

TD Clustering

ENSTA ParisTech INT-22

Exercice 1 : K-means

Utilisez l'algorithme du k-means et la distance euclidienne pour regrouper les 8 exemples suivants en 3 clusters : A1(2,10), A2(2,5), A3(8,4), A4(5,8), A5(7,5), A6(6,4), A7(1,2), A8(4,9).

La matrice de distance basée sur la distance Euclidienne est fournie ci-dessous :

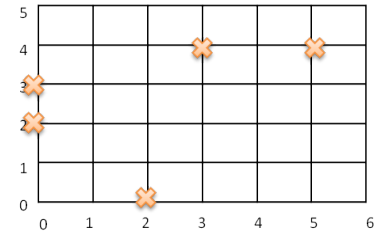
	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

On considère comme centre de classes à l'initialisation les points A1, A4 et A7.

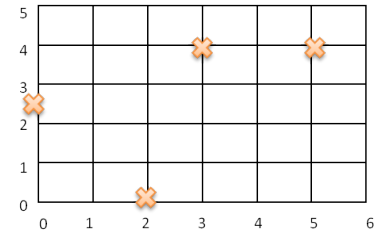
- Déroulez une itération de l'algorithme de k-means pour ces données et cette initialisation et donnez :
 - Les nouveaux clusters
 - Les centres de chaque cluster
 - Faites une représentation graphique montrant les points d'étude, les clusters et les centres des clusters
- Combien d'itérations supplémentaires sont-elles nécessaires pour converger ? Dessinez les résultats de chaque itération nécessaire.

Exercice 2 : Classification hiérarchique

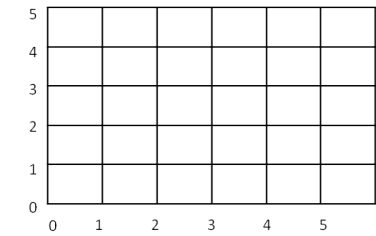
La succession des quatre dessins suivants correspond aux étapes successives d'une classification hiérarchique ascendante des cinq points $M_1(2,0)$, $M_2(0,1)$, $M_3(0,2)$, $M_4(3,4)$ et $M_5(5,4)$ progressivement regroupées en classes de deux ou trois points dont les centres de gravité sont notés G_6 , G_7 , G_8 . On suppose que les cinq points initiaux sont tous affectés du poids 1. La distance choisie pour cette classification, qui apparait dans les quatre matrices de distance, est l'écart de Ward.



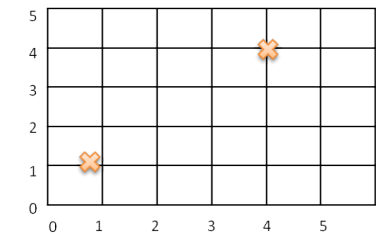
	M1	M2	M3	M4	M5
M1	0.00	2.50	4.00	8.50	12.50
M2	2.50	0.00	0.50	9.00	17.00
M3	4.00	0.50	0.00	6.50	14.50
M4	8.50	9.00	6.50		
M5	12.50			2.00	0.00



	M1	G6	M4	M5
M1	0.00	4.17	8.50	12.50
G6	4.17	0.00		20.83
M4	8.50		0.00	2.00
M5	12.50	20.83	2.00	0.00



	M1	G6	G7
M1	0.00	4.17	13.33
G6	4.17	0.00	22.25
G7	13.33	22.25	0.00



	G8	G7
G8	0.00	24.13
G7	24.13	0.00

- Compléter le troisième dessin en y plaçant les trois points devant y figurer et indiquer sur les quatre dessins le nom des points.
- Compléter les six distances manquantes dans les matrices de distances.
- Précisez les coordonnées des points G_6 , G_7 , G_8
- Calculer les coordonnées du centre de gravité G_9 des cinq points.
- Tracer un dendrogramme résumant cette classification.