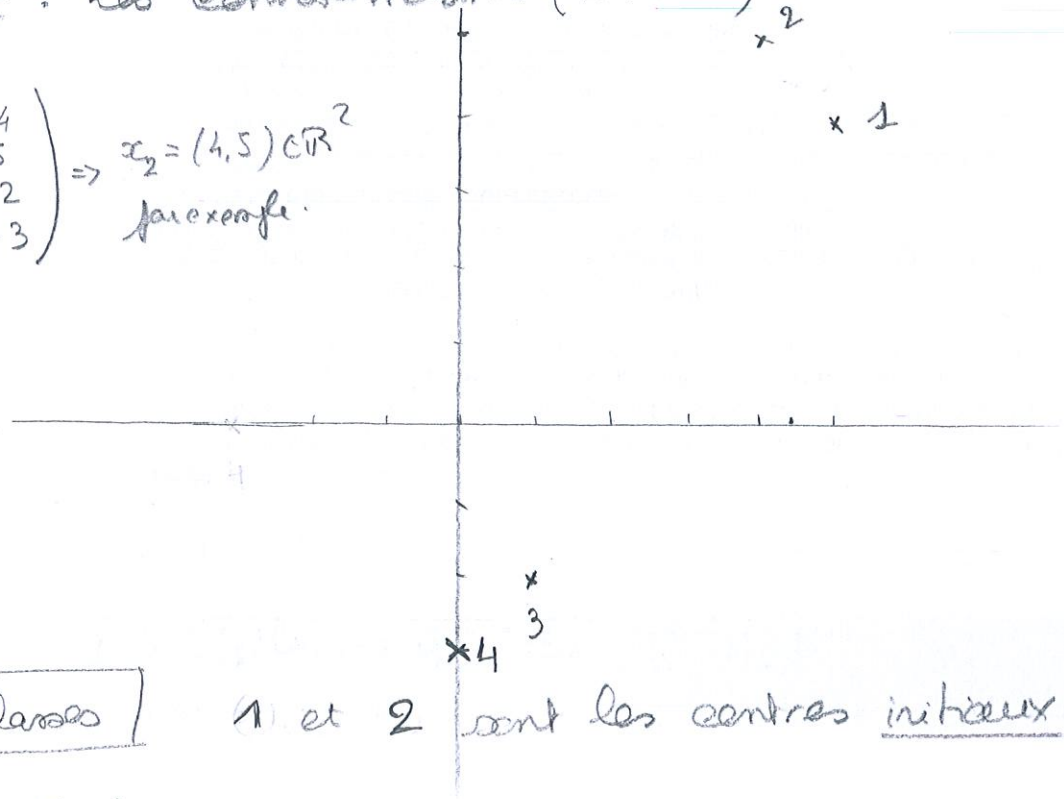
TP: les algo de classification

Exercice 1 : les centres-mobiles (k-means)

Données :

$$X = \begin{pmatrix} 5 & 4 \\ 4 & 5 \\ 1 & -2 \\ 0 & -3 \end{pmatrix} \Rightarrow x_2 = (4,5) \in \mathbb{R}^2$$

par exemple.



$k = 2$ classes — 1 et 2 sont les centres initiaux de C_1 et C_2

(a) Initialisation

$g_1 = (5,4)$

$g_2 = (4,5)$

(b) Affectation :

$1 \rightarrow C_1$

$2 \rightarrow C_2$

$3 \rightarrow C_1$ car $d^2(3,g_1) = (1-5)^4 + (-2-4)^2 = 4^2 + 6^2 = 52$

$d^2(3,g_2) = (1-4)^2 + (-2-5)^2 = 3^2 + 7^2 = 58$

$4 \rightarrow C_1$ car $d^2(4,g_1) = 5^2 + 7^2 = 74$

$d^2(4,g_2) = 4^2 + 8^2 = 80$

$\Rightarrow C_1 = \{1,3,4\}$ $C_2 = \{2\}$ test = 1 (au moins 1 individu a changé de classe)

(c) Representation : $w_i = 1$, $i = 1-4$ $\Rightarrow \mu_k = m_k$ = nb. d'individu dans C_k.

$$g_1 = \frac{1}{3}(x_1 + x_3 + x_4) = \frac{1}{3}((5,4)+(1,-2)+(0,-3)) = \left(\frac{5+1+0}{3}, \frac{4-2-3}{3}\right)$$
$$= (2, -1/3)$$

$$g_2 = x_2 = (4,5)$$

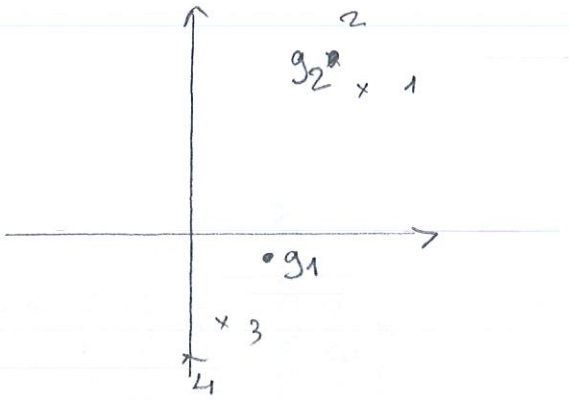
(b) Affectation

$1 \to C_2$ car $d(1,g_2) < d(1,g_1)$

$2 \to C_2$

$3 \to C_1$ car $d(3,g_1) < d(3,g_2)$

$4 \to C_1$

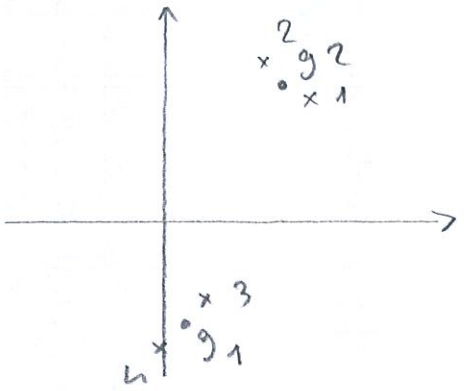


$\Rightarrow C_1 = \{3,4\}$ $C_2 = \{1,2\}$ test = 1 (car m 1 indiv. a changé de classe)

(c) Representat° :

$$g_1 = \frac{x_3 + x_4}{2} = \left(\frac{1}{2}, -\frac{5}{2}\right)$$

$$g_2 = \frac{x_1 + x_2}{2} = \left(\frac{9}{2}, \frac{9}{2}\right)$$



(b) Affectation

$1 \to C_2$
$2 \to C_2$ $\Rightarrow C_1 = \{3,4\}$ $C_2 = \{1,2\}$ test = 0 (personne n'a changé de classe)
$3 \to C_1$
$4 \to C_1$

$\Rightarrow$ ARRET

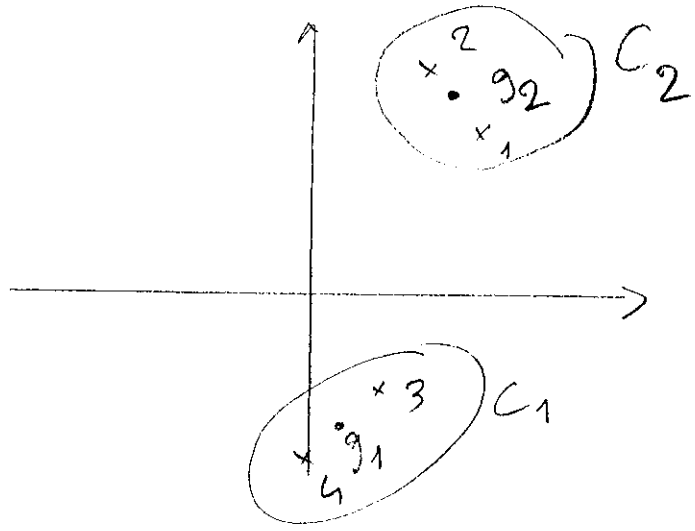
$\Rightarrow$ Sortie de l'algo :

$P = (C_1, C_2)$ avec $C_1 = \{3,4\}$
$C_2 = \{1,2\}$.

Qualité de cette partition en terme d'inertie intra-classe.

avec $M = I$, $w_i = 1 \ \forall i$

$$W = \underbrace{I(C_1)}_{\text{inertie}} + I(C_2) = \underbrace{d^2(3,g_1) + d^2(4,g_1)}_{I(C_1)} + \underbrace{d^2(1,g_2) + d^2(2,g_2)}_{I(C_2)}$$



$$d^2(3,g_1) = \left(1 - \frac{1}{2}\right)^2 + \left(-2 + \frac{5}{2}\right)^2 = 0,5^2 + 0,5^2 = 0,5$$

$$d^2(4,g_1) = \left(0 - \frac{1}{2}\right)^2 + \left(-3 + \frac{5}{2}\right)^2 = 0,5^2 + 0,5^2 = 0,5$$

$$d^2(1,g_2) = \left(5 - \frac{9}{2}\right)^2 + \left(4 - \frac{9}{2}\right)^2 = 0,5^2 + 0,5^2 = 0,5$$

$$d^2(2,g_2) = d^2(1,g_2) = 0,5$$

$$\Rightarrow \boxed{W = 2}$$

En terme de pourcentage d'inertie avec $w_i = 1$

$$T = \underbrace{\sum_{i=1}^{n} w_i d^2(x_i, g)}_{\text{inertie totale}} = \sum_{j=1}^{p} \sum_{i=1}^{n} w_i \left(x_{ij} - g_j\right)^2 = 67$$

$T = W + B$

W : Min, B : Max

$$\Rightarrow \frac{T - W}{T} = \frac{B}{T} = \text{prop}^\circ \text{ d'inertie expliquée}$$

$$= \frac{67 - 2}{67} = \frac{65}{67} = 0,9701$$

$\Rightarrow 97,01\%$

Exercice 2 : la classification hiérarchique

* Mesure d'agrégation du lien max : $D(A,B) = \max_{i \in A, j \in B} d(i,j)$ (distance euclidienne)

* indice h : $\begin{cases} h(A \cup B) = D(A,B) \\ h(\{i\}) = 0 \quad \forall i = 1 \dots n \end{cases}$

(a) Initialisation : $P = (\{1\}\{2\}\{3\}\{4\})$: partition des singletons.

$D(\{1\},\{2\}) = d(1,2) = 1.4$
$D(\{1\},\{3\}) = d(1,3) = 7.2$
$D(\{1\},\{4\}) = d(1,4) = 8.6$
$D(\{2\},\{3\}) = d(2,3) = 7.6$
$D(\{2\},\{4\}) = d(2,4) = 8.9$
$D(\{3\},\{4\}) = d(3,4) = 1.4$

⇒ Matrice de distances :

	1	2	3	4
1	0	1.4	7.2	8.6
2	1.4	0	7.6	8.9
3	7.2	7.6	0	1.4
4	8.6	8.9	1.4	0

(Mtx symétrique)

⇒ Agrège $\{1\}$ et $\{2\}$ au $\{3\}$ et $\{4\}$

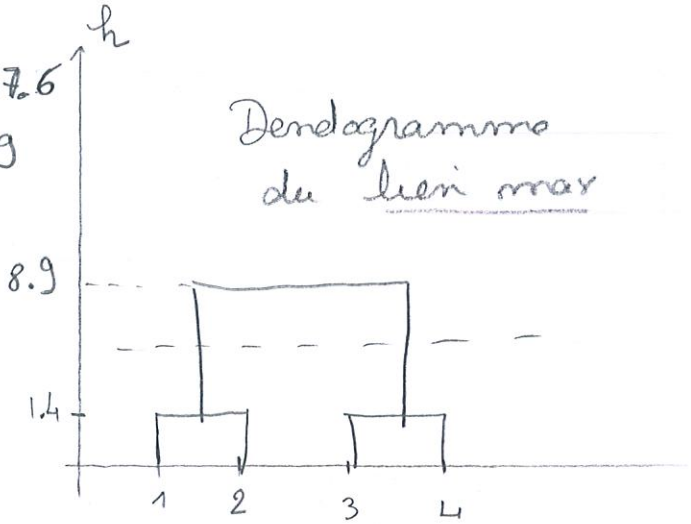
$h(\{1,2\}) = D(\{1\},\{2\}) = 1.4$

(b) $P = (\{1,2\},\{3\},\{4\})$

$D(\{1,2\},\{3\}) = \max(d(1,3), d(2,3)) = 7.6$
$D(\{1,2\},\{4\}) = \max(d(1,4), d(2,4)) = 8.9$
$D(\{3\},\{4\}) = 1.4$

⇒ Aggrège $\{3\}$ et $\{4\}$

$h(\{3,4\}) = 1.4$



(b') $P = (\{1,2\}\{3,4\})$

⇒ Agrège $\{1,2\}$ avec $\{3,4\}$

$h(\{1,2,3,4\}) = D(\{1,2\},\{3,4\}) = \max(d(1,3), d(1,4), d(2,3), d(2,4)) = 8,9$

* Mesure d'agrégation de Ward : $D(A,B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} d^2(g_A, g_B)$

$\mu_A = \sum_{i \in A} w_i$

* On prend $w_i = 1$, $i = 1 \dots 4$.

(a) Initialisat° : $P = (\{1\}, \{2\}, \{3\}, \{4\})$

$D(\{1\}, \{2\}) = \frac{1}{2} d^2(1,2) = 1$

$\vdots$

$D(\{3\}, \{4\}) = \frac{1}{2} d^2(3,4) = 1$

On agrège $\{1\}$ et $\{2\}$ et $h\{1,2\} = 1$

⇒ Matrice des distances au carré

$$\begin{pmatrix} 0 & & & \\ 2 & 0 & & \\ 52 & 58 & 0 & \\ 74 & 80 & 2 & 0 \end{pmatrix}$$

(b) $P = (\{1,2\}, \{3\}, \{4\})$

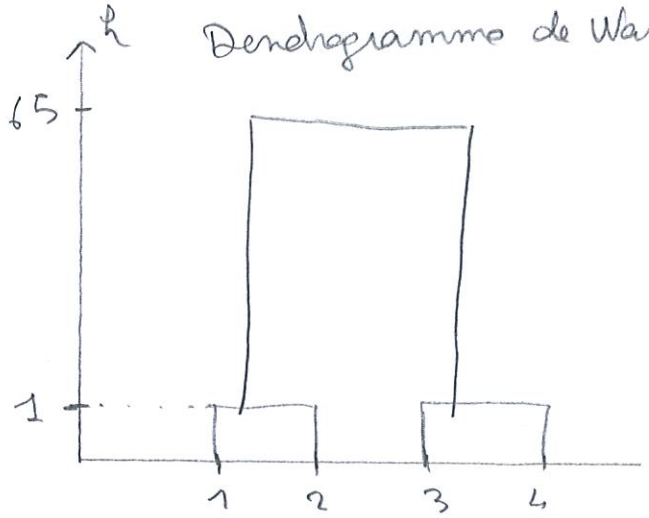
$D(\{1,2\}, \{3\}) = \frac{2}{3} d^2(g_2, 3)$

avec $g_2 = \frac{x_1 + x_2}{2}$

$D(\{1,2\}, \{4\}) = \frac{2}{3} d^2(g_2, 4)$

$D(\{3\}, \{4\}) = 1$

Dendrogramme de Ward



On agrège $\{3\}$ et $\{4\}$ et $h\{3,4\} = 1$

(b') $P = (\{1,2\}, \{3,4\})$

On agrège $\{1,2\}$ avec $\{3,4\}$ et

$h(\{1,2,3,4\}) = D(\{1,2\}, \{3,4\}) = \frac{4}{4} d^2(g_2, g_1) = 65$

avec $g_2 = (4,5 \quad 4,5)$

$g_1 = (0,5, -2,5)$

Pour Ward :

Dans R : avec hclust, on retrouve $2 * h$

⚠ il faut passer d^2.