivants :

$$D = \begin{bmatrix} 1.51 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.98 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.61 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.54 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.01 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.74 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.23 \end{bmatrix} \quad L = \begin{bmatrix} 0.51 & 0 & 0 & 0 & -0.14 & 0 & -0.37 & 0 \\ 0 & 0.98 & 0 & 0 & 0 & -0.37 & 0 & -0.61 \\ 0 & 0 & 0.64 & -0.61 & -0.01 & -0.02 & 0 & 0 \\ 0 & 0 & -0.61 & 0.61 & 0 & 0 & 0 & 0 \\ -0.14 & 0 & -0.01 & 0 & 0.54 & -0.01 & -0.37 & -0.01 \\ 0 & -0.37 & -0.02 & 0 & -0.01 & 1.01 & 0 & -0.61 \\ -0.37 & 0 & 0 & 0 & -0.37 & 0 & 0.74 & 0 \\ 0 & -0.61 & 0 & 0 & -0.01 & -0.61 & 0 & 1.23 \end{bmatrix}.$$

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

Valeurs propres de L rangées par ordre croissant

0	0.01	0.02	0.67	1.12	1.23	1.37	1.84

Vecteurs propres associés aux valeurs propres (chaque colonne correspond à un vecteur propre)

0.35	0.46	-0.09	0.70	0.40	0.01	0.00	0.00
0.35	-0.20	0.42	0.01	-0.01	0.05	-0.71	0.40
0.35	-0.37	-0.48	0.00	0.02	-0.71	-0.06	-0.01
0.35	-0.38	-0.49	0.01	-0.02	0.70	0.05	0.00
0.35	0.43	-0.08	-0.71	0.43	0.02	-0.01	0.00
0.35	-0.20	0.40	0.00	0.01	-0.05	0.70	0.42
0.35	0.45	-0.09	-0.03	-0.81	-0.02	0.01	0.00
0.35	-0.20	0.41	0.01	0.00	0.01	0.02	-0.82

A partir de toutes ces informations, quel nombre de classes semble être le plus adapté aux données ? En déduire géométriquement (sans faire de calculs) une partition des données via l'algorithme de classification spectrale non normalisé.