Classification

Exemple: On considère le tableau X de données suivant:

	X_1	X_2
I_1	2	2
I_2	7.5	4
I_3	3	3
I_4	0.5	5
I_5	6	4

On cherche à faire une classification hiérarchique ascendante en utilisant la distance euclidienne et la méthode d'agrégation de Ward.

On note $N_I = \{I_1, I_2, I_3, I_4, I_5\}$ le nuage des individus à classer.

- **Étape 1:** $P_5 = \{I_1, I_2, I_3, I_4, I_5\}$

Classification

- Matrice des distances (euclidienne) 5×5 entre les individus:

	I_1	I_2	I_3	I_4	I_5
I_1	0	5.85	1.41	3.35	4.47
I_2	5.85	0	4.61	7.07	1.50
I_3	1.41	4.61	0	3.20	3.16
I_4	3.35	7.07	3.20	0	5.59
I_5	4.47	1.5	3.16	5.59	0

- **Étape 2:** Matrice des distances de Ward 5×5:

	I_1	I_2	I_3	I_4	I_5
I_1	0	17.12	1	5.62	10
I_2	17.12	0	10.62	25	1.12
I_3	1	10.62	0	5.12	5
I_4	5.62	25	5.12	0	15.62
I_5	10	1.12	5	15.62	0

Classification

Par exemple:

$$d_w(I_1, I_2) = \frac{1 \times 1}{1+1} \times 5.85^2 = 17.12$$

La plus petite valeur ($\neq 0$) dans le tableau des distances de Ward est 1 entre l'individu I_1 et I_3, donc on agrège ces deux individus dans le groupe $G_1 = \{I_1, I_3\}$ et on obtient une nouvelle partition $P_4 = \{I_2, I_4, I_5, G_1\}$

Le centre de gravité associé à G_1 est le point g_1 de coordonnées:
$g_1 = (\frac{2+3}{2}, \frac{2+3}{2}) = (2.5, 2.5)$

L'inertie intraclasse de P_4 est:

$$\begin{aligned} I_{intra}(P_4) &= \frac{1}{n} \sum_{k=1}^{1} \sum_{i \in G_1} d^2(I_i, g_1) \\ &= \frac{1}{n}(d^2(I_1, g_1) + d^2(I_3, g_1)) \\ &= \frac{1}{5}((2-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (3-2.5)^2) = 0.2 \end{aligned}$$

- **Étape 3:** Nouvelle matrice des distances 4×4 de Ward:

	I_2	I_4	I_5	G_1
I_2	0	25	1.12	18.16
I_4	25	0	15.62	6.83
I_5	1.12	15.62	0	9.66
G_1	18.16	6.83	9.66	0

Avec $d_w(I_2, G_1) = \frac{1\times 2}{1+2}((7.5-2.5)^2 + (4-2.5)^2) = 18.16$

Classification

La plus petite valeur dans le tableau des distances 4×4 de Ward est 1.12 entre l'individu I_2 et I_5, donc on agrège ces deux individus dans le groupe $G_2 = \{I_2, I_5\}$ et on obtient une nouvelle partition $P_3 = \{I_4, G_1, G_2\}$

Le centre de gravité associé à G_2 est le point g_2 de coordonnées:
$g_2 = (\frac{7.5+6}{2}, \frac{4+4}{2}) = (6.75, 4)$

L'inertie intraclasse de P_3 est:

$$
\begin{aligned}
I_{intra}(P_3) &= \frac{1}{n} \sum_{k=1}^{2} \sum_{i \in G_k} d^2(I_i, g_k) \\
&= \frac{1}{n} \left(\sum_{i \in G_1} d^2(I_i, g_1) + \sum_{i \in G_2} d^2(I_i, g_2) \right) \\
&= 0.2 + \frac{1}{5}((7.5 - 6.75)^2 + (4 - 4)^2 + (6 - 6.75)^2 + (4 - 4)^2) \\
&= 0.425
\end{aligned}
$$

- **Étape 4:** Nouvelle matrice des distances 3×3 de Ward:

	I_4	G_1	G_2
I_4	0	6.83	26.7
G_1	6.83	0	20.31
G_2	26.7	20.31	0

Avec $d_w(G_1, G_2) = \frac{2\times 2}{2+2}((6.75 - 2.5)^2 + (4 - 2.5)^2) = 20.31$

Classification

La plus petite valeur dans le tableau des distances 3×3 de Ward est 6.83 entre l'individu I_4 et G_1, donc les individus I_4 et G_1 sont les plus proches. On les regroupe pour former le groupe $G_3 = \{I_4, G_1\}$ et on obtient une nouvelle partition $P_2 = \{G_3, G_2\}$

Le centre de gravité associé à G_3 est le point g_3 de coordonnées:
$g_3 = (\frac{2+3+0.5}{3}, \frac{2+3+5}{3}) = (1.833, 3.333)$

L'inertie intraclasse de P_2 est:

$$\begin{aligned} I_{intra}(P_2) &= \frac{1}{n} \sum_{i \in G_2} d^2(I_i, g_2) + \frac{1}{n} \sum_{i \in G_3} d^2(I_i, g_3) \\ &= 1.79 \end{aligned}$$

- **Étape 5:** Nouvelle matrice des distances 2×2 de Ward:

	G_2	G_3
G_2	0	29.54
G_3	29.54	0

Avec $d_w(G_2, G_3) = \frac{2\times3}{2+3}((6.75 - 1.833)^2 + (4 - 3.333)^2) = 29.54$

Classification

Il ne reste plus que 2 éléments G_2 et G_3, on les regroupe

$\Rightarrow G_4 = \{G_2, G_3\}$. Cela donne la partition $P_1 = \{G_4\}$.

L'inertie intraclasse de P_1 est égale à l'inertie totale du nuage:

$$\begin{aligned} I_{intra}(P_1) &= I_{totale}(N_I) \\ &= \frac{1}{n}\sum_{i=1}^{n} d^2(I_i, g) \\ &= 7.7 \end{aligned}$$

Avec g est le centre de gravité du nuage N_I.

Méthodes de classification:

2) Méthode des centres mobiles

La méthode des centres mobiles ou la méthode K-means est fondée sur une méthode de partitionnement directe des individus connaissant par avance le nombre de classes attendues.

Soit $X = (x_{ij})_{i=1,\ldots,n\ ;j=1,\ldots,p}$ une matrice d'observations. On choisit a priori le nombre de classes K (avec $K \leq n$). On note g_k le centre de gravité de la classe k.

Algorithme des kmeans

Étape 0: Choisir le nombre de classes K puis choisir K points (individus) au hasard parmi les n individus.

$\Rightarrow$ Ces K individus servent de centres initiaux des classes.

Étape 1: Allouer l'individu I_i à la classe k telle que $d(I_i, g_k) \leq d(I_i, g_l)$ pour tout $l \neq k$.

Étape 2: Recalculer les centres de gravité g_k des K classes.

Étape 3: Répéter les étapes 1 et 2 jusqu'à la stabilité des centres (les centres ne bougent plus)

Remarque: (autre critère d'arrêt)

L'algorithme est itéré jusqu'à ce que le critère de variance interclasse ne croisse plus de manière significative.

Illustration: (Voir le fichier "exemple.ppt")

Exemple:

On reprend l'exemple précédent,

	X_1	X_2
I_1	2	2
I_2	7.5	4
I_3	3	3
I_4	0.5	5
I_5	6	4

Soit $N_I = \{I_1, I_2, I_3, I_4, I_5\}$ le nuage des individus à classer.

On cherche à regrouper les individus en $K = 2$ classes.

Classification

Étape 0: Soit $K = 2$, on considère, par exemple, les deux individus I_1 et I_5 comme des centres initiaux, c-à-d $g_1^0 = I_1 = (2, 2)$ et $g_2^0 = I_5 = (6, 4)$

Étape 1: Tableau des distances entre les individus et les centres,

	I_1	I_2	I_3	I_4	I_5
g_1^0	0	5.85	1.41	3.35	4.47
g_2^0	4.47	1.5	3.16	5.59	0

Donc, on obtient les deux groupes suivant:

$G_1 = \{I_1, I_3, I_4\}$ et $G_2 = \{I_2, I_5\}$

Étape 2: Recalculer les centres de gravité:

On considère deux nouveaux centres, g_1^1 et g_2^1, lesquels sont les centres de gravité des deux groupes G_1 et G_2.

Donc

$g_1^1 = (\frac{2+3+0.5}{3}, \frac{2+3+5}{3}) = (1.83, 3.33)$ et

$g_2^1 = (\frac{7.5+6}{2}, \frac{4+4}{2}) = (6.75, 4)$

Étape 3: Tableau des distances entre les individus et les nouveaux centres,

	I_1	I_2	I_3	I_4	I_5
g_1^1	1.34	5.71	1.21	2.13	4.22
g_2^1	5.15	0.75	3.88	6.32	0.75

D'où les deux groupes :

$G_1 = \{I_1, I_3, I_4\}$ et $G_2 = \{I_2, I_5\}$

On retrouve la même classification que l'étape précédente, on arrête l'algorithme.