Contrôle écrit - Clustering

Durée : 1h30
Documents non autorisés, Calculettes autorisées
Répondre directement sur les feuilles

Nom :

Prénoms :

Questions de cours

1. Préciser la différence qui existe entre la classification automatique (clustering) et la discrimination.

2. Soit $\mathbf{C} = (c_{ik})_{i=1,\ldots,n\ k=1,\ldots,K}$ une partition floue d'un ensemble de n individus $\{x_1, \ldots, x_n\}$ en K classes. Dire en quoi cette partition floue pourrait différer d'une partition classique $\mathbf{Z} = (z_{ik})$ de ce même ensemble.

3. A partir de la partition floue $\mathbf{C} = (c_{ik})$ de n individus en K classes, définir une partition $\mathbf{Z} = (z_{ik})$.

4. Citer deux critères numériques permettant de quantifier l'homogénéité d'une partition $\mathbf{Z} = (z_{ik})$ de n individus en K classes. A quelle(s) difficulté(s) peut-on être confronté dans l'optimisation de ces critères ?

5. Quelle(s) propriété(s) théorique(s) nous assure(nt) que la partition obtenue à la convergence de l'algorithme des kmeans correspond bien à un optimum (local ou global) du critère d'inertie intra-classe ?

6. Quel lien existe entre l'algorithme des kmeans et l'algorithme de classification ascendante hiérarchique utilisé avec le critère de Ward ?

7. Soit $\mathbf{\Omega}$ un ensemble de données partitionné en $K = 2$ ensembles $\mathbf{\Omega_1}$ et $\mathbf{\Omega_2}$. Les caractéristiques des ensembles $\mathbf{\Omega}$, $\mathbf{\Omega_1}$ et $\mathbf{\Omega_2}$ sont donnés par le tableau ci-dessous. Quelle relation y a t-il entre l'inertie I de $\mathbf{\Omega}$ d'une part et les inerties I_1, I_2 d'autre part ?

ensemble	effectif	centre	inertie
$\mathbf{\Omega}$	n	$\mathbf{g}$	I
$\mathbf{\Omega_1}$	n_1	$\mathbf{g}_1$	I_1
$\mathbf{\Omega_2}$	n_2	$\mathbf{g}_2$	I_2

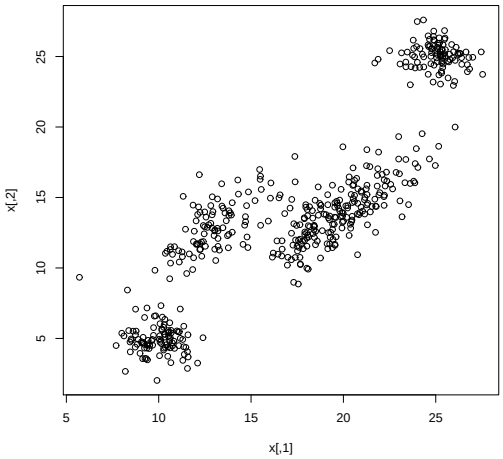
8. Expliquer le principe de la version séquentielle de l'algorithme des kmeans.

9. Citer les similitudes et les différences qui existent entre l'algorithme SOM (Self Organizing Map) et la version séquentielle des kmeans ?

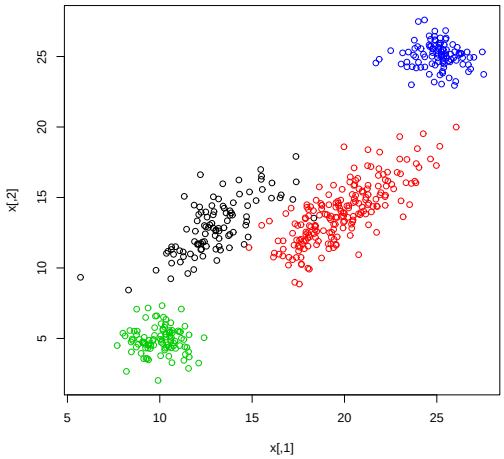
Exercice 1

On considère l'ensemble des 500 observations suivantes, décrites par deux variables numériques et constitué de quatre classes.

Données

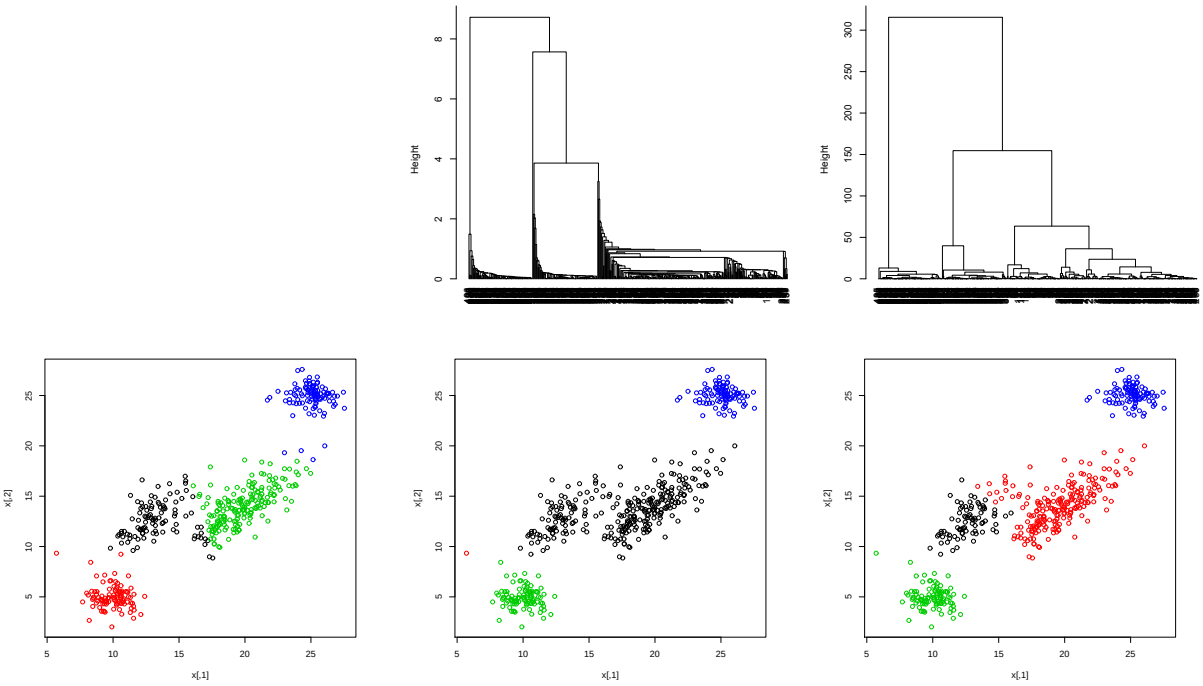


Vraies classes



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

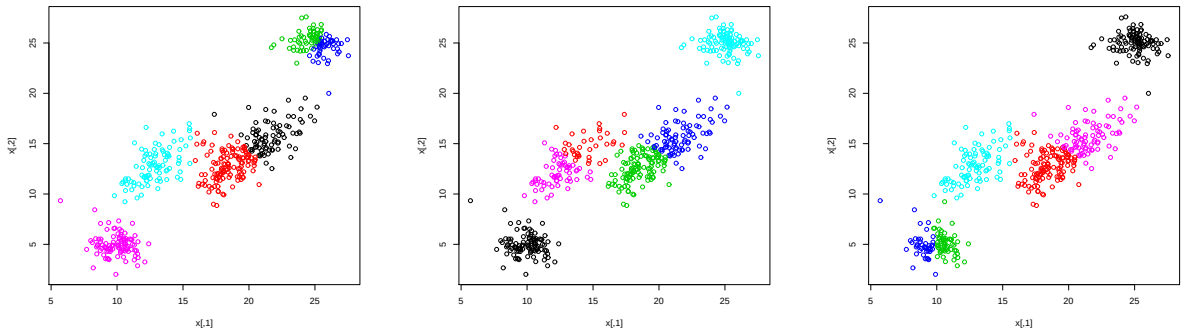
kmean CAH-min CAH-moyen



1. Interpreter les différences entre ces partitions.

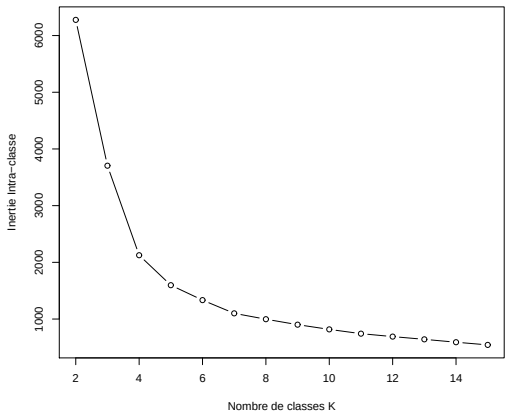
2. Afin de trouver la meilleure partition des données par l'algorithme des kmeans, celui-ci a été lancé avec une nombre de classe K variant de 1 à 15. De plus, pour chaque nombre de classes, plusieurs lancers consécutifs ont été effectués afin d'observer la stabilité de l'algorithme. Il a été fait le constat que des lancers consécutifs de l'algorithme, pour K fixé, conduisaient à des partitions différentes. Pour $K = 6$ notamment, on a obtenu les trois partitions suivantes

$K = 6$: lancer 1 $K = 6$: lancer 2 $K = 6$: lancer 3



(a) Expliquer à quoi est due cette variabilité dans les résultats et proposer une méthode pour palier ce problème.

(b) Ayant résolu le problème liés aux multiples solutions, on a finalement réussi à obtenir une unique partition pour chaque valeur de K. Afin de trouver le nombre de classes adapté, l'inertie intra-classe a été représentée en fonction du nombre de classes.



Expliquer pourquoi le choix du nombre de classes approprié par minimisation du critère d'inertie intra-classe n'est pas judicieux. Quelle solution pourrait être envisagée ?

Exercice 2

On considère un ensemble de 6 observations distantes les unes des autres selon le tableau de distances euclidiennes suivantes :

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	2	5	17	20	32
x_2	2	0	1	25	18	34
x_3	5	1	0	36	25	45
x_4	17	25	36	0	13	9
x_5	20	18	25	13	0	4
x_6	32	34	45	9	4	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 adéquate des données.

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

4. On considère maintenant l'ultramétrique suivante définie sur l'ensemble des données $\{x_1, \ldots, x_6\}$.

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	5	5	72	72	72
x_2	5	0	1	72	72	72
x_3	5	1	0	72	72	72
x_4	72	72	72	0	13	13
x_5	72	72	72	13	0	4
x_6	72	72	72	13	4	0

Représenter la hiérarchie indicée associée à cette ultramétrique.