Contrôle écrit - Apprentissage non supervisé - Clustering

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

Nom :

Prénoms :

Questions de cours (5 points)

1. Dans une boîte à moustaches (appelée aussi diagramme en boîte), quel est le pourcentage de données qui se situe :

(a) dans l'intervalle compris entre les deux extrémités de la boîte

(b) entre $-\infty$ et l'extrémité supérieure de la boîte

2. Répondre par vrai ou faux aux questions suivantes.

(a) La variance est une mesure de dispersion autour de la médiane.
☐ VRAI ☐ FAUX

(b) La covariance est une mesure de dépendance entre deux variables.
☐ VRAI ☐ FAUX

(c) En général, si deux variables ont un coefficient de corrélation nul, alors elles sont indépendantes.
☐ VRAI ☐ FAUX

(d) Si deux variables dépendent linéairement l'une de l'autre, alors leur coefficient de corrélation est proche de 1 ou de -1.
☐ VRAI ☐ FAUX

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

4. Si l'objectif visé est la classification, les résultats généralement fournis par l'algorithme SOM (Self Organizing Map) suffisent-ils à effectuer cette tâche ? Sinon comment ces derniers peuvent être complétés pour atteindre cet objectif ?

5. A partir d'une partition floue $\mathbf{C} = (c_{ik})$ de n individus en K classes, comment peut-on définir une partition $\mathbf{Z} = (z_{ik})$.

6. Généralement, la partition fournie par la méthode des nuées dynamiques résulte t-elle de l'optimisation locale ou globale d'un critère d'inertie ?

7. Préciser la différence entre la classification spectrale et les méthodes itératives usuelles de partitionnement.

8. Proposer une heuristique qui pourrait être associée aux algorithmes de classification spectrale pour choisir le nombre de classes ?

Exercice 1 (4 points)

On considère un ensemble de 5 observations distantes les unes des autres selon le tableau de distances euclidiennes suivant :

	x_1	x_2	x_3	x_4	x_5
x_1	0	2	7.5	8.5	10
x_2	2	0	5.5	6.5	8
x_3	7.5	5.5	0	1	2.5
x_4	8.5	6.5	1	0	1.5
x_5	10	8	2.5	1.5	0

1. En utilisant le critère d'agrégation du lien minimum, construire la hiérarchie indicée associée à cette matrice de distances.

2. Déduire, à partir de cette représentation, une partition des données.

3. Calculer sous la forme d'une matrice, l'ultramétrique associée à la hiérarchie indicée obtenue dans la question **1**.

Exercice 2 (6 points)

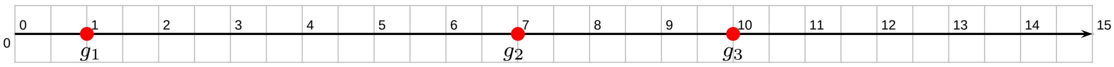
On considère l'échantillon suivant, de 6 individus décrits par 1 variable quantitative :

$$X = \begin{pmatrix} 3 \\ 11 \\ 8 \\ 5.5 \\ 13.5 \\ 0 \end{pmatrix}$$

En supposant que les données sont reçues de manière séquentielle ($x_1 = 3$ puis $x_2 = 11$ puis $x_3 = 8 \ldots$), appliquer la version séquentielle de l'algorithme des k-means pour trouver une partition des données en 3 classes. *Attention : toutes les étapes listées ci-dessous ne sont pas nécessairement utiles.*

Étape 0 : Initialisation aléatoire des centres

$$g_1 = 1 \quad g_2 = 7 \quad g_3 = 10$$

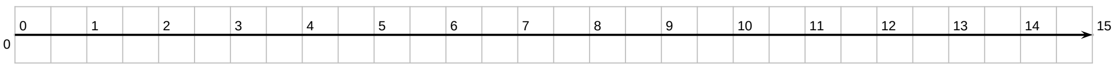


Étape 1 : · · · · · · · · ·

. . . .
. . . .
. . . .

Étape 2 : · · · · · · · · ·

. . . .
. . . .
. . . .

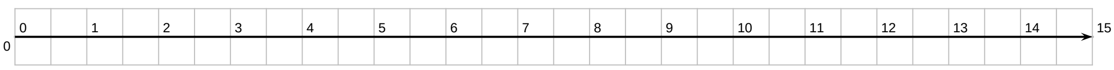


Étape 3 : · · · · · · · · ·

. . . .
. . . .
. . . .

Etape 4 : · · · · · · · · ·

. . . .
. . . .
. . . .

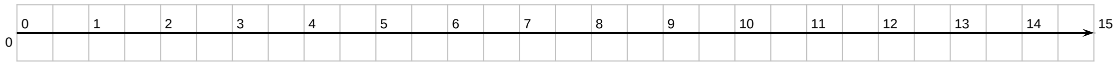


Étape 5 : ·········

....
....
....

Étape 6 : ·········

....
....
....

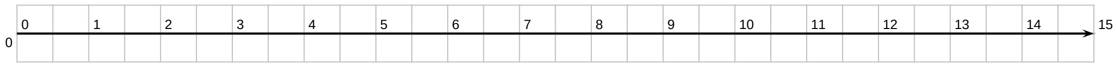


Étape 7 : ·········

....
....
....

Étape 8 : ·········

....
....
....

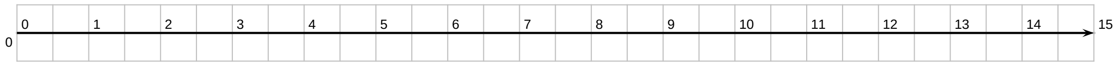


Étape 9 : ·········

....
....
....

Étape 10 : ·········

....
....
....

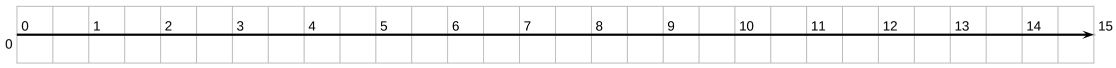


Étape 11 : ·········

....
....
....

Étape 12 : ·········

....
....
....

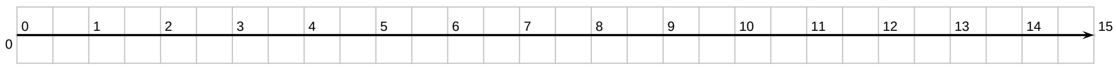


Étape 13 : ·········

....
....
....

Étape 14 : ·········

....
....
....

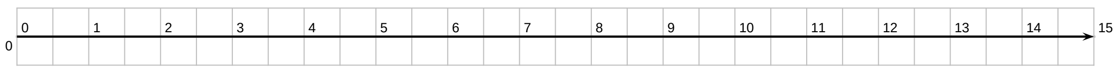


Étape 15 : ·········

....
....
....

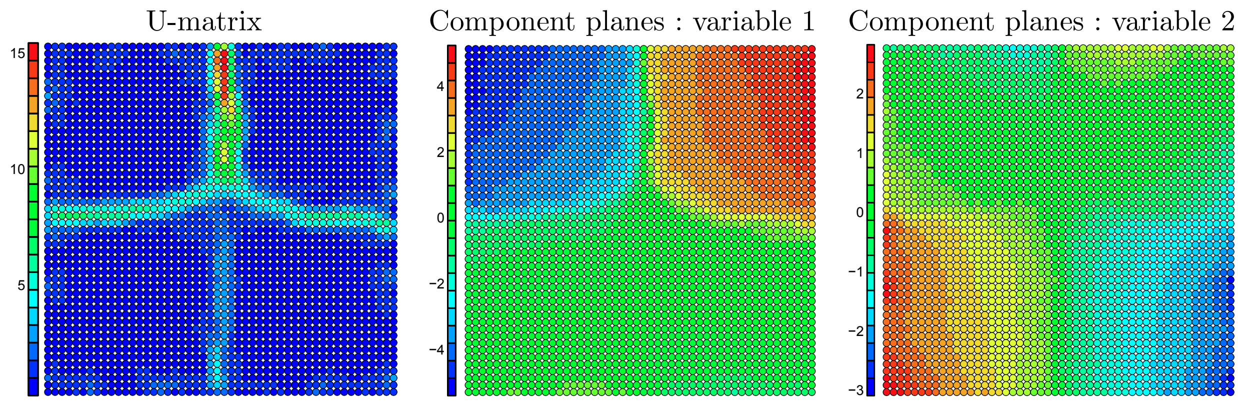
Étape 16 : ·········

....
....
....



Exercice 3 (5 points)

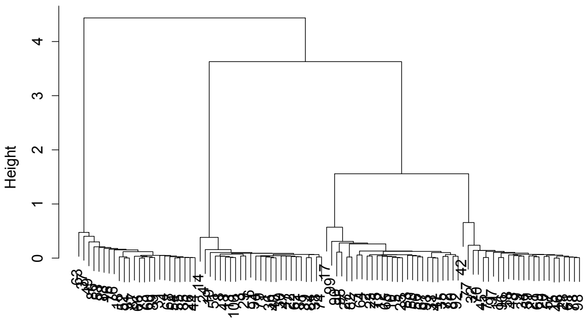
Sur un jeu de données décrites par deux variables quantitatives et constitué de **classes sphériques de même proportion**, on a lancé l'algorithme SOM (Self Organizing Map) avec une grille 50×50. Les résultats obtenus sont donnés par les trois cartes ci-dessous.



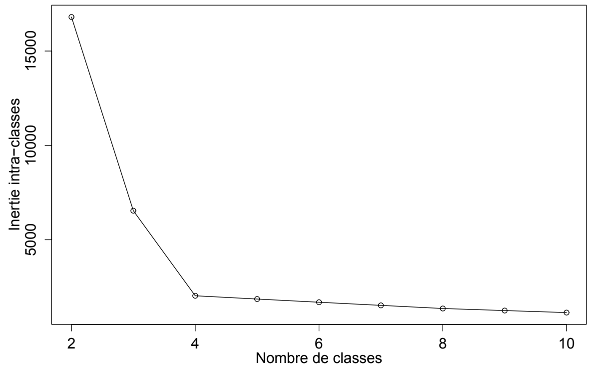
2. Donner une interprétation des résultats fournis par ces cartes.

Sur le même jeu de données, l'algorithme de classification ascendante hiérarchique (CAH) a été lancé avec le critère d'agrégation du lien minimum. 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



k-means



3. Définir le critère d'agrégation du lien minimum.

4. Quel(s) nombre(s) de classes sont suggérés par ces deux méthodes. Commentez.

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