ACT6100	Analyse de données
H2019	Série 2

Analyse de classification

1. Soit un jeu de données contenant les points $x_1 = (0, 0)$, $x_2 = (1, 0)$ et $x_3 = (5, 5)$.
   (a) Calculer la matrice des distances euclidiennes entre les individus.
   (b) Utiliser un *agglomerative hierarchichal algorithm* (ascendante) pour déterminer les groupes dans ce jeu de données. Utiliser les distances euclidiennes entre les individus et les distances simples entre les groupes.
   (c) En utilisant des poids égaux, $p_i = 1/3$, $i = 1, 2, 3$, déterminer le centre de gravité du nuage. Représenter sur un plan les trois points et le centre de gravité.
2. À partir de huit points $x_1, \ldots, x_8 \in \mathbb{R}^2$, on obtient la matrice des distances euclidiennes suivantes :

$$D_1 = \begin{bmatrix} 0 & 10 & 53 & 73 & 50 & 98 & 41 & 65 \\ & 0 & 25 & 41 & 20 & 80 & 37 & 65 \\ & & 0 & 2 & 1 & 25 & 18 & 34 \\ & & & 0 & 5 & 17 & 20 & 32 \\ & & & & 0 & 36 & 25 & 45 \\ & & & & & 0 & 13 & 9 \\ & & & & & & 0 & 4 \\ & & & & & & & 0 \end{bmatrix}.$$

   Utiliser un *agglomerative hierarchichal algorithm* pour déterminer les groupes dans ce jeu de données. Utiliser les distances euclidiennes entre les individus et les distances simples entre les groupes. Dessiner le dendrogramme. Quelle séparation semble la plus intéressante ?
3. Refaire l'exercice 1 (b), mais en utilisant l'algorithme de Ward pour déterminer les groupes. Utiliser des poids égaux. Pour chacune des étapes, donner l'inertie totale, l'inertie *within* et l'inertie *between*.
4. Une étude a été réalisée afin de déterminer l'espérance de vie des hommes et des femmes de plusieurs villes. Les résultats sont présentés dans la base de données *LifeE.csv* disponible sur le site *Moodle* du cours. Les variables sont
   — **Ville_annee** : la ville et l'année ;
   — **MXX** : l'espérance de vie d'un homme âgé de XX années ($XX = 0, 25, 50, 75$) ; et
   — **FXX** : l'espérance de vie d'une femme âgée de XX années ($XX = 0, 25, 50, 75$).
   En utilisant R et l'analyse de classification, commenter les affirmations suivantes. Justifier.
   (a) L'emplacement géographique d'une ville semble être un facteur d'homogénéité pour l'espérance de vie de ses habitants.
   (b) Malgré le fait qu'ils soient éloignés géographiquement, les habitants de la ville de Düsseldorf et les habitants de la ville de Montréal forment un groupe plus homogène que les habitants de la ville de Montréal et les habitants de la ville de Toronto.
   (c) À partir d'un cluster regroupant toutes les villes, c'est le retrait de la ville de Djakarta qui occasionnera la plus grande hausse de l'inertie *between* (I_B).
5. Un échantillon bivarié de 10 observations est partitionné en deux groupes : le groupe A_1 (4 observations) et le groupe A_2 (6 observations). Chaque point de l'échantillon reçoit le même poids, 1/10. Le centre de gravité du groupe A_1 est $g_1 = (1, 3)$, et celui de l'échantillon complet $A_1 \cup A_2$ est $g = (4, 2.4)$. De plus, l'inertie du groupe A_1 vaut $I_1 = 5$ et celle du groupe A_2 vaut $I_2 = 6$.

(a) Quel est le centre de gravité du groupe A_2 ?

(b) Calculer l'inertie totale de l'échantillon (distance euclidienne).

6. Cet exercice est la suite du dernier exercice de la série 1. Après standardisation des six variables, le jeu de données *brain* est soumis à l'algorithme de classification de Ward, fournissant le dendrogramme présenté à la Figure 1. L'apparition des jumeaux est évidente. Que représente l'axe vertical *height* du dendrogramme ? Par exemple, l'aggrégation des

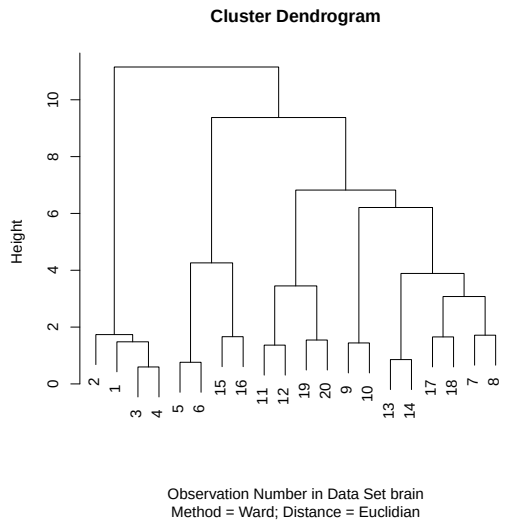

FIGURE 1 – Dendrogramme.

clusters {5,6} et {15,16} se fait à une hauteur d'environ 4. Que signifie cette valeur ?

7. Dans le cadre de l'algorithme de classification par ré-allocation dynamique, MAT8594

(a) vérifier que pour tout $j = 1, \ldots, p$, on a

$$\frac{1}{\text{card}(C_k)} \sum_{i,i' \in C_k} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} (x_{ij} - \bar{x}_{kj})^2 ,$$

où

$$\bar{x}_{kj} = \frac{1}{\text{card}(C_k)} \sum_{i \in C_k} x_{ij};$$

(b) vérifier qu'il est possible d'écrire la fonction à minimiser comme étant

$$\sum_{k=1}^{K} W(C_k) = 2 \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i \in C_k} (x_{ij} - \bar{x}_{kj})^2 ;$$

(c) déduire que la fonction à minimiser ne peut que rester ou diminuer de valeur à chaque itération de l'algorithme.

8. La base de données *USArrests* disponible dans la librairie *datasets* sur R contient des statistiques sur les crimes violents pour 50 états américains.

(a) Utiliser l'algorithme de Ward (avec poids de $1/n$, n étant la taille de la base de données [1]) avec la distance euclidienne afin d'obtenir le dendrogramme complet.

(b) À partir de la sous-question précédente, construire 10 groupes et calculer le taux moyen de meurtre par groupe.

1. Appliquer la fonction `hclust()` sur $D^2/(2n)$, où D est la matrice des distances.

(c) Utiliser maintenant un algorithme de classification par ré-allocation dynamique pour diviser la base de données en 10 groupes. Utiliser les centroïdes des groupes formés à la première sous-question comme point de départ. Vérifier que les résultats sont identiques avec les deux approches. Pourquoi ?

9. On considère les images utilisées dans l'**Application 2 : les petits chats** présentée dans les notes de cours sur l'analyse en composantes principales. Importer les images en R et les formater (200×200). Décrire/Expliquer comment on peut utiliser un algorithme de classification par ré-allocation dynamique afin de séparer cette base de données. Utiliser ensuite cet algorithme afin de diviser la base de données en trois groupes. Analyser brièvement les groupes obtenus.

Réponses

1. (a)

$$D_1 = \begin{bmatrix} 0 & 1 & \sqrt{50} \\ 1 & 0 & \sqrt{41} \\ \sqrt{50} & \sqrt{41} & 0 \end{bmatrix}.$$

   (b) On regroupe x_1 et x_2 pour ensuite ajouter x_3.

   (c) $g = (2, 5/3)$

2. Le dendrogramme est présenté à la figure 2. Une division en trois groupes semble la plus

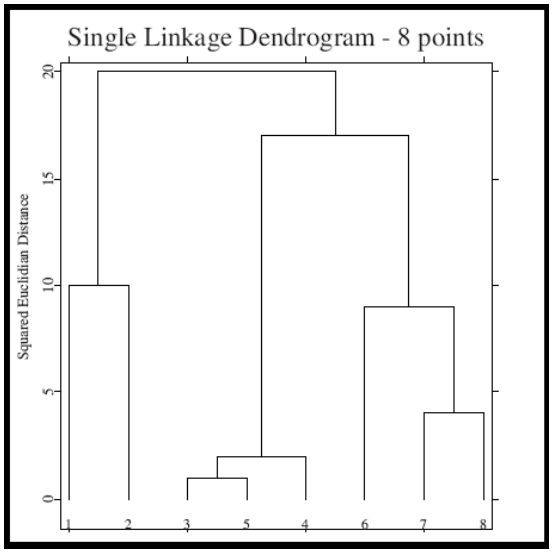

FIGURE 2 – Dendrogramme

   pertinente.

3. On regroupe x_1 et x_2 pour ensuite ajouter x_3. Les gains/pertes d'intertie sont décrits dans la Table 1.

Groupe(s)	inertie totale (I_T)	inertie *between* (I_B)	inertie *within* (I_W)
P, Q et R	10.22	10.22	0
$P \cup Q$ et R	10.22	$181/18 = 10.05$	$1/6 = 0.17$
$P \cup Q \cup R$	10.22	0	10.22

TABLE 1 – Décomposition de l'inertie

4. (a) Faux

   (b) Vrai

   (c) Vrai

5. (a) (6,2)

   (b) 17.24

6. La hauteur représente le niveau d'aggrégation, ici la perte de l'inertie *between* après l'aggrégation des clusters concernés.

7. Démonstration

8. (a) -

   (b) -

   (c) -

9. -