Calcul Scientifique et Apprentissage Automatique

TD
Apprentissage Non Supervisé - Clustering

Elana Courtines
courtines.e@gmail.com
https://github.com/irinacake

Séance 1 - 4 octobre 2022

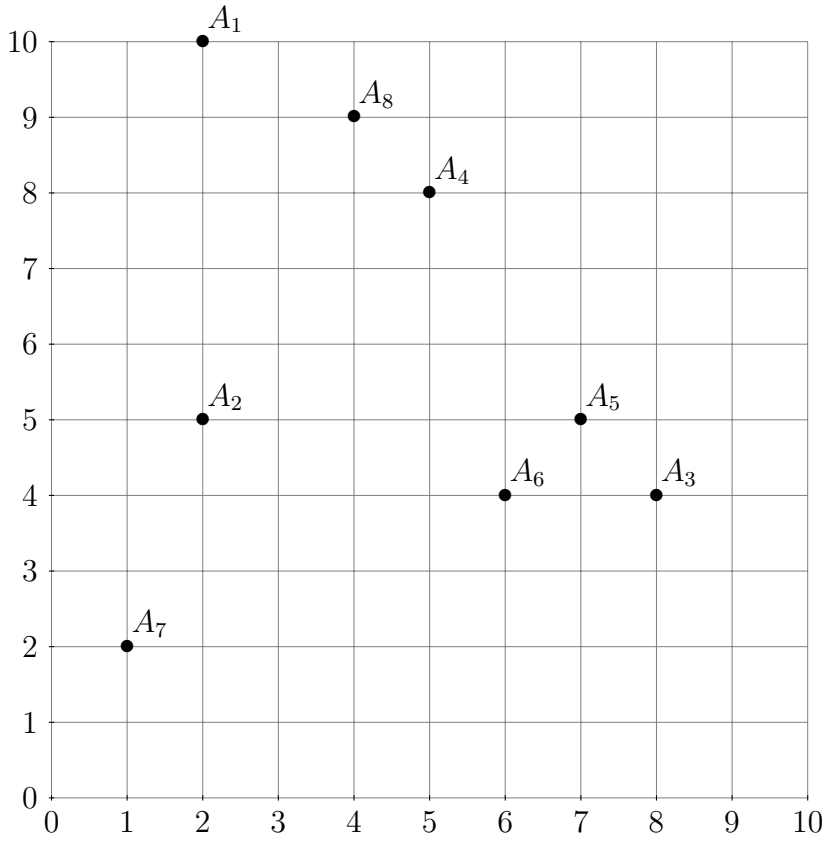
Sandrine Mouysset - sandrine.mouysset@irit.fr

Exercice 1

Avec les points suivants :

$$A1 = (2, 10), A2 = (2, 5), A3 = (8, 4), A4 = (5, 8)$$

$$A5 = (7, 5), A6 = (6, 4), A7 = (1, 2), A8 = (4, 9)$$



Kppv

Question 1 Calculer la matrice des distances euclidiennes au carré entre les points $(A_i)_{i=1..8}$:

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	25	72	13	50	52	65	5
A_2		0	37	18	25	17	10	20
A_3			0	25	2	4	53	41
A_4				0	13	17	52	2
A_5					0	2	45	25
A_6						0	29	29
A_7							0	58
A_8								0

Question 2 : Réaliser la méthode des K-ppv avec k = 1 et un seuil de 16 :

Pour ce faire, on parcourt la liste des points X_i dans l'ordre et :

1. On déterminer le point le plus proche ;
2. Si aucun des points n'est dans une classe, on en crée une et on y ajoute les deux points ;
3. Si ce point le plus proche appartient déjà à une classe, on y ajoute X_i

D'où :

- $\forall j \neq 1,\ d(A_1, A_8) \leq d(A_1, A_j)$
  or, $d(A_1, A_8) = 5 \leq seuil$
  d'où : $C^1 = \{A_1, A_8\}$
- $\forall j \neq 2,\ d(A_2, A_7) \leq d(A_2, A_j)$
  or, $d(A_2, A_7) = 10 \leq seuil$
  d'où : $C^2 = \{A_2, A_7\}$
- $\forall j \neq 3,\ d(A_3, A_5) \leq d(A_3, A_j)$
  or, $d(A_3, A_5) = 2 \leq seuil$
  d'où : $C^3 = \{A_3, A_5\}$
- $\forall j \neq 4,\ d(A_4, A_8) \leq d(A_4, A_j)$
  or, $d(A_4, A_8) = 2 \leq seuil$
  d'où : $C^1 = \{A_1, A_4, A_8\}$
- $\forall j \neq 5,\ d(A_5, A_3) = d(A_5, A_6) \leq d(A_1, A_j)$
  or, $d(A_5, A_3) = d(A_5, A_6) = 2 \leq seuil$
  d'où : $C^3 = \{A_3, A_5, A_6\}$
- *les étapes pour* A_6, A_7 *et* A_8 *ne changent rien au résultat*

Au final on a :

- $C^1 = \{A_1, A_4, A_8\}$
- $C^2 = \{A_2, A_7\}$
- $C^3 = \{A_3, A_5, A_6\}$

Kmeans

Question 1 : Réaliser une itération de la classification par K-means en prenant des centres initiaux $C^1 = A_1,\ C^2 = A_4$ et $C^3 = A_7$, c'est à dire $m_1 = (2, 10),\ m_2 = (4, 9),\ m_3 = (1, 2)$

Itération 1 :
Étape 1, "étiquetage de tous les points" :

- Pour A_1 : étant confondu avec C^1, $C^1 = \{A_1\}$
- Pour A_2 : $\forall j = 1, 2, 3,\ d(A_2, C^3) \leq d(A_2, C^j)$
  d'où : $C^3 = \{A_2\}$
- Pour A_3 : $\forall j = 1, 2, 3,\ d(A_3, C^2) \leq d(A_3, C^j)$
  d'où : $C^2 = \{A_3\}$

- Pour A_4 : étant confondu avec C^2, $C^2 = \{A_3, A_4\}$
- Pour A_5 : $\forall j = 1, 2, 3,\ d(A_5, C^2) \leq d(A_5, C^j)$
  d'où : $C^2 = \{A_3, A_4, A_5\}$
- Pour A_6 : $\forall j = 1, 2, 3,\ d(A_6, C^2) \leq d(A_6, C^j)$
  d'où : $C^2 = \{A_3, A_4, A_5, A_6\}$
- Pour A_7 : étant confondu avec C^3, $C^3 = \{A_2, A_7\}$
- Pour A_8 : $\forall j = 1, 2, 3,\ d(A_8, C^2) \leq d(A_8, C^j)$
  d'où : $C^2 = \{A_3, A_4, A_5, A_6, A_8\}$

Étape 2, mise à jour des centres :

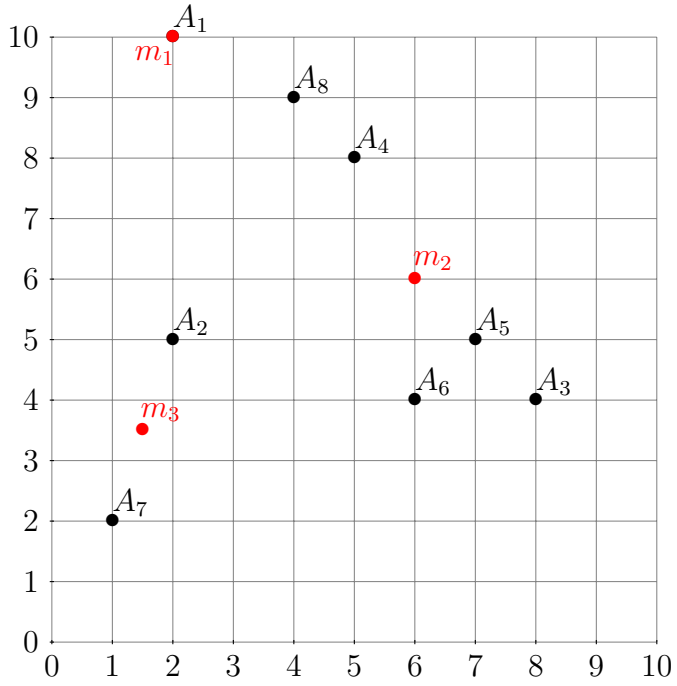
- $C^1 = milieu(A_1)$, d'où $m_1 = (2, 10)$
- $C^2 = milieu(A_3, A_4, A_5, A_6, A_8)$
  d'où m_2

$$= \left(\frac{x_{A_3} + x_{A_4} + x_{A_5} + x_{A_6} + x_{A_8}}{5} \ , \ \frac{y_{A_3} + y_{A_4} + y_{A_5} + y_{A_6} + y_{A_8}}{5} \right)$$
$$= \left(\frac{8 + 5 + 7 + 6 + 4}{5} \ , \ \frac{4 + 8 + 5 + 4 + 9}{5} \right)$$
$$= (6,\ 6)$$

- $C^3 = milieu(A_2, A_7)$
  d'où m_3

$$= \left(\frac{x_{A_2} + x_{A_7}}{2} \ , \ \frac{y_{A_2} + y_{A_7}}{2} \right)$$
$$= \left(\frac{2 + 1}{2} \ , \ \frac{5 + 2}{2} \right)$$
$$= (1.5,\ 3.5)$$

On a alors :



Matrice des distances euclidiennes au carré mise à jour pour les nouveaux centres :

	C^1	C^2	C^3
A_1	0	32	30.5
A_2	25	17	2.5
A_3	72	8	42.5
A_4	13	5	32.5
A_5	50	2	32.5
A_6	52	4	20.5
A_7	65	41	2.5
A_8	5	13	36.5

Itération 2 :
Étape 1, "étiquetage de tous les points" :

- Pour A_1 : $\forall j = 1, 2, 3,\ d(A_1, C^1) \leq d(A_1, C^j)$
  d'où : $C^1 = \{A_1\}$
- Pour A_2 : $\forall j = 1, 2, 3,\ d(A_2, C^3) \leq d(A_2, C^j)$
  d'où : $C^3 = \{A_2\}$
- Pour A_3 : $\forall j = 1, 2, 3,\ d(A_3, C^2) \leq d(A_3, C^j)$
  d'où : $C^2 = \{A_3\}$
- Pour A_4 : $\forall j = 1, 2, 3,\ d(A_4, C^2) \leq d(A_4, C^j)$
  d'où : $C^2 = \{A_3, A_4\}$
- Pour A_5 : $\forall j = 1, 2, 3,\ d(A_5, C^2) \leq d(A_5, C^j)$
  d'où : $C^2 = \{A_3, A_4, A_5\}$
- Pour A_6 : $\forall j = 1, 2, 3,\ d(A_6, C^2) \leq d(A_6, C^j)$
  d'où : $C^2 = \{A_3, A_4, A_5, A_6\}$
- Pour A_7 : $\forall j = 1, 2, 3,\ d(A_7, C^3) \leq d(A_7, C^j)$
  d'où : $C^3 = \{A_2, A_7\}$
- Pour A_8 : $\forall j = 1, 2, 3,\ d(A_8, C^1) \leq d(A_8, C^j)$
  d'où : $C^1 = \{A_1, A_8\}$

Étape 2, mise à jour des centres :

- $C^1 = milieu(A_1, A_8)$
  d'où m_1

  $= \left(\dfrac{x_{A_1} + x_{A_8}}{2} \ , \ \dfrac{y_{A_1} + y_{A_8}}{2}\right)$

  $= \left(\dfrac{2+4}{2} \ , \ \dfrac{10+9}{2}\right)$

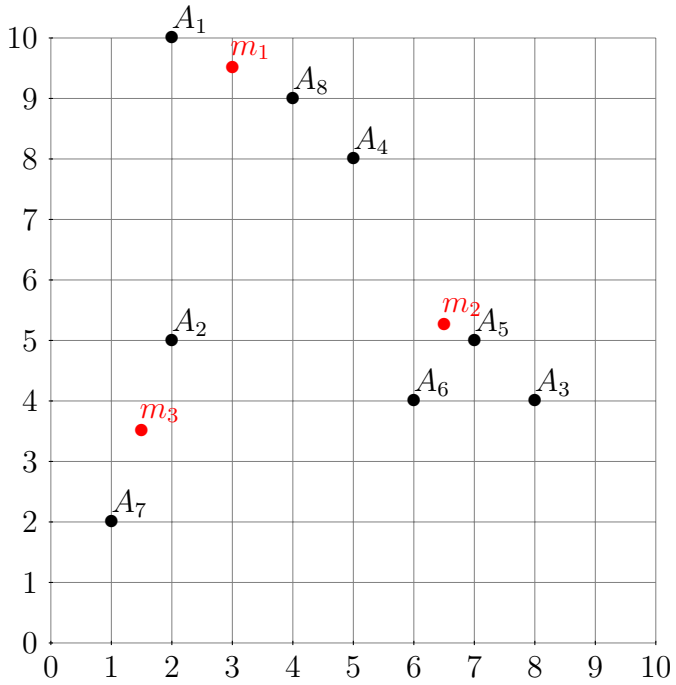
  $= (3,\ 9.5)$
- $C^2 = milieu(A_3, A_4, A_5, A_6)$
  d'où m_2

  $= \left(\dfrac{x_{A_3} + x_{A_4} + x_{A_5} + x_{A_6}}{4} \ , \ \dfrac{y_{A_3} + y_{A_4} + y_{A_5} + y_{A_6}}{4}\right)$

$$= \left(\frac{8+5+7+6}{4} \quad , \quad \frac{4+8+5+4}{4} \right)$$
$$= (6.5,\ 5.25)$$

- C^3 n'a pas changé, donc $m_3 = (1.5,\ 3.5)$

On a alors :



2. Obtient-on la même partition que pour les Kppv ?
À l'itération 1, non. Mais à partir de l'itération 3 (non réalisée ici), oui.

Classification Hiérarchique

Rappel de la matrice des distances calculée plus haut :

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	25	72	13	50	52	65	5
A_2		0	37	18	25	17	10	20
A_3			0	25	2	4	53	41
A_4				0	13	17	52	2
A_5					0	2	45	25
A_6						0	29	29
A_7							0	58
A_8								0

Question 1 : Classification hiérarchique par lien simple :

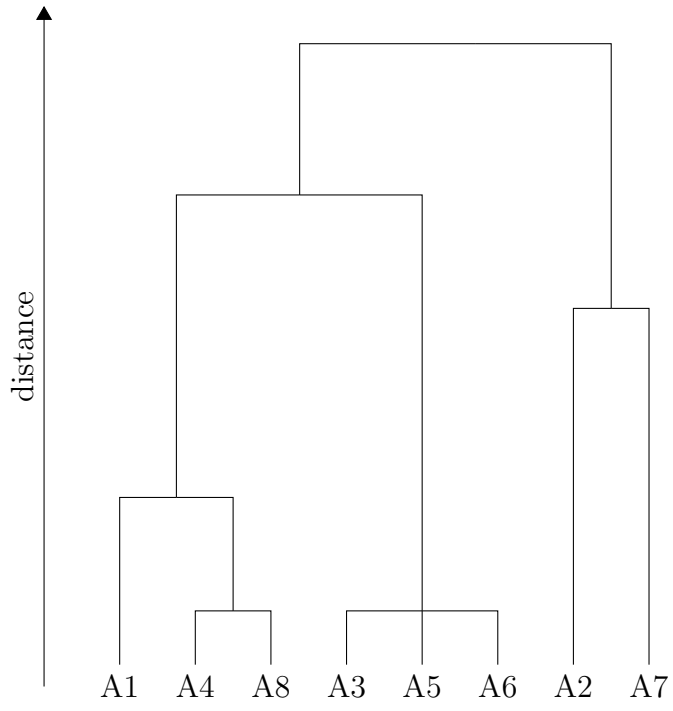
d	liens	groupes	k
d=0		$\{A_1\}\{A_2\}\{A_3\}\{A_4\}\{A_5\}\{A_6\}\{A_7\}\{A_8\}$	k=8
d=1		$\{A_1\}\{A_2\}\{A_3\}\{A_4\}\{A_5\}\{A_6\}\{A_7\}\{A_8\}$	
d=2	$d(A_3, A_5) =$ $d(A_5, A_6) = d(A_4, A_8)$	$\{A_1\}\{A_2\}\{A_3, A_5, A_6\}\{A_4, A_8\}\{A_7\}$	k=5
d=3		$\{A_1\}\{A_2\}\{A_3, A_5, A_6\}\{A_4, A_8\}\{A_7\}$	
d=4		$\{A_1\}\{A_2\}\{A_3, A_5, A_6\}\{A_4, A_8\}\{A_7\}$	
d=5	$d(A_1, A_8)$	$\{A_1, A_4, A_8\}\{A_2\}\{A_3, A_5, A_6\}\{A_7\}$	k=4
...		...	
d=10	$d(A_2, A_7)$	$\{A_1, A_4, A_8\}\{A_2, A_7\}\{A_3, A_5, A_6\}$	k=3
...		...	
d=13	$d(A_4, A_5)$	$\{A_1, A_4, A_8, A_3, A_5, A_6\}\{A_2, A_7\}$	k=2
...		...	
d=17	$d(A_2, A_6)$	$\{A_1, A_4, A_8, A_3, A_5, A_6, A_2, A_7\}$	k=1

Classification hiérarchique par lien complet :

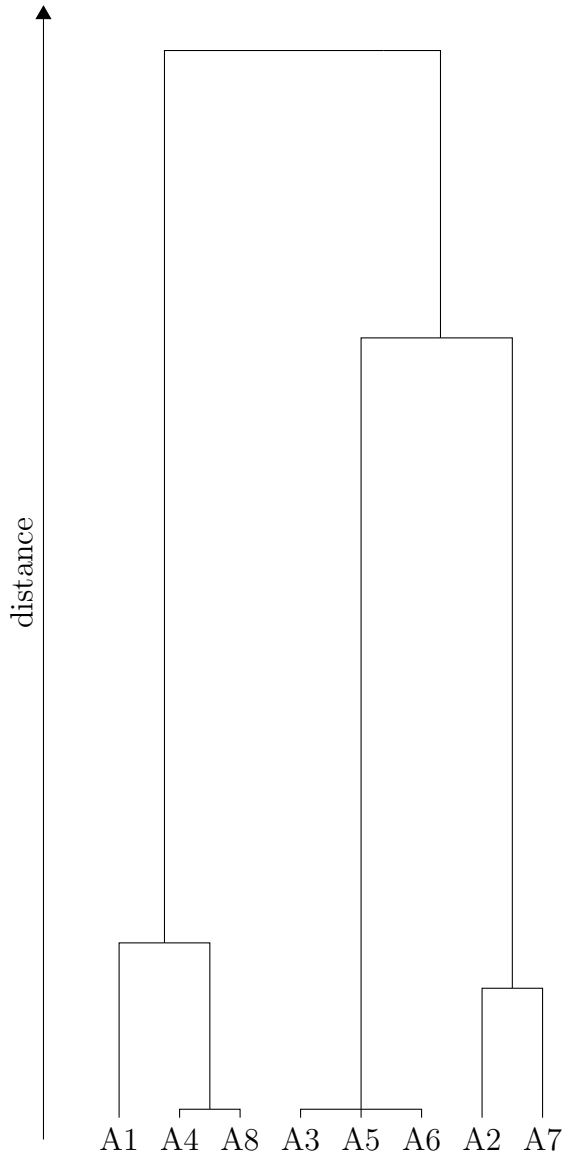
d	liens	groupes	k
d=0		$\{A_1\}\{A_2\}\{A_3\}\{A_4\}\{A_5\}\{A_6\}\{A_7\}\{A_8\}$	k=8
d=1		$\{A_1\}\{A_2\}\{A_3\}\{A_4\}\{A_5\}\{A_6\}\{A_7\}\{A_8\}$	
d=2	$d(A_3, A_5) =$ $d(A_5, A_6) = d(A_4, A_8)$	$\{A_1\}\{A_2\}\{A_3, A_5, A_6\}\{A_4, A_8\}\{A_7\}$	k=5
...		...	
d=10	$d(A_2, A_7)$	$\{A_1\}\{A_2, A_7\}\{A_3, A_5, A_6\}\{A_4, A_8\}$	k=4
...		...	
d=13	$d(A_1, A_4)$	$\{A_1, A_4, A_8\}\{A_2, A_7\}\{A_3, A_5, A_6\}$	k=3
...		...	
d=53	$d(A_3, A_7)$	$\{A_1, A_4, A_8\}\{A_2, A_7, A_3, A_5, A_6\}$	k=2
...		...	
d=72	$d(A_1, A_3)$	$\{A_1, A_4, A_8, A_2, A_7, A_3, A_5, A_6\}$	k=1

Dendrogrammes résultants :

Liens Simples



Liens Complets



Exercice 2 : ACP et Classification

Cet exercice reprend l'exercice 4 du TD d'Algèbre Linéraire, dans lequel on a obtenu :

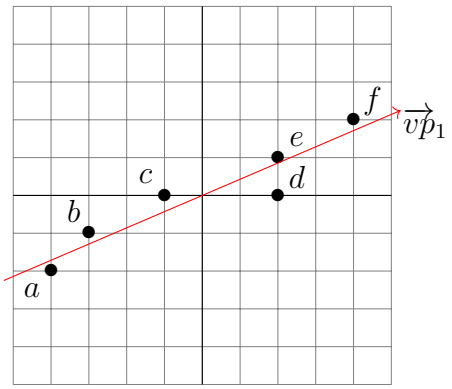
$$X = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} \qquad g = [0, 0] \qquad \Sigma = \begin{bmatrix} 50 & 21 \\ 21 & 10 \end{bmatrix} \qquad X_1 = \begin{bmatrix} 7 \\ 3 \end{bmatrix} \text{ le } \overrightarrow{vp} \text{ de } \Sigma \text{ associé à } \lambda_1 \text{ (1er axe).}$$

On effectue d'abord la projection des points sur le premier axe principal :

$$C_1 = X * X_1 = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} * \begin{bmatrix} 7 \\ 3 \end{bmatrix} = \begin{bmatrix} -34 \\ -24 \\ -7 \\ 14 \\ 17 \\ 34 \end{bmatrix} \begin{matrix} \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f} \end{matrix}$$

On en déduit alors la matrice des distances suivante :

	a	b	c	d	e	f
a	0	10	27	48	51	68
b		0	17	38	41	58
c			0	21	24	41
d				0	3	20
e					0	17
f						0



Enfin, on peut itérer sur les points pour kppv avec un seuil de 20 :

- $\forall x \neq a,\ d(a,b) \leq d(a,x)$
  or, $d(a,b) = 10 \leq 20$
  d'où : $C^1 = \{a, b\}$
- $\forall x \neq b,\ d(b,a) \leq d(b,x)$
  or, $d(b,a) = 10 \leq 20$
  d'où : $C^1 = \{a, b\}$
- $\forall x \neq c,\ d(c,b) \leq d(c,x)$
  or, $d(c,b) = 17 \leq 20$
  d'où : $C^1 = \{a, b, c\}$
- $\forall x \neq d,\ d(d,e) \leq d(d,x)$
  or, $d(d,e) = 5 \leq 20$
  d'où : $C^2 = \{d, e\}$
- $\forall x \neq e,\ d(e,d) \leq d(e,x)$
  or, $d(e,d) = 5 \leq 20$
  d'où : $C^2 = \{d, e\}$
- $\forall x \neq f,\ d(f,e) \leq d(f,x)$
  or, $d(f,e) = 17 \leq 20$
  d'où : $C^2 = \{d, e, f\}$

D'où $C^1 = \{a, b, c\}$ et $C^2 = \{d, e, f\}$