Master 1 Informatique – ENSICAEN 2A

Module « Apprentissage »

TD numéro 2

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

	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						0	$\sqrt{29}$	$\sqrt{29}$
x_7							0	$\sqrt{58}$
x_8								0

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

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 10×10.

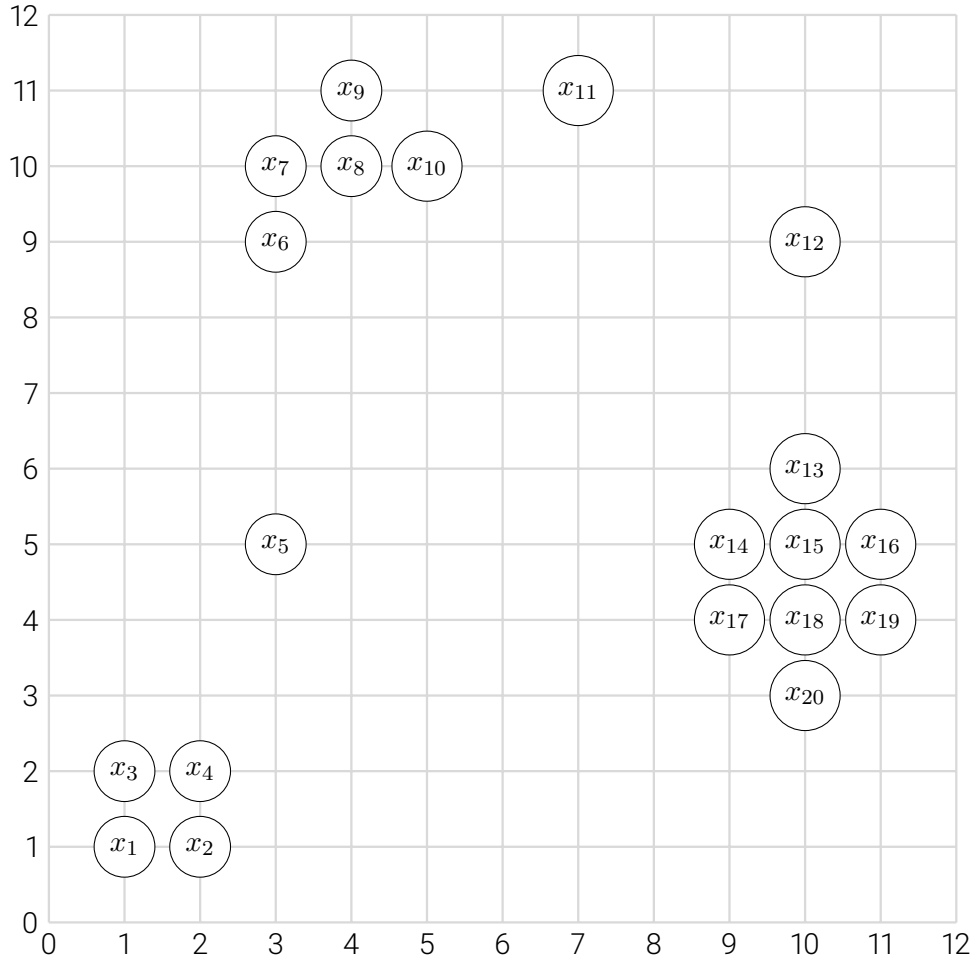
Combien faut-il d'itérations pour que l'algorithme converge ? L'inertie intra-cluster diminue-t-elle ?

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.

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.


12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10 11 12
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
x_{10}
x_{11}
x_{12}
x_{13}
x_{14}
x_{15}
x_{16}
x_{17}
x_{18}
x_{19}
x_{20}