# TD : Non-supervisée

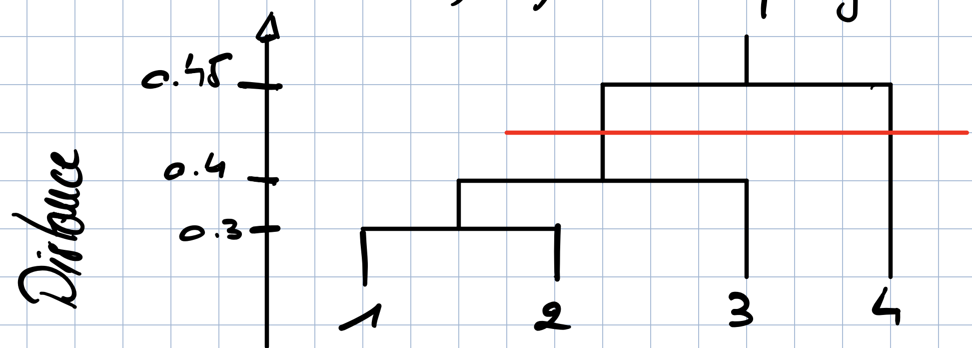
## Exercice 1 :

Matrice de distance.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 0.3 | 0.4 | 0.7 |
| 2 | 0.3 | 0 | 0.5 | 0.8 |
| 3 | 0.4 | 0.5 | 0 | 0.45 |
| 4 | 0.7 | 0.8 | 0.45 | 0 |

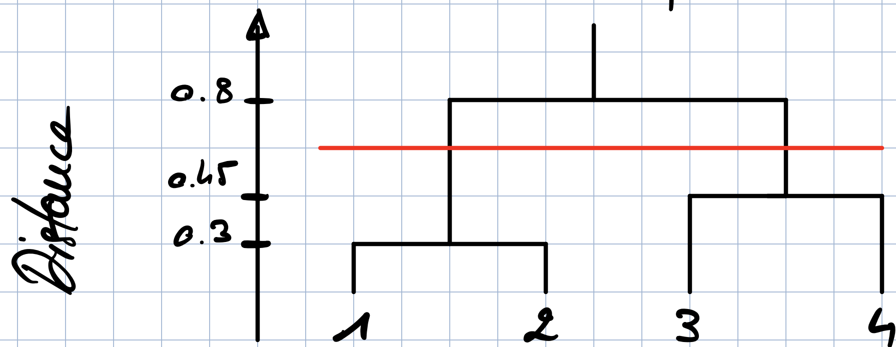
1. Méthode du plus proche voisin :

$$d(A,B) = \min \{ d_{ij} : i \in A, j \in B \}.$$




2. Méthode du voisin le plus distant :

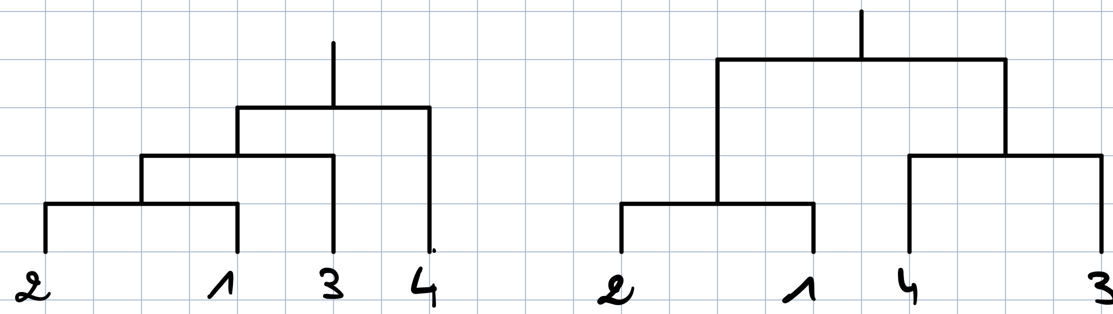
$$d(A,B) = \max \{ d_{ij} : i \in A, j \in B \}$$



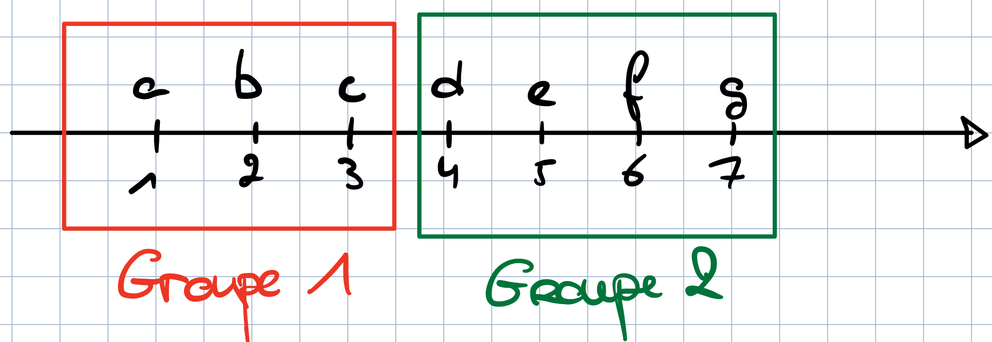
3. Si on coupe le premier arbre de sorte à avoir 2 groupes :
$\{1,2,3\}, \{4\}$

Si on coupe le deuxième arbre de sorte à avoir 2 groupes :
$\{1,2\}, \{3,4\}$.

4.




## Exercice 2 :




1. $S(\{c\}) = \dfrac{b_c - a_c}{\max(a_c, b_c)}$

avec $a_c$ = distance moyenne entre $\{c\}$ et $\{a\}$ et $\{b\}$.

$d(a,c) = 2$ , $d(b,c) = 1 \rightarrow a_c = \frac{1}{2}(2+1) = \frac{3}{2}$

$b_c$ = distance moyenne entre $\{c\}$ et $\{d\}, \{e\}, \{f\}$ et $\{g\}$.

$d(d,c) = 1$, $d(e,c) = 2$, $d(f,c) = 3$ et $d(g,c) = 4$.

$\rightarrow b_c = \frac{1}{4}(1+2+3+4) = \frac{5}{2}$.

donc $S(\{c\}) = \dfrac{\frac{5}{2} - \frac{3}{2}}{5/2} = \dfrac{2}{5} \rightarrow$ c bien classé.

2. $\text{Pseudo\_}R^2 = \dfrac{I_{inter}}{I_{tot}}$. $G = \{d\}$.

$$I_{tot} = \frac{1}{n}\sum_{i=1}^{n} d^2(X_i, G)$$
$$= \frac{1}{7}\left(3^2+2^2+1^2+0+1^2+2^2+3^2\right)$$
$$= \frac{28}{7} = 4$$

$$I_{inter} = \frac{1}{n}\sum_{k=1}^{K} n_k\, d^2(G_k, G) \quad \text{avec } G_1 = \{b\},\ G_2 = \{5.5\}$$
$$= \frac{1}{7}\left(3\times 2^2 + 4\times\left(\frac{3}{2}\right)^2\right) = \frac{21}{7} = 3.$$

$$I_{intra} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k} d^2(X_i, G_k) = \frac{1}{7}\left(1^2+0^2+1^2+\left(\frac{3}{2}\right)^2+\left(\frac{1}{2}\right)^2+\left(\frac{3}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)$$
$$= \frac{7}{7} = 1$$

$\text{Pseudo-}R^2 = \dfrac{3}{4} \longrightarrow$ le clustering est bon.

Exercice 3 :

| Obs | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_{i1}$ | -1 | -0.5 | 0 | 0.5 | 1 |
| $x_{i2}$ | -1 | 0 | 0.5 | -0.5 | 1 |

1. On assigne aléatoirement les obs. 1, 2 et 5 au groupe 1 et les observations 3 et 5 au groupe 2.

Donc $C(1) = C(2) = C(5) = 1$

$C(3) = C(4) = 2$

2. $$\mu_1 = \frac{1}{3}\left\{\begin{pmatrix}-1\\-1\end{pmatrix}+\begin{pmatrix}-0.5\\0\end{pmatrix}+\begin{pmatrix}1\\1\end{pmatrix}\right\} = \begin{pmatrix}-1/6\\0\end{pmatrix}$$

$$\mu_2 = \frac{1}{2}\left\{\begin{pmatrix}0\\0.5\end{pmatrix}+\begin{pmatrix}0.5\\-0.5\end{pmatrix}\right\} = \begin{pmatrix}1/4\\0\end{pmatrix}$$

3.

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $d(i,\mu_1)$ | 1.69 | 0.11 | 0.28 | 0.69 | 2.36 |
| $d(i,\mu_2)$ | 2.56 | 0.56 | 0.31 | 0.31 | 1.87 |

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \sum_{j:C(j)=k} d(X_i, X_j)$$

$$= d(X_1, X_2) + d(X_1, X_4) + d(X_2, X_4) + d(X_3, X_4)$$

4. On assigne l'obs $i$ au groupe le plus proche, donc si $d(i,\mu_1) < d(i,\mu_2)$, on assigne au groupe 1, et au groupe 2, sinon.

Groupe 1 devient : $\{1, 2, 3\}$

Groupe 2 devient : $\{4, 5\}$.

5. $$\mu_1 = \frac{1}{3}\left\{\begin{pmatrix}-1\\-1\end{pmatrix} + \begin{pmatrix}-0.5\\0\end{pmatrix} + \begin{pmatrix}0\\0.5\end{pmatrix}\right\} = \begin{pmatrix}-1/2\\-1/6\end{pmatrix}.$$

$$\mu_2 = \frac{1}{2}\left\{\begin{pmatrix}0.5\\-0.5\end{pmatrix} + \begin{pmatrix}1\\1\end{pmatrix}\right\} = \begin{pmatrix}3/4\\1/4\end{pmatrix}.$$

6.

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $d(i,\mu_1)$ | 0.94 | 0.03 | 0.69 | 1.11 | 3.61 |
| $d(i,\mu_2)$ | 4.63 | 1.63 | 0.63 | 0.63 | 0.63 |

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \sum_{j:C(j)=k} d(X_i, X_j)$$

$$= d(X_1, X_2) + d(X_1, X_3) + d(X_2, X_3) + d(X_4, X_5)$$

Les groupes deviennent :

Groupe 1 → $\{1, 2\}$

Groupe 2 → $\{3, 4, 5\}$.

Si on refait les calculs, les obs. ne changent plus de groupe après cet étape.