USTHB, Faculté d'Informatique
Département d'Intelligence Artificielle
Et des Sciences de Données
Master S2I, 2ème année

Bab-Ezzouar, le 22 Janvier 2023

Corrigé-type EMD de 'Data mining'

(Les documents ne sont pas autorisés)

Exercice 1. (10 pts)
Considérer les 8 points euclidiens suivants :
A1(2 ;10), A2(8 ;4), A3(5 ;8), A4(7 ;5), A5(6 ;4), A6(1 ;2), A7(4 ;9) et A8(5 ;9).

Matrice des distances (2 pts)

	A1	A2	A3	A4	A5	A6	A7	A8
A1								
A2	8.4							
A3	3.6	5						
A4	7.1	1.4	3.6					
A5	7.2	2	4.1	1.4				
A6	8.1	7.2	7.2	6.7	5.4			
A7	2.2	6.4	1.4	5	5.4	7.6		
A8	3.1	5.8	1	4.5	5.1	8.1	1	

d(A1, A2) = $\sqrt{(2-8)^2+(10-4)^2}$ = 8.4
d(A1, A3) = $\sqrt{(2-5)^2+(10-8)^2}$ = 3.6
d(A1, A4) = $\sqrt{(2-7)^2+(10-5)^2}$ = 7.1
d(A1, A5) = $\sqrt{(2-6)^2+(10-4)^2}$ = 7.2
d(A1, A6) = $\sqrt{(2-1)^2+(10-2)^2}$ = 8.1
d(A1, A7) = $\sqrt{(2-4)^2+(10-9)^2}$ = 2.2
d(A1, A8) = $\sqrt{(2-5)^2+(10-9)^2}$ = 3.1

d(A2, A3) = $\sqrt{(8-5)^2+(4-8)^2}$ = 5
d(A2, A4) = $\sqrt{(8-7)^2+(4-5)^2}$ = 1.4
d(A2, A5) = $\sqrt{(8-6)^2+(4-4)^2}$ = 2
d(A2, A6) = $\sqrt{(8-1)^2+(4-2)^2}$ = 7.2
d(A2, A7) = $\sqrt{(8-4)^2+(4-9)^2}$ = 6.4
d(A2, A8) = $\sqrt{(8-5)^2+(4-9)^2}$ = 5.8

d(A3, A4) = $\sqrt{(5-7)^2+(8-5)^2}$ = 3.6
d(A3, A5) = $\sqrt{(5-6)^2+(8-4)^2}$ = 4.1
d(A3, A6) = $\sqrt{(5-1)^2+(8-2)^2}$ = 7.2
d(A3, A7) = $\sqrt{(5-4)^2+(8-9)^2}$ = 1.4
d(A3, A8) = $\sqrt{(5-5)^2+(8-9)^2}$ = 1

d(A4, A5) = $\sqrt{(7-6)^2+(5-4)^2}$ = 1.4
d(A4, A6) = $\sqrt{(7-1)^2+(5-2)^2}$ = 6.7
d(A4, A7) = $\sqrt{(7-4)^2+(5-9)^2}$ = 5
d(A4, A8) = $\sqrt{(7-5)^2+(5-9)^2}$ = 4.5

d(A5, A6) = $\sqrt{(6-1)^2+(4-2)^2}$ = 5.4
d(A5, A7) = $\sqrt{(6-4)^2+(4-9)^2}$ = 5.4
d(A5, A8) = $\sqrt{(6-5)^2+(4-9)^2}$ = 5.1

d(A6, A7) = $\sqrt{(1-4)^2+(2-9)^2}$ = 7.6
d(A6, A8) = $\sqrt{(1-5)^2+(2-9)^2}$ = 8.1

d(A7, A8) = $\sqrt{(4-5)^2+(9-9)^2}$ = 1

1) En considérant eps=$\sqrt{2}$ et MinPts= 2, déterminer les clusters engendrés par DBSCAN. (3 pts)

Voisinage (A1) = {A1}
Pas de nouveau point donc A1 est un outlier.

Voisinage (A2) = {A2, A4}
Voisinage (A4) = {A4, A2, A5}
Voisinage(A5) = {A4, A5}
Pas de nouveau point donc C1 = {A2, A4, A5}

Voisinage(A3) = {A3, A7, A8}
Voisinage(A7) = {A3, A7, A8}
Voisinage(A8) = {A3, A7, A8}
Pas de nouveau point donc C2 = {A3, A7, A8}

Voisinage(A6) = {A6}
Pas de nouveau point donc A6 est un outlier.

2) Appliquer l'algorithme AGNES et dessiner le dendrogramme. (2 pts)
Initialement, chaque point constitue un cluster. Nous avons donc :
C1 = {A1}, C2 = {A2}, …, C8 = {A8}

La plus petite distance entre les clusters est égale à 1. Elle lie les clusters C3 et C8 d'une part et les clusters C7 et C8 d'autre part.
Fusionnons C3 et C8. Nous obtenons le cluster C9 = {A3, A8}
Utilisons le linkage par centroïde.
Le centroïde de C9 = {(5+5)/2 ; (8+9)/2} = {5 ; 8.5}

Calculons maintenant les distances entre C9 et les autres clusters :
d(C9, C1) = $\sqrt{(5-2)^2+(8.5-10)^2} = \sqrt{3^2+1.5^2} = \sqrt{9+2.25} = \sqrt{11.25}$ = 3.35
d(C9, C2) = $\sqrt{(5-8)^2+(8.5-4)^2} = \sqrt{3^2+4.5^2} = \sqrt{9+20.25} = \sqrt{29.25}$ = 5.40
d(C9, C4) = $\sqrt{(5-7)^2+(8.5-5)^2} = \sqrt{2^2+3.5^2} = \sqrt{4+12.25} = \sqrt{16.25}$ = 4.03
d(C9, C5) = $\sqrt{(5-6)^2+(8.5-4)^2} = \sqrt{1^2+4.5^2} = \sqrt{1+20.25} = \sqrt{21.25}$ = 4.60
d(C9, C6) = $\sqrt{(5-1)^2+(8.5-2)^2} = \sqrt{4^2+6.5^2} = \sqrt{16+42.25} = \sqrt{58.25}$ = 7.63
d(C9, C7) = $\sqrt{(5-4)^2+(8.5-9)^2} = \sqrt{1^2+0.5^2} = \sqrt{1+0.25} = \sqrt{1,25}$ = 1.11
La plus petite distance entre 2 clusters en considérant tous les clusters 2 à 2 de C1, C2, C4, C5, C6,C7 et C9 est égale à 1.11.
On fusionne donc C9 avec C7 et on obtient C10 = {A3, A7, A8}.
Le centroïde de C10 = {(5+4+5)/3 ; (8+9+9)/3}={4.66 ; 8.66}

Calculons maintenant les distances entre C10 et les autres clusters :
d(C10, C1) = $\sqrt{(4.66-2)^2+(8.66-10)^2} = \sqrt{2.66^2+1.34^2} = \sqrt{7.07+1.79} = \sqrt{8.86}$ = 2.97
d(C10, C2) = $\sqrt{(4.66-8)^2+(8.66-4)^2} = \sqrt{3.34^2+4.66^2} = \sqrt{11.15+21.71} = \sqrt{32.86}$ = 5.73
d(C10, C4) = $\sqrt{(4.66-7)^2+(8.66-5)^2} = \sqrt{2.34^2+3.66^2} = \sqrt{5.47+13.39} = \sqrt{18.86}$ = 4.34
d(C10, C5) = $\sqrt{(4.66-6)^2+(8.66-4)^2} = \sqrt{1.34^2+4.66^2} = \sqrt{1.79+21.71} = \sqrt{23.50}$ = 4.84

d(C10, C6) = $\sqrt{(4.66-1)^2+(8.66-2)^2} = \sqrt{3.66^2+6.66^2} = \sqrt{13.39+44.38} = \sqrt{57.77}$ = 7.60
La plus petite distance entre 2 clusters en considérant tous les clusters 2 à 2 de C1, C5, C6 et C10 est égale à 1.4.
On a le choix entre fusionner C4 et C2 ou C4 et C5.
Fusionnons C2 avec C4 et on obtient C11 = {A2, A4,}.
Le centroïde de C11 = {(8+7)/2 ; (4+5)/2}={7.50 ; 4.50}

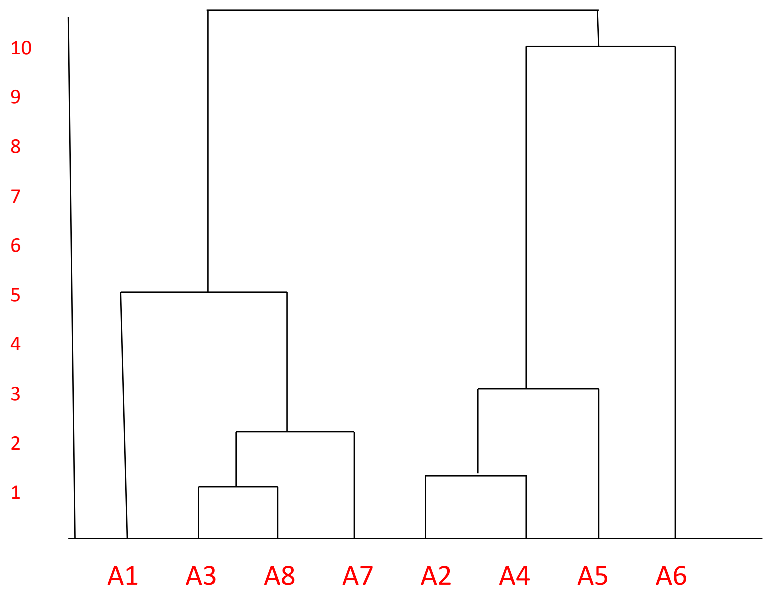
Calculons maintenant les distances entre C11 et les autres clusters :
d(C11, C1) = $\sqrt{(7.50-2)^2+(4.50-10)^2} = \sqrt{5.50^2+5.50^2} = \sqrt{30.25+30.25} = \sqrt{60.50}$ = 7.77
d(C11, C5) = $\sqrt{(7.50-6)^2+(4.50-4)^2} = \sqrt{1.50^2+0.50^2} = \sqrt{2.25+0.25} = \sqrt{2.50}$ = 1.58
d(C11, C6) = $\sqrt{(7.50-1)^2+(4.50-2)^2} = \sqrt{6.5^2+2.5^2} = \sqrt{42.25+6.25} = \sqrt{48.50}$ = 6,96
d(C11,C10) = $\sqrt{(7.50-4.66)^2+(4.50-8.66)^2} = \sqrt{2.84^2+4.16^2} = \sqrt{8.06+17.30} = \sqrt{25.36}$ = 5.03
La plus petite distance entre 2 clusters en considérant tous les clusters 2 à 2 de C1, C5, C6, C10 et C11 est égale à 1.58 qui relie C11 à C5.
Fusionnons C11 avec C5 et on obtient C12 = {A2, A4, A5}.
Le centroïde de C12 = {(8+7+6)/3 ; (4+5+4)/3}={7 ; 4.33}.

Calculons maintenant les distances entre C12 et les autres clusters :
d(C12, C1) = $\sqrt{(7-2)^2+(4.33-10)^2} = \sqrt{5^2+5.64^2} = \sqrt{25+31.80} = \sqrt{60.50}$ = 7.53
d(C12, C6) = $\sqrt{(7-1)^2+(4.33-2)^2} = \sqrt{6.5^2+2.5^2} = \sqrt{42.25+6.25} = \sqrt{48.50}$ = 6,96
La plus petite distance entre 2 clusters en considérant tous les clusters 2 à 2 de C1, C6 et C12 est égale à 6.96 qui relie C12 à C6.
Fusionnons C12 avec C6 et on obtient C13 = {A2, A4, A5, A6}.
Le centroïde de C13 est égal à {(8+7+6+1)/4 ; (4+5+4+2)/4} = {5.50 ; 3.75}
Les clusters calculés jusqu'à maintenant sont :
C10 = {A3, A7, A8} et C13 = {A2, A4, A5, A6}. Il nous reste à calculer les distances suivantes :
d(C10, C1) = $\sqrt{(4.66-2)^2+(8.66-10)^2} = \sqrt{2.66^2+1.34^2} = \sqrt{7.07+1.79} = \sqrt{8.86}$ = 2.97
d(C13, C1) = $\sqrt{(5.50-2)^2+(3.75-10)^2} = \sqrt{3.50^2+6.25^2} = \sqrt{12.25+39.06} = \sqrt{51.31}$ = 7.16
La plus petite distance entre C1 et C10 d'une part et C1 et C13 d'autre part est égal à 2.97. On fusionne alors C1 avec C10 et on obtient C14 = {A1, A3, A7, A8}. Le centroïde de C14 ={(2+5+4+5)/4 ; (10+8+9+9)/4}={4 ; 9}

Les deux clusters englobant tous les points obtenus sont :
C13 = {A2, A4, A5, A6}.
C14 = {A1, A3, A7, A8}.
d(C13, C14) = $\sqrt{(5.50-4)^2+(3.75-9)^2} = \sqrt{1.50^2+5.25^2} = \sqrt{2.25+27{,}56} = \sqrt{29.81}$ = 5.46

Il reste à fusionner ces 2 clusters et on obtient :
C15 = {A1, A2, A3, A4, A5, A6, A7, A8}.

Dendrogramme: (2 pts)



3) Dresser un tableau comparatif des 2 méthodes. (1 pt)

DBSCAN	**AGNES**
Density-based clustering	Hierarchical clustering
Needs 2 parameters : eps and MinPts	Does not need parameters
Spatial clustering	Nested clusters
-	dendogram

<u>Exercice</u> 2. (10 pts)

Considérer les données suivantes de 10 patients dans un hôpital.

Patient	**Age**	**T1**	**T2**	**Risque**
1	Jeune	0	1	Low
2	Jeune	1	1	High
3	Adulte	0	0	Low
4	Senior	1	0	High
5	Senior	0	1	Average
6	Jeune	0	0	Low
7	Adulte	1	0	Average
8	Adulte	1	1	Average
9	Senior	0	0	Low
10	Senior	1	1	High

Pour prédire le risque d'attraper la COVID-19, l'âge du patient ainsi que les tests T1 et T2 sont pris en compte. Le risque est l'attribut classe. L'âge est discrétisé en 3 valeurs (jeune, adulte et senior). T1 et

T2 ont des valeurs booléennes (0 : négatif et 1 : positif). Le risque est évalué selon 3 valeurs (Low : faible, Average : moyen et High : élevé).

1) Quel est le risque du patient X ayant les attributs respectifs suivants : (jeune, 1,0), si l'on applique la méthode 4-NN ? (4 pts)

X = (jeune, 1, 0)

d(X,1) = 1/3
d(X,2) = 2/3
d(X,3) = 1/3
d(X,4) = 2/3
d(X,5) = 0
d(X,6) = 2/3
d(X,7) = 2/3
d(X,8) = 1/3
d(X,9) = 1/3
d(X,10) = 1/3

Les 4 plus proches voisins sont : 2, 4, 6 et 7.
Risque(2) = 'High'
Risque(4) = 'High'
Risque(6) = 'Low'
Risque(7) = 'Average'

Par conséquent X appartient à la classe (risque = 'High').

2) Quel est le risque du même patient si l'on applique la classification bayésienne naïve ? (4 pts)

Classes :
C1 : risque = 'Low'
C2 : risque = 'Average'
C3 : risque = 'High'

X = (jeune, 1,0)

P(C1) = 4/10 = 0.4
P(C2) = 3/10 = 0.3
P(C3) = 3/10 = 0.3

Compute $P(X|C_i)$ for each class
P(âge = 'jeune' | risque = 'Low') = 2/4 = 0.5
P(âge = 'jeune' | risque = 'Average') = 0/3 = 0
P(âge = 'jeune' | risque = 'High') = 1/3 = 0.33

P(T1 = '1' | risque = 'Low') = 0/4 = 0
P(T1 = '1' | risque = 'Average') = 2/3 = 0.66
P(T1 = '1' | risque = 'High') = 3/3 = 1

P(T2 = '0' | risque = 'Low') = 3/4 = 0.75
P(T2 = '0' | risque = 'Average') = 1/3 = 0.33
P(T2 = '0' | risque = 'High') = 1/3 = 0.33

$P(X|C_i)$: P(X|risque = 'Low') = 0.5 * 0 * 0.75 = 0
P(X|risque = 'Average') = 0 * 0.66 * 0.33 = 0
P(X|risque = 'High') = 0.33 * 1 * 0.33 = 0.11

$P(X|C_i)*P(C_i)$: P(X|risque = 'Low') * P(risque = 'Low') = 0 * 0.4 = 0
P(X|risque = 'Average') * P(risque = 'Average') = 0 * 0.3 = 0
P(X|risque = 'High') * P(risque = 'High') = 0.11 * 0.3 = 0.033

Par conséquent, X appartient à la classe (risque = 'High')

3) Dresser un tableau comparatif des 2 méthodes. (2 pts)

4-NN	**Bayésienne Naive**
Based on 'Lazy' classification	Based on statistics
Depends on the value of K	No parameter needed
Simple implementation	Simple implementation
-	Time consuming
Effectiveness depends on k	Effectiveness to be improved