Module 6
Regroupement

Sommaire

Dernière mise à jour le 1er octobre 2021

6.1 Principe du regroupement

Le **regroupement**, aussi appelé **agrégation** (*clustering*), est une méthode d'apprentissage machine non supervisée (*unsupervised learning*), qui a pour objectif de former des groupes ou agrégats (*clusters*) d'objets similaires à partir d'un ensemble hétérogène d'objets (aussi appelées individus ou points).

Formellement, soit un ensemble de données formé de N objets décrits par D variables. Le regroupement consiste à trouver une répartition de ces données en K groupes de sorte que les objets à l'intérieur d'un même groupe soient similaires. Le nombre de groupes K peut être déterminé a priori.

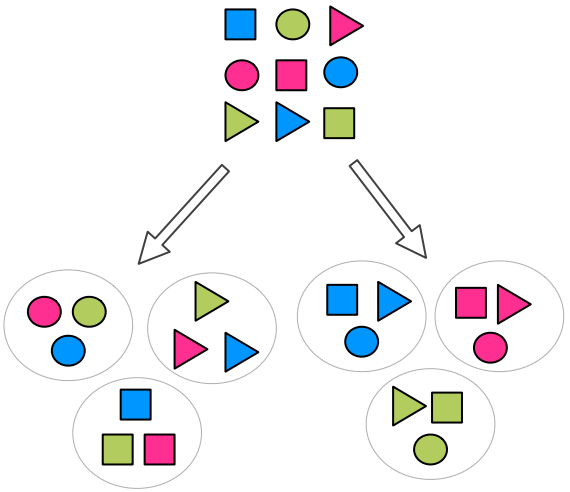

Figure 6.1: Regroupement des données en trois groupes selon la forme (à gauche) et selon la couleur (à droite).

▶ **Exemple 6.1** Prenons l'exemple de la figure 6.1. Nous disposons d'un ensemble d'objets caractérisés par leur forme géométrique et leur couleur. Dans ce cas de figure, nous avons $N = 9$ objets et $K = 2$ variables correspondant à la forme et la couleur. Les techniques de regroupement permettent d'obtenir trois groupes en se basant, soit, sur la forme uniquement (groupes illustrés à gauche), soit, en se basant sur la couleur (groupes illustrés à droite).

Le regroupement trouve son application dans plusieurs domaines, notamment :

- **Marketing** pour segmenter le marché en groupes de clients distincts à partir de bases de données d'achats.
- **Environnement** pour déterminer des zones terrestres similaires dans une base de données d'observation de la terre.
- **Assurance** pour déterminer des groupes d'assurés distincts qui présentent un nombre important de réclamations.
- **Biologie** pour identifier les espèces proches et présenter des arbres généalogiques.

6.2 Évaluation de la qualité du regroupement

Une bonne méthode de regroupement produit des groupes d'individus homogènes entre eux et distincts des autres groupes. Ceci se traduit par :

- Une similarité intra-groupe importante, pour obtenir les groupes les plus homogènes possible.
- Une similarité inter-groupe faible afin d'obtenir des groupes bien distincts.

La qualité des regroupements peut s'évaluer grâce à plusieurs indicateurs comme l'inertie intra-groupe, l'inertie inter-groupe et l'indice de silhouette.

6.2.1 Inertie intra-groupe

Soit $\mathcal{D} = \{\mathbf{x}_i\}_{i=1...N}$ un ensemble de données contenant N individus. Chaque individu $\mathbf{x}_i = (x_i^1, ..., x_i^D)'$ est décrit par D caractéristiques. Autrement dit, un individu est un point de l'espace $\mathbb{R}^D$ caractérisé par un vecteur $\mathbf{x}_i = (x_i^1, ..., x_i^D)'$.

On suppose que l'ensemble des données appartenant à $\mathcal{D}$ est réparti dans K groupes G_k, $k \in \{1, 2, ...K\}$.

Pour chaque groupe G_k, l'inertie est la variance des individus formant le groupe. Elle est déterminée selon :

$$J_k = \sum_{i \in G_k} d^2(\mathbf{x}_i, \mu_k) \tag{6.1}$$

où μ_k désigne le centre de gravité du groupe G_k.

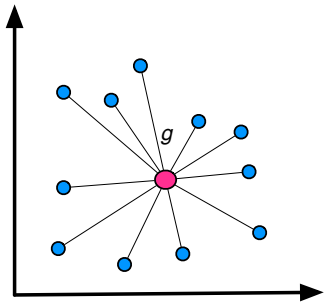

Figure 6.2: Inertie intragroupe.

L'inertie intragroupe est donc la somme des inerties de tous les groupes :

$$J_w = \sum_{k=1}^{K} \sum_{i \in G_k} d^2(\mathbf{x}_i, \mu_k) = \sum_{i \in G_k} J_k \tag{6.2}$$

(L'indice w fait référence à *within*).
L'inertie intra-groupe permet, donc, de mesurer la concentration des individus du groupe autour du centre de gravité. On peut ainsi conclure que plus la valeur de l'inertie intra-groupe est faible, plus les individus sont concentrés autour du centre de gravité du groupe en question.

La figure 6.2 représente un ensemble de données dans l'espace $\mathbb{R}^2$. Le point g au centre représente le centre de gravité.

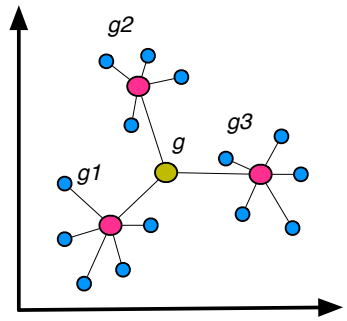

Figure 6.3: Inertie inter-groupes

6.2.2 Inertie inter-groupe

Soit μ la moyenne de toutes les données indépendamment de leur groupe d'appartenance :

$$\mu = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_i \tag{6.3}$$

où N désigne le nombre d'individus.
L'inertie inter-groupe :

$$J_b = \sum_{c} N_c d^2(\mu_c, \mu) \tag{6.4}$$

(L'indice b fait référence à *between*).
L'inertie inter-groupe mesure l'éloignement des centres des groupes entre eux. Ainsi, plus l'inertie inter-groupe est grande, plus les groupes sont bien distincts.

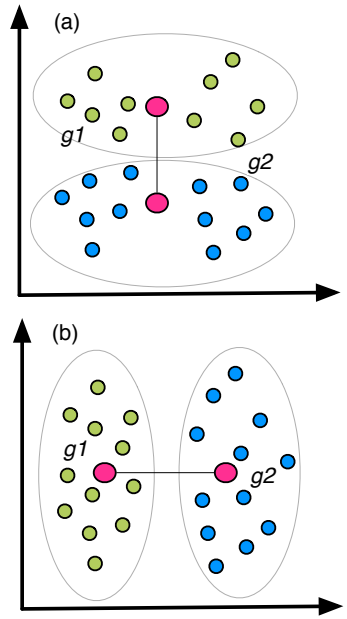

Figure 6.4: Deux partitions différentes d'un même ensemble de données.

En conclusion, une bonne partition des données réduit le plus possible l'inertie intra-groupe et favorise au maximum l'inertie inter-groupe.

La figure 6.4 illustre sur un même ensemble de données deux structures en deux groupes possibles. Nous obtenons en (b) une bonne partition et en (a) une moins bonne structure puisque l'inertie inter-groupe est plus petite.

6.2.3 Indice de silhouette

L'indice de silhouette est un indicateur efficace pour valider une méthode de regroupement [4]. Pour tout individu i de l'ensemble des données, l'indice de silhouette est défini par la formule suivante :

$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{6.5}$$

où a_i est la dissimilarité moyenne entre l'individu i et tous les autres individus du groupe G_j auquel il appartient (la dissimilarité est expliquée à la section 6.4.3), et b_i est le minimum des dissimilarités moyennes entre l'individu i et tous les autres individus des groupes G_k $(k = 1, 2, ..., K$ avec $k \neq j)$ [2].

La représentation graphique de l'indice de silhouette (figure 6.5) montre une bonne séparation d'un ensemble de données en quatre groupes. Sur l'axe des ordonnées sont représentées les données des différents groupes selon le regroupement utilisé. L'axe des abscisses représente la valeur de l'indice de silhouette calculé selon l'équation 6.5. L'indice de silhouette est positif pour la majorité des échantillons, sauf pour 6 individus (1 appartenant à G1 et 5 à G2). Ce qui démontre l'homogénéité des données au sein des différents groupes. Pour plus de détails, vous pouvez vous référer à l'exemple de la Section 6.6.2.

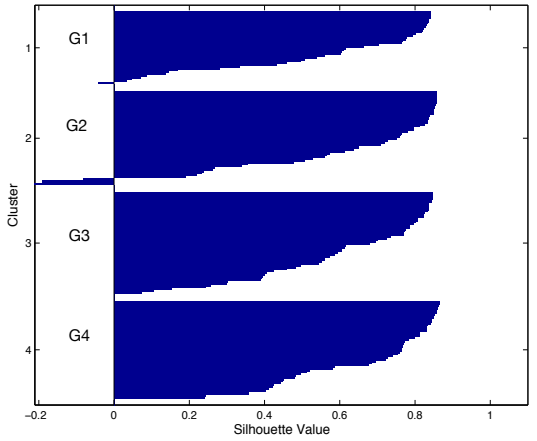

Figure 6.5: Indice de silhouette d'un ensemble de données en 4 groupes.

6.3 Détermination du nombre optimal de regroupement

Les critères présentés à la section 6.2 permettent de répondre à la question d'évaluation de la qualité d'un regroupement. Une question complémentaire à cette dernière est : comment déterminer le nombre optimal de regroupement ?

6.3.1 La méthode du coude

La méthode du coude (*Elbow method*) est la plus connue dans la littérature. Son principe consiste à mesurer et représenter la variation de l'inertie intra-groupe en fonction du nombre k de groupe. Le « coude » est le point correspondant au nombre de regroupement à partir duquel l'inertie intra-groupe ne diminue plus significativement ce qui correspond, alors, au nombre optimal de groupes. Le diagramme de la Figure 6.6 représente la variation de l'inertie intra-groupes en fonction du nombre de regroupement. Dans cet exemple, le coude est situé à 4 groupes donc le nombre optimal de regroupement est fixé à 4.

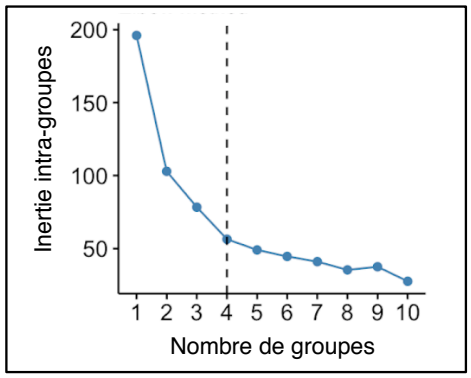

Figure 6.6: Méthode du coude.

6.3.2 La méthode de silhouette

L'indice de silhouette peut aussi permettre de déterminer le nombre optimal de groupe. Ce dernier correspond au nombre k de groupe qui maximise la silhouette moyenne. Dans l'exemplede la Figure , le coude est situé à 4 groupes donc le nombre optimal de regroupement est fixé à 4.

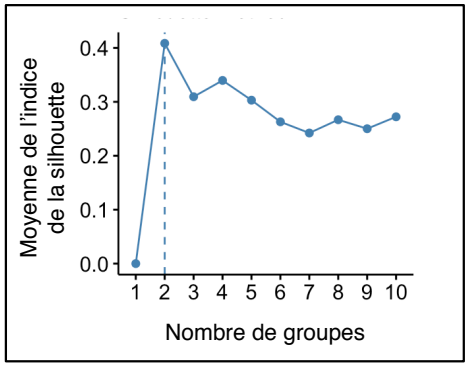

Figure 6.7: Méthode de silhouette.

6.4 Distance, similarité et dissimilarité

Comme on l'a mentionné précédemment, une bonne méthode de groupement maximise la ressemblance et la similarité entre les données à l'intérieur de chaque groupe, et minimise

celles entre les données des groupes différents, d'où l'importance de présenter les notions de distance, de similarité et de dissimilarité entre les données.[1].

6.4.1 Distance

La distance d définie sur un ensemble E est une application de $E \times E$ dans $\mathbb{R}^+$ vérifiant les propriétés suivantes (décrites aussi dans [3]) :

- Symétrie
$$\forall (x, y) \in E \times E,\ d(x, y) = d(y, x)$$
- Séparation
$$\forall (x, y) \in E \times E,\ d(x, y) = 0 \Leftrightarrow x = y$$
- Inégalité triangulaire
$$\forall (x, y, z) \in E^3,\ d(x, z) \leq d(x, y) + d(y, z)$$

6.4.2 Similarité

La similarité s est une mesure de proximité de $E \times E$ dans $\mathbb{R}^+$ qui vérifie les propriétés suivantes :

- Symétrie
$$\forall (x, y) \in E \times E,\ s(x, y) = s(y, x)$$
- Maximale
$$\forall (x, y) \in E \times E,\ s(x, x) \geq s(x, y)$$
- Majoration
$$\forall (x, y) \in E \times E,\ s(x, y) \leq S$$
où S est la valeur de majoration.

6.4.3 Dissimilarité

La dissimilarité d est une application de $E \times E$ dans $\mathbb{R}^+$ qui vérifie les propriétés suivantes :

- Symétrie
$$\forall (x, y) \in E \times E,\ s(x, y) = s(y, x)$$
- Séparation
$$\forall (x, y) \in E \times E,\ d(x, y) = 0 \Leftrightarrow x = y$$
- Majoration
$$\forall (x, y) \in E \times E,\ s(x, y) \leq S$$
où S est la valeur de majoration.

La dissimilarité et la similarité évoluent dans des sens opposés. Si s désigne la similarité et d est la dissimilarité, alors

$$\forall (x, y) \in E \times E, d(x, y) = S - s(x, y)$$

On remarque donc que plus la similarité est forte, plus la dissimilarité est faible ; et inversement.

6.4.4 Mesure de distance entre variables

La mesure de la distance entre variables dépend de la nature de ces dernières. Dans ce qui suit, nous en citons les mesures les plus fréquemment utilisées.

Variables quantitatives

Soient $\mathbf{x}_i = (x_i^1, ..., x_i^d)'$ et $\mathbf{x}_j = (x_j^1, ..., x_j^d)'$ deux variables de dimension P. La distance $d(\mathbf{x}_i, \mathbf{x}_j)$ entre $\mathbf{x}_i$ et $\mathbf{x}_j$ peut être calculée selon les mesures suivantes :

- La distance de Minkowski :

$$d^q(\mathbf{x}_i, \mathbf{x}_j) = \sum_{p=1}^{P} (x_i^p - x_j^p)^q$$

  où q est entier positif.

- La distance euclidienne est un cas particulier de la distance Minkowski avec $q = 2$:

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{p=1}^{P} (x_i^p - x_j^p)^2$$

- La distance de Manhattan est obtenue à partir de la distance de Minkowski en posant $q = 1$:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{p=1}^{P} |x_i^p - x_j^p|$$

- La distance de Mahalanobis prend en considération la matrice de covariance Σ :

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)^t$$

  Σ est la matrice de covariance.

Variables binaires

Soient deux variables binaires de dimension D. Soient deux individus A et B décrits, respectivement, par les variables binaires $\mathbf{x}_i = (x_i^1, ..., x_i^D)'$ et $\mathbf{x}_j = (x_j^1, ..., x_j^D)'$ de dimension D. La similarité entre $\mathbf{x}_i$ et $\mathbf{x}_j$ peut être calculée à partir de la table de contingence qui comprend, à l'intersection des lignes et des colonnes, les effectifs des deux variables.

		Individu B		
		1	0	somme
Individu A	1	a	b	$a+b$
	0	c	d	$c+d$
	somme	$a+c$	$b+d$	p

La somme $(a+d)$ indique la concordance entre les variables et la somme $(b+c)$ indique la discordance.

Dans le cas de variables binaires symétriques, le coefficient de correspondance simple est souvent utilisé pour mesurer la similarité.

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{b+c}{a+b+c+d}$$

Dans le cas de variables binaires asymétriques, le coefficient de Jaccard est fréquemment utilisé pour mesurer la similarité entre individus [3].

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{a}{a+b+c} \tag{6.6}$$

▶ **Exemple 6.2** Soit le tableau suivant qui décrit l'état de trois patients décrit par 6 variables. La première variable décrit le sexe. Elle prend la valeur `M` ou `F`. Il s'agit d'une variable binaire symétrique. Ensuite, on trouve les variables Fièvre et Toux qui prennent les valeurs `Oui` ou `Non`. Il s'agit de variables binaires asymétriques. De même pour les variables Test-1 à Test-4 qui prennent les valeurs `Positif` ou `Négatif`.

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Myriam	`F`	`Oui`	`Non`	`Positif`	`Négatif`	`Négatif`	`Négatif`
Charlotte	`F`	`Oui`	`Non`	`Positif`	`Négatif`	`Positif`	`Négatif`
Alexandre	`M`	`Oui`	`Oui`	`Négatif`	`Négatif`	`Négatif`	`Négatif`

Le tableau de valeurs est transformé en un tableau de valeurs binaires qui prend les valeurs `0` (`Non` ou `Négatif`) et `1` (`Oui` ou `Positif`) .

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Myriam	F	1	0	1	0	0	0
Charlotte	F	1	0	1	0	1	0
Alexandre	M	1	1	0	0	0	0

On désire calculer la distance entre les individus Myriam et Charlotte. On détermine le tableau de contingence (Myriam, Charlotte) qui comprend, à l'intersection des lignes et des colonnes, les effectifs des deux variables, c'est-à-dire le nombre de fois (le nombre de correspondances entre les valeurs de `0` et de `1`) :

		Charlotte	
		1	0
Myriam	1	2	0
	0	1	3

En utilisant l'équation (6.6), la distance est alors égale à :

$$\mathrm{d}(\text{Myriam, Charlotte}) = \frac{2}{2+0+1} = 0.66$$

Variables qualitatives nominales

La mesure de la distance entre variables nominales nécessite leur transformation au préalable.
Soit un individu A décrit par la variable $\mathbf{x}_i = (x_i^1, ..., x_i^D)'$ de dimension K. Pour chaque composante j, $j = 1, 2, ...D$, la valeur de la composante x_i^j est remplacée par son rang r_i^j, avec $r_i^j = 1, 2, ...D$. La variable x_i^j est ensuite transformée en une variable z_i^j selon :

$$z_i^j = \frac{r_i^j - 1}{D - 1}$$

La valeur de z_i^j est comprise entre 0 et 1. On peut, par conséquent, calculer les distances entre individus en les considérant comme des variables numériques.

Variables qualitatives ordinales

Dans le cas de variables qualitatives ordinales, les variables sont remplacées par leur rang, puis la distance est calculée entre les valeurs des rangs.

6.4.5 Mesures de proximité entre groupes

La **proximité** entre deux groupes G_1 et G_2 est évaluée par la distance $d(G_1, G_2)$ qui peut être mesurée selon :

- Le diamètre minimum défini par :

$$D_{\min}(G_1, G_2) = \min\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2\}$$

- Le diamètre maximum défini par :

$$D_{\max}(G_1, G_2) = \max\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2\}$$

- La distance moyenne doit être pondérée afin de prendre en considération la taille des groupes, c'est-à-dire le nombre d'individus de chacun d'eux :

$$D_{moy} = \frac{1}{n_1 n_2} \sum_{\mathbf{x}_i \in G_1} \sum_{\mathbf{x}_j \in G_2} d(\mathbf{x}_i, \mathbf{x}_j)$$

  où n_1 et n_2 correspondent aux nombres d'individus dans G_1 et G_2.

- La distance de Ward utilise une analyse de la variance afin d'évaluer les distances entre groupes :

$$D_{Ward} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} d(\mu_1, \mu_2)$$

  où μ_1 et μ_2 sont les moyennes dans G_1 et G_2 respectivement.

6.5 Regroupement hiérarchique

Le regroupement hiérarchique consiste à déterminer des groupes d'individus qui se forment par agglomération progressive. On en distingue deux types : le regroupement hiérarchique ascendant et le regroupement hiérarchique descendant.

Les méthodes de regroupement hiérarchique produisent souvent une représentation graphique des résultats sous la forme d'un **dendrogramme**, qui est un arbre dont les feuilles sont les individus alignés sur l'axe des abscisses. L'axe vertical indique les distances entre les individus.
Le nombre de groupe obtenu suite à une regroupement hiérarchique en coupant l'arbre à un certain niveau (L1 ou L2 selon la figure 6.8). Ce dernier peut être optimisé en utilisant la variabilité inter et intra-groupes.

▶ **Exemple 6.3** La figure 6.8 est un dendogramme qui représente le regroupement de 5 individus. Si la distance de coupure est fixée à `L1=3`, alors nous obtenons 3 groupes {`1, 2`}, {`3`} et {`4, 5`}. Alors que si la distance est fixée à `L2=6`, alors nous obtenons deux groupes {`1, 2, 3`} et {`4, 5`}.

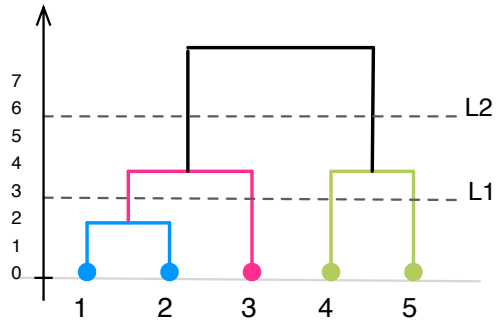

Figure 6.8: Un dendrogramme.

6.5.1 Regroupement hiérarchique ascendant

Le principe du regroupement hiérarchique **ascendant** ou **agglomératif** considère, au départ, que chaque individu est un regroupement en soit. Ensuite, à chaque étape, des groupes sont constitués par agrégation des deux individus les plus proches de la partition de l'étape précédente. Le processus de regroupement s'arrête lorsqu'on arrive à former un seul groupe (il s'agit donc d'une démarche *bottom-up*).
Le dendrogramme issu du regroupement agglomératif est construit de manière ascendante : les N individus sont regroupés, progressivement, en partant de n groupes jusqu'à l'obtention de 1 groupe.
Le regroupement hiérarchique ascendant peut être décrit par les lignes de pseudo-codes suivantes dans lesquelles on considère que chaque individu est un groupe indépendant et qu'on cherche à déterminer K regroupements à partir d'un ensemble de n individus.

Regroupement hiérarchique ascendant :

1. Placer chaque individu dans son propre groupe.
2. Calculer une liste (un tableau) des distances inter-groupes et la trier dans l'ordre croissant.
3. **Répéter**
4. Regrouper les deux groupes les plus proches au sens de la distance entre groupes choisis.
5. Mettre à jour le tableau de distances en remplaçant les deux groupes formés.
6. **Jusqu'à** l'obtention d'un seul groupe.

6.5.2 Regroupement hiérarchique descendant

Contrairement au regroupement hiérarchique ascendant, le regroupement **descendant** ou **divisif** commence à partir d'un ensemble de données et le subdivise en sous-ensembles pour générer une séquence de regroupements ordonnée. Il s'agit donc d'une approche *top-down*.

Le dendrogramme issu du regroupement divisif est construit de manière descendante : les n individus sont divisés, progressivement, à partir d'un seul groupe jusqu'à N groupes.

6.5.3 Choix du critère de ressemblance

Les méthodes hiérarchiques (ascendante et descendante) diffèrent entre elles par le choix du critère de ressemblance servant à lier les groupes. On distingue les critères de lien simple, complet et moyen.

Regroupement à lien simple

Le principe du regroupement à **lien simple** aussi appelé **saut minimum** (*simple linkage*) consiste à classer les couples de données par ordre de similarité croissante et à regrouper ceux dont la similarité est la plus forte (Figure 6.9 - a).

Soient deux individus A et B qui proviennent respectivement de 2 groupes $G1$ et $G2$ et qui sont décrits par les variables x et y. Un lien simple est établi entre A et B si :

$$d_S(G1, G2) = \min_{x \in G1, y \in G2} \{d(x, y)\}$$

Regroupement à lien complet

Le principe du regroupement à **lien complet**, aussi appelé **saut maximum** (*complete linkage*), consiste à classer les couples de données par ordre de similarité décroissante et

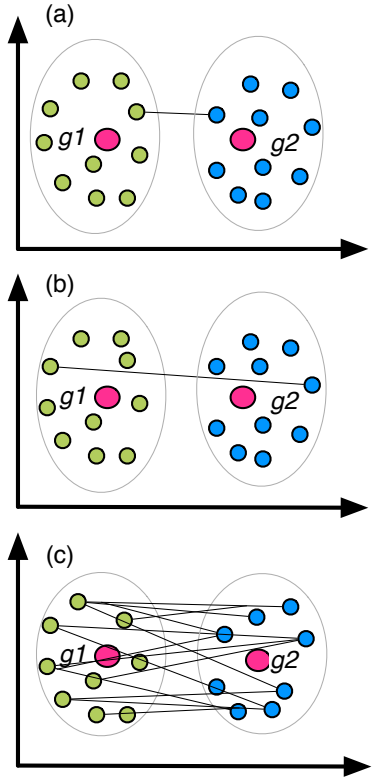

Figure 6.9: Regroupement à (a) lien simple, (b) lien complet et (c) lien moyen.

à regrouper ceux dont la similarité est la plus faible. Un lien complet est établi entre A et B si :

$$d_C(G_1, G_2) = \max_{x \in G1, y \in G2} \{d(x, y)\}$$

Regroupement à lien moyen

Le principe du regroupement à **lien moyen** (*average linkage*) consiste à fusionner les deux groupes les plus proches pour ce qui est de leur distance moyenne. Cette distance est la moyenne calculée entre toutes les paires d'individus des deux groupes :

$$d_M(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{x \in G_1} \sum_{y \in G_2} d(x, y)$$

avec n_1 et n_2, qui sont, respectivement, le nombre d'individus dans G_1 et G_2.

▶ **Exemple 6.4** L'objectif de cet exemple est d'appliquer le regroupement hiérarchique descendant. Soit l'ensemble de données décrit dans le tableau suivant :

Individus	A	B	C	D	E	F
x	27	51	52	33	45	22
y	28	44	40	38	20	57

Il s'agit d'un ensemble de 6 individus caractérisés par deux variables quantitatives (figure 6.10). Le regroupement de ces données en utilisant les liens simple, complet et moyen est illustré dans la figure 6.11.

Les individus sont codés selon le codage du tableau suivant :

Individus	A	B	C	D	E	F
Indices	1	2	3	4	5	6

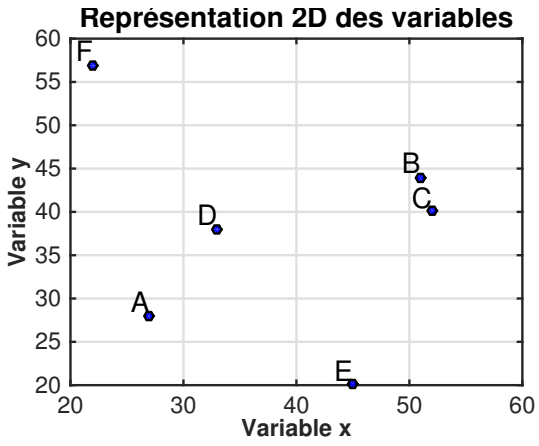

Figure 6.10: Représentation des données dans l'espace 2D.

Selon la figure 6.11, sur l'axe des abscisses, on trouve les indices des individus tels qu'ils sont décrits dans le tableau précédent. Par exemple, par un lien simple, on relie d'abord les individus B (d'indice 2) et C (d'indice 3) en raison de leur proximité.

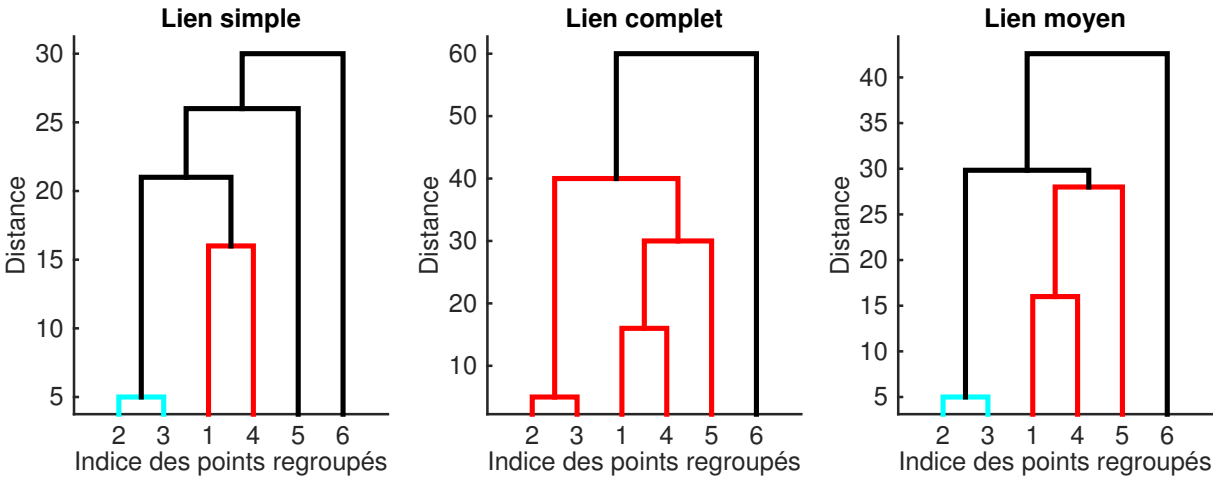

Figure 6.11: Regroupement des données par les liens simple, complet et moyen.

6.6 Regroupement par partition

6.6.1 Principe du regroupement par partition

Les méthodes de regroupement par partition consistent à construire une partition unique en K groupes à partir des N individus à regrouper. Il existe de multiples méthodes de partitionnement des données, telles que la méthode $K-$moyennes, la méthode des $C-$medoids et les nuées dynamiques. Dans la suite de ce module, nous traitons uniquement de la méthode K-moyennes.

6.6.2 Méthode $K-$moyennes

L'algorithme $K-$moyennes ($K-means$) utilise une technique d'affinement itératif qui consiste à améliorer progressivement la qualité des groupes selon un indice de similarité entre les différents individus. Le paramètre d'entrée de l'algorithme $K-$moyennes est le nombre de groupes désiré K. La distance euclidienne est souvent utilisée pour mesurer le rapprochement des groupes.

Algorithme $K-$moyennes : On cherche à déterminer K regroupements à partir d'un ensemble de N individus.

1. Choisir aléatirement K centres initiaux G_i avec $i = 1, 2, ...K$.
4. **Répéter**
2. Répartir chaque individu dans le groupe G_i le plus proche.
3. Recalculer les nouveaux centres G_i.
4. **Jusqu'à ce que** les centres deviennent stables.

▶ **Exemple 6.5** Soit $E = \{1, 2, 3, 6, 7, 8, 13, 15, 17\}$ un ensemble de 9 individus décrits par leur position, comme illustré dans la figure 6.12.

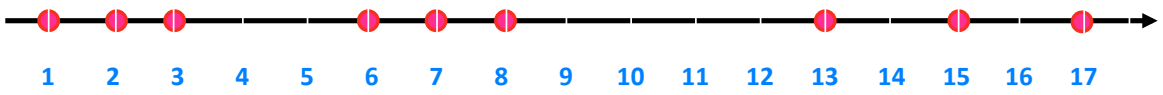

Figure 6.12: Représentation des données de l'ensemble E dans l'espace 1D.

Ces individus sont codés selon le codage du tableau suivant :

Tableau 6.1: Ensemble des individus et leur indice correspondant

x	1	2	3	6	7	8	13	15	17
Indices	1	2	3	4	5	6	7	8	9

On désire créer trois groupes à partir de l'ensemble E. L'algorithme $K-$moyennes permet d'obtenir le résultat illustré dans la figure 6.13.

Exécution de l'algorithme $K-$moyennes

Soit $A = \{1, 2, 3, 6, 7, 8, 13, 15, 17\}$. Nous considérons 3 individus au hasard (Ligne 1 de l'algorithme $K-$moyennes). Supposons que ça soit 1, 2 et 3. Nous avons alors $\mu_1 = 1$, $\mu_2 = 2$ et $\mu_3 = 3$

À l'itération 1 :

x	1	2	3	6	7	8	13	15	17
$d^2(x, \mu_1)$	0	1	4	25	36	49	144	196	256
$d^2(x, \mu_2)$	1	0	1	16	25	36	121	169	225
$d^2(x, \mu_3)$	4	1	0	9	16	25	100	144	196

Afin de simplifier les calculs, nous présentons dans ce tableau les carrés des distances.
Pour $x = \{1\}$, la distance minimale est par rapport à μ_1.
Pour $x = \{2\}$, la distance minimale est par rapport à μ_2.
Pour $x = \{3, 6, 7, 8, 13, 15, 17\}$, la distance minimale est par rapport à μ_3.

D'où $G1 = \{1\}$, $G2 = \{2\}$ et $G3 = \{3, 6, 7, 8, 13, 15, 17\}$

Nous calculons par la suite les nouvelles valeurs des moyennes :

$$\mu_1 = 1$$

$$\mu_2 = 2$$

$$\mu_3 = \frac{3+6+7+8+13+15+17}{7} = 9.8$$

À l'itération 2 :

x	1	2	3	6	7	8	13	15	17
$d^2(x, \mu_1)$	0	1	4	25	36	49	144	196	256
$d^2(x, \mu_2)$	1	0	1	16	25	36	121	169	225
$d^2(x, \mu_3)$	77,44	60,84	46,24	14,44	7,84	3,24	10,24	27,04	51,84

Pour $x = \{1\}$, la distance minimale est par rapport à μ_1.
Pour $x = \{2, 3\}$, la distance minimale est par rapport à μ_2.
Pour $x = \{6, 7, 8, 13, 15, 17\}$, la distance minimale est par rapport à μ_3.

D'où $G1 = \{1\}$, $G2 = \{2, 3\}$ et $G3 = \{6, 7, 8, 13, 15, 17\}$

Et ainsi de suite jusqu'à la convergence, c'est-à-dire jusqu'à ce que les valeurs des moyennes ne changent plus.
À la convergence, nous obtenons :
$G1 = \{1, 2, 3\}$, $G2 = \{6, 7, 8\}$ et $G3 = \{13, 15, 17\}$ avec les moyennes : $\mu_1 = 2$, $\mu_2 = 7$ et $\mu_3 = 15$

Analyse visuelle

Ce résultat est prévisible par une simple analyse visuelle des données qui permet de faire ressortir les trois groupes.

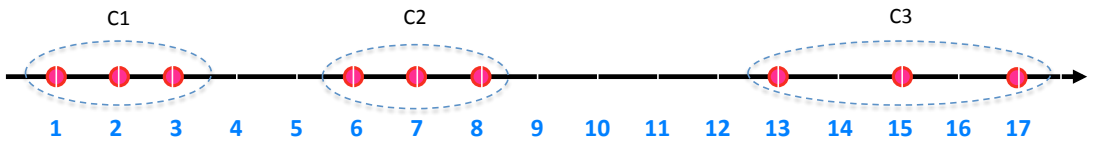

Figure 6.13: Représentation des données dans l'espace 1D.

Analyse avec R

L'exécution de l'algorithme $K-$moyennes en utilisant la fonction `kmeans` du logiciel R fournit le groupe d'appartenance de chaque individu $E = \{3, 3, 3, 1, 1, 1, 2, 2, 2\}$. La figure 6.14 illustre la silhouette de l'ensemble E ; elle montre une bonne séparation des trois groupes qui contiennent chacun trois individus. Les chiffres sur l'axe des ordonnées, à gauche, indiquent l'indice des individus par groupe (tels que décrit à la figure 6.14). L'indice des groupes est indiqué sur l'axe des ordonnées à droite.

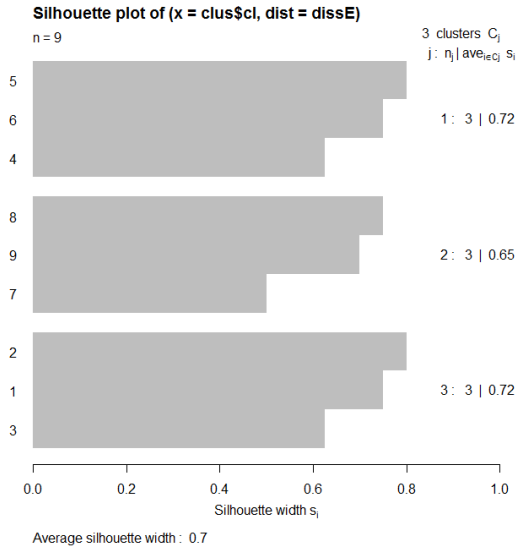

Figure 6.14: La silhouette de l'ensemble des données E.

Avantages et faiblesses de l'algorithme $K-$moyennes

L'algorithme $K-$moyennes est très populaire dans les applications scientifiques et industrielles en raison de la facilité de sa mise en oeuvre. Cependant, cette facilité donne lieu à des faiblesses. En effet, l'algorithme $K-$moyennes prend en compte les données aberrantes (points extrêmes en dehors des groupes), ce qui entraîne un biais lors du calcul des moyennes et par conséquent pour la détermination des centres des groupes. De plus, la complexité de l'algorithme est de l'ordre de $\Theta(ICN)$, où I est le nombre d'itérations, K le nombre de groupes et N le nombre d'individus. Finalement cette algorithme impose un choix de K fixe.

6.7 Annexes

Dans cette section, nous citons des fonctions utiles pour les regroupement en utilisant le logiciel R. Nous reprenons, en général, textuellement la description fournie dans le site web de R (https://cran.r-project.org).

- `cluster` : Ce paquet contient les méthodes d'analyse de regroupements. Il s'agit d'une mise en oeuvre de l'ouvrage de Peter Rousseeuw, Anja Struyf et Mia Hubert, sur la base de Kaufman et Rousseeuw (1990)"Finding Groups in Data".

- `hclust` : Regroupement hiérarchique.
  https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html

- `kmeans` : Regroupement $K-$moyennes.
  https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Références

[1] Classification non supervisée. http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-classif.pdf. Consulté : 3 décembre 2018.

[2] N. Mezghani, A. Fuentes, N. Gaudreault, A. Mitiche, N. Hagemeister, R. Aissaoui, and J.A De Guise. Identification of knee frontal plane kinematic patterns in normal gait by principal component analysis. *Journal of Mechanics in Medicine and Biology*, 13(3) :17 – 24, 2013.

[3] J.P. Nakache and J. Confais. *Approche pragmatique de la classification : arbres hiérarchiques, partitionnements*. Ed. Technip, Paris, 2004.

[4] P. Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1) :53–65, 1987.

