**Master 1 Informatique – ENSICAEN 2A**

**Module « Apprentissage »**

**TD numéro 2 Correction**

*Apprentissage non supervisé - Clustering*

---

# 1 Classification Ascendante Hiérarchique

On dispose d'un ensemble $\mathcal{X} = \{x_1, \cdots, x_7\}$ ainsi qu'une mesure $d$ définie sur l'ensemble des couples de $\mathcal{X}$, dont les valeurs sont précisées sur le tableau ci-dessous.

| $d$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ |
|---|---|---|---|---|---|---|---|
| $x_1$ | 0 | 2 | 4.5 | 5.5 | 7.5 | 9.5 | 4 |
| $x_2$ | 2 | 0 | 2.5 | 3.5 | 5.5 | 7.5 | 4 |
| $x_3$ | 4.5 | 2.5 | 0 | 3 | 5 | 7 | 6.5 |
| $x_4$ | 5.5 | 3.5 | 3 | 0 | 2 | 4 | 7.5 |
| $x_5$ | 7.5 | 5.5 | 5 | 2 | 0 | 4 | 9.5 |
| $x_6$ | 9.5 | 7.5 | 7 | 4 | 4 | 0 | 5.5 |
| $x_7$ | 4 | 4 | 6.5 | 7.5 | 9.5 | 5.5 | 0 |

1. La mesure $d$ est-elle une distance métrique ? Pour être une métrique la distance doit respecter :

$$d(x,y) = 0 \Leftrightarrow x = y \quad \text{séparation} \quad (1)$$
$$d(x,y) = d(y,x) \quad \text{symmétrique} \quad (2)$$
$$d(x,y) \leq d(x,z) + d(z,y) \quad \text{inégalité triangulaire} \quad (3)$$
$$d(x,y) \geq 0 \quad \text{positivité} \quad (4)$$

C'est bien le cas, par exemple : $d(x_5, x_7) = 9.5 \leq d(x_5, x_2) + d(x_2, x_7) = 5.5 + 4 = 9.5$

2. Construisez et représentez graphiquement la hiérarchie obtenue par une CAH avec complete linkage (agrégation du lien maximum).

Les étapes de la fusion :

| d | Clusters |
|---|---|
| 0 | $\{x_1\}\{x_2\}\{x_3\}\{x_4\}\{x_5\}\{x_6\}\{x_7\}$ |
| 2 | $\{x_1, x_2\}\{x_3\}\{x_4\}\{x_5\}\{x_6\}\{x_7\}$ |
| 2 | $\{x_1, x_2\}\{x_3\}\{x_4, x_5\}\{x_6\}\{x_7\}$ |
| 4 | $\{x_1, x_2, x_7\}\{x_3\}\{x_4, x_5\}\{x_6\}$ |
| 4 | $\{x_1, x_2, x_7\}\{x_3\}\{x_4, x_5, x_6\}$ |
| 6.5 | $\{x_1, x_2, x_7, x_3\}\{x_4, x_5, x_6\}$ |
| 9.5 | $\{x_1, x_2, x_7, x_3, x_4, x_5, x_6\}$ |

La mise à jour de la matrice avec le lien maximal :

| $d$ | $\{x_1, x_2\}$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ |
|---|---|---|---|---|---|---|
| $\{x_1, x_2\}$ | 0 | 4.5 | 5.5 | 7.5 | 9.5 | 4 |
| $x_3$ | | 0 | 3 | 5 | 7 | 6.5 |
| $x_4$ | | | 0 | 2 | 4 | 7.5 |
| $x_5$ | | | | 0 | 4 | 9.5 |
| $x_6$ | | | | | 0 | 5.5 |
| $x_7$ | | | | | | 0 |

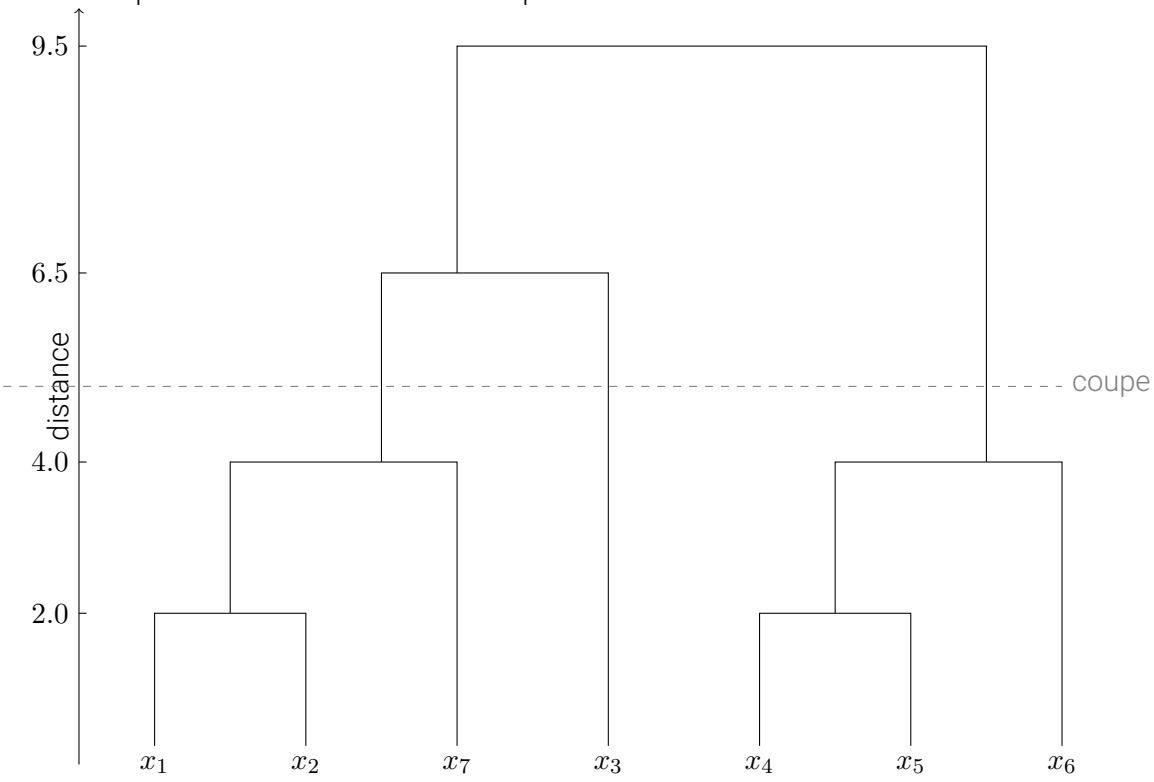
| $d$ | $\{x_1, x_2\}$ | $x_3$ | $\{x_4, x_5\}$ | $x_6$ | $x_7$ |
|---|---|---|---|---|---|
| $\{x_1, x_2\}$ | 0 | 4.5 | 7.5 | 9.5 | 4 |
| $x_3$ | | 0 | 5 | 7 | 6.5 |
| $\{x_4, x_5\}$ | | | 0 | 4 | 9.5 |
| $x_6$ | | | | 0 | 5.5 |
| $x_7$ | | | | | 0 |

| $d$ | $\{x_1, x_2, x_7\}$ | $x_3$ | $\{x_4, x_5\}$ | $x_6$ |
|---|---|---|---|---|
| $\{x_1, x_2, x_7\}$ | 0 | 6.5 | 9.5 | 9.5 |
| $x_3$ | | 0 | 5 | 7 |
| $\{x_4, x_5\}$ | | | 0 | 4 |
| $x_6$ | | | | 0 |

| $d$ | $\{x_1, x_2, x_7\}$ | $x_3$ | $\{x_4, x_5, x_6\}$ |
|---|---|---|---|
| $\{x_1, x_2, x_7\}$ | 0 | 6.5 | 9.5 |
| $x_3$ | | 0 | 7 |
| $\{x_4, x_5, x_6\}$ | | | 0 |

| $d$ | $\{x_1, x_2, x_7, x_3\}$ | $\{x_4, x_5, x_6\}$ |
|---|---|---|
| $\{x_1, x_2, x_7, x_3\}$ | 0 | 9.5 |
| $\{x_4, x_5, x_6\}$ | | 0 |

3. Donnez la partition en $3$ classes obtenue par la hiérarchie.




## 2 K-moyennes

On dispose d'un ensemble de points 2D $\mathcal{X} = \{x_1, \cdots, x_8\}$ que l'on souhaite regrouper en $3$ clusters à l'aide de la méthode des $k$-moyennes. Les exemples sont $x_1 = (2, 10)$, $x_2 = (2, 5)$, $x_3 = (8, 4)$, $x_4 = (5, 8)$, $x_5 = (7, 5)$, $x_6 = (6, 4)$, $x_7 = (1, 2)$, $x_8 = (4, 9)$, On utilise la distance Euclidienne pour et cela donne la matrice ci-dessous des distances (non remplie sous la diagonale) :

On suppose que les centres initiaux sont $x_1$, $x_4$ et $x_7$ et nommés $S_1$, $S_2$, et $S_3$. Faites une itération de l'algorithme des $k$-moyennes. À chaque itération vous spécifierez :

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 0 | $\sqrt{25}$ | $\sqrt{36}$ | $\sqrt{13}$ | $\sqrt{50}$ | $\sqrt{52}$ | $\sqrt{65}$ | $\sqrt{5}$ |
| $x_2$ | | 0 | $\sqrt{37}$ | $\sqrt{18}$ | $\sqrt{25}$ | $\sqrt{17}$ | $\sqrt{10}$ | $\sqrt{20}$ |
| $x_3$ | | | 0 | $\sqrt{25}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{53}$ | $\sqrt{41}$ |
| $x_4$ | | | | 0 | $\sqrt{13}$ | $\sqrt{17}$ | $\sqrt{52}$ | $\sqrt{2}$ |
| $x_5$ | | | | | 0 | $\sqrt{2}$ | $\sqrt{45}$ | $\sqrt{35}$ |
| $x_6$ | | | | | | | $\sqrt{29}$ | $\sqrt{29}$ |
| $x_7$ | | | | | | | 0 | $\sqrt{58}$ |
| $x_8$ | | | | | | | | 0 |

1. Les exemples affectés à chaque cluster,
2. Les centres des clusters,
3. L'inertie intra cluster,
4. Les exemples et clusters sur une grille $10 \times 10$.

Combien faut-il d'itérations pour que l'algorithme converge ? L'inertie intra-cluster diminue-t-elle ?
On utilise la distance Euclidienne $\|x-y\|_2 = \left(\sum_{i=1}^{d} |x_i - y_i|^2\right)^{1/2}$ donnée par la matrice.

**Itération 1:**

$x_1$ : C'est le centre $C_1$, $\Rightarrow x_1 \in C_1$

$x_2$ :

$$d(x_2, S_1) = d(x_2, x_1) = \sqrt{25} = 5$$
$$d(x_2, S_2) = d(x_2, x_4) = \sqrt{18} = 4.24$$
$$d(x_2, S_3) = d(x_2, x_7) = \sqrt{10} = 3.16 \rightsquigarrow \text{le plus petit} \Rightarrow x_2 \in C_3$$

$x_3$ :

$$d(x_3, S_1) = d(x_3, x_1) = \sqrt{36} = 6$$
$$d(x_3, S_2) = d(x_3, x_4) = \sqrt{25} = 5 \rightsquigarrow \text{le plus petit} \Rightarrow x_3 \in C_2$$
$$d(x_3, S_3) = d(x_3, x_7) = \sqrt{53} = 7.28$$

$x_4$ : C'est le centre $C_2$, $\Rightarrow x_4 \in C_2$

$x_5$ :

$$d(x_5, S_1) = d(x_5, x_1) = \sqrt{50} = 7.07$$
$$d(x_5, S_2) = d(x_5, x_4) = \sqrt{13} = 3.60 \rightsquigarrow \text{le plus petit} \Rightarrow x_5 \in C_2$$
$$d(x_5, S_3) = d(x_5, x_7) = \sqrt{45} = 6.70$$

$x_6$ :

$$d(x_6, S_1) = d(x_6, x_1) = \sqrt{52} = 7.21$$
$$d(x_6, S_2) = d(x_6, x_4) = \sqrt{17} = 4.12 \rightsquigarrow \text{le plus petit} \Rightarrow x_6 \in C_2$$
$$d(x_6, S_3) = d(x_6, x_7) = \sqrt{29} = 5.38$$

$x_7$ : C'est le centre $C_3$, $\Rightarrow x_7 \in C_3$
$x_8$ :

$$d(x_8, S_1) = d(x_8, x_1) = \sqrt{5}$$
$$d(x_8, S_2) = d(x_8, x_4) = \sqrt{2} \rightsquigarrow \text{le plus petit} \Rightarrow x_8 \in C_2$$
$$d(x_8, S_3) = d(x_8, x_7) = \sqrt{58}$$

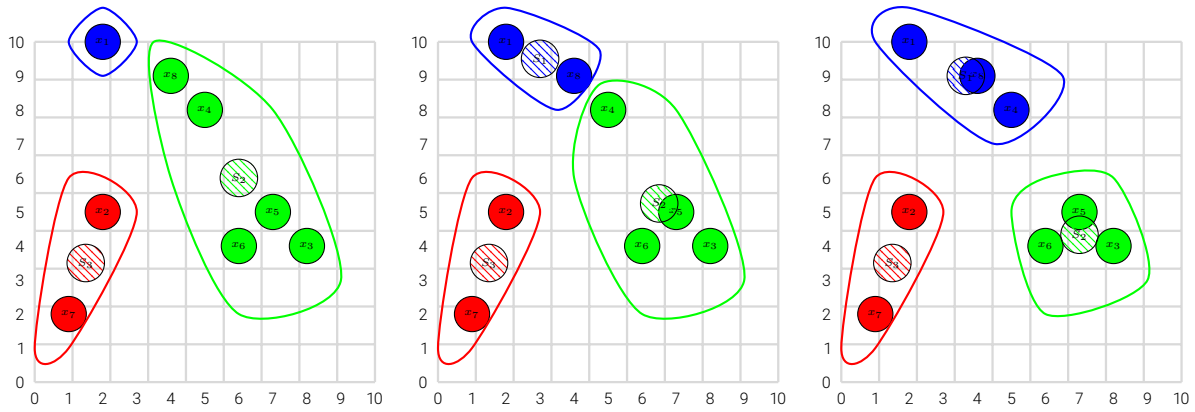
Les clusters sont $C_1 = \{x_1\}, C_2 = \{x_3, x_4, x_5, x_6, x_8\}, C_3 = \{x_2, x_7\}$.
Les nouveaux centres des clusters sont :
$S_1 = x_1 = (2, 10)$
$S_2 = ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)$
$S_3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$
L'inertie Intra-cluster est $J_w = \sum_{k=1}^{K} \sum_{x \in C_k} d^2(x, \mu_k)$.
$J_w = 0 + [(8-6)^2 + (4-6)^2 + (5-6)^2 + (8-6)^2 + (7-6)^2 + (5-6)^2 + (6-6)^2 + (4-6)^2 + (4-6)^2 + (9-6)^2] + [(2-1.5)^2 + (5-3.5)^2 + (1-1.5)^2 + (2-3.5)^2] = 0 + 32 + 5 = 35$

**Itération 2:**

Les clusters sont $C_1 = \{x_1, x_8\}, C_2 = \{x_3, x_4, x_5, x_6\}, C_3 = \{x_2, x_7\}$.
Les nouveaux centres des clusters sont :
$S_1 = ((2+4)/2, (10+9)/2) = (3, 9.5)$
$S_2 = ((8+5+7+6)/4, (4+8+5+4)/4) = (6.5, 5.25)$
$S_3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$
L'inertie Intra-cluster est $J_w = [(2-3)^2 + (10-9.5)^2 + (4-3)^2 + (9-9.5)^2] + [(8-6.5)^2 + (4-5.25)^2 + (5-6.5)^2 + (8-5.25)^2 + (7-6.5)^2 + (5-5.25)^2 + (6-6.5)^2 + (4-5.25)^2] + [(2-1.5)^2 + (5-3.5)^2 + (1-1.5)^2 + (2-3.5)^2] = 2.5 + 15.75 + 5 = 23.25$

**Itération 3:**

Les clusters sont $C_1 = \{x_1, x_4, x_8\}, C_2 = \{x_3, x_5, x_6\}, C_3 = \{x_2, x_7\}$.
Les nouveaux centres des clusters sont :
$S_1 = ((2+5+4)/3, (10+8+9)/3) = (3.66, 9)$
$S_2 = ((8+7+6)/3, (4+5+4)/3) = (7, 4.33)$
$S_3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$
L'inertie Intra-cluster est $J_w = [(2-3.66)^2 + (10-9)^2 + (5-3.66)^2 + (8-9)^2 + (4-3.66)^2 + (9-9)^2] + [(8-7)^2 + (4-4.33)^2 + (7-7)^2 + (5-4.33)^2 + (6-7)^2 + (4-4.33)^2] + [(2-1.5)^2 + (5-3.5)^2 + (2-1.5)^2 + (2-3.5)^2] = 6.66 + 2.66 + 5 = 14.32$



Itération 1 Itération 2 Itération 3

## 3 GMM

On dispose d'une base de $100$ exemples répartis dans $3$ clusters modélisés par des Gaussiennes. Le cluster $A$ contient $30\%$ des points. Sa moyenne est $\mu_A = (2, 2)$ et sa matrice de covariance est $\Sigma_A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$. Le cluster $B$ contient $20\%$ des points. Sa moyenne est $\mu_B = (5, 3)$ et sa matrice de covariance est $\Sigma_B = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$. Le cluster $C$ contient $50\%$ des points. Sa moyenne est $\mu_C = (1, 4)$ et sa matrice de covariance est $\Sigma_C = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}$. Calculez les probabilités d'appartenance du point $p = (2.5, 3.0)$ aux clusters $A$, $B$ et $C$.

On sait que :

Les probabilités d'appartenance sont estimées par $\gamma_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x_i}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x_i}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$

On a des distrcibutions Gaussiennes : $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}$

Les probabilités a priori sont $\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$. On part avec des affectations dures des points des clusters et donc les probabilités a priori sont les proportions de chaque cluster.

```
import math
import numpy as np
from numpy.linalg import det
p=np.array([2.5,3])
muA=np.array([2,2])
covA=np.array([[3,0],[0,3]])
muB=np.array([5,3])
covB=np.array([[2,1],[1,4]])
muC=np.array([1,4])
covC=np.array([[16,0],[0,4]])
NA=math.exp(-0.5*(p-muA).T.dot(inv(covA)).dot(p-muA))/(2*math.pi*(det(covA)**0.
                                                    5))
NB=math.exp(-0.5*(p-muB).T.dot(inv(covB)).dot(p-muB))/(2*math.pi*(det(covB)**0.
                                                    5))
NC=math.exp(-0.5*(p-muC).T.dot(inv(covC)).dot(p-muC))/(2*math.pi*(det(covC)**0.
                                                    5))
N=0.3*NA+0.2*NB+0.5*NC
pA=0.3*NA/N
pB=0.2*NB/N
pC=0.5*NC/N
print (pA, pB, pC)
```

## 4 DBSCAN

On dispose d'en ensemble de points 2D $\mathcal{X} = \{x_1, \cdots, x_{20}\}$ que l'on découper à l'aide de l'algorithme DBSCAN. Les points sont répartis comme cela est présenté sur la figure suivante. Vous utiliserez la distance de Manhattan entre les points $d_M(x_i, x_j) = \|x_i - x_j\|_1 = \left(\sum_{k=1}^{2} |x_i^k - x_j^k|\right)$. À l'aide de DBSCAN, déterminez quels points sont des points core, border ou noise. Les paramètres de DBSCAN seront : $\epsilon = 2$ et $minPts = 3$. Les points sont supposés tirés au hasard selon leur numérotation.

$x_1$ est un core point (3 points dans son voisinage) et $x_2$, $x_3$ et $x_4$ sont density reachable de $x_1 \Rightarrow$ 1 cluster est créé).

$x_5$ est un noise point (pas de points dans son voisinage).
$x_6$ est un border point (seulement 2 points dans le voisinage).
$x_7$ est un core point et $x_6$, $x_8$, $x_9$ et $x_{10}$ sont density reachable de $x_7 \Rightarrow$ 1 cluster est créé).
$x_{11}$ est un noise point (pas de points dans son voisinage).
$x_{12}$ est un noise point (pas de points dans son voisinage).
$x_{13}$ est un core point ($3$ points dans son voisinage) et $x_{14}$ à $x_{20}$ sont density reachable $\Rightarrow$ 1 cluster est créé).