Examen Final (2)

Exercice 2 (6 points)

Soient huit points donnés dans le plan.

a	b	c	d	e	f	g	h
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

a) Appliquer la méthode k-means pour partager ces données en deux groupes. **(4 pts)**

b) Vérifier que la méthode k-means tend à réduire la somme des moindres carrés suivante : **(2 pts)**

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

Où m_i est le centre du cluster C_i et p les points de chaque cluster.

Exercice 3 (08 points)

Le dendrogramme est une représentation graphique d'une classification (clustering) hiérarchique par un arbre.

1- Le dendrogramme d'une classification ascendante (ou descendante) est-il unique ? Si OUI dites Comment ? Si NON dites pourquoi ? **(2 Pts)**

2- Comment déterminer le nombre de classes (clusters) à partir du dendrogramme ? Cette méthode est-elle exacte ou approximative ? Quel est le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité? **(2 Pts)**

3- Considérer la matrice de similarité suivante de cinq documents d1, d2, d3, d4 et d5. Déterminer le dendrogramme résultant de l'application du text clustering hiérarchique ascendant en utilisant le « **lien maximum** ». **(4 Pts)**

	d1	**d2**	**d3**	**d4**	**d5**
d1	0	0.5	0.5	0.6	0.8
d2	0.5	0	0.7	0.6	0.5
d3	0.5	0.7	0	0.6	0.5
d4	0.6	0.6	0.6	0	0.9
d5	0.8	0.5	0.5	0.9	0

BONNE REUSSITE

Dr. Tahar Mehenni

Examen Final (2)
Correction

Exercice 2 (6 points)

Soient huit points donnés dans le plan.

a	b	c	d	e	f	g	h
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

a) Appliquer la méthode k-means pour partager ces données en deux groupes. **(4 pts)**

It0 : choix de deux centres : m1=a=(1,3) et m2=e=(1,2), C1={a }, C2={e }

It1 : calcul des distances et affectation aux clusters C1, C2

$d^2(m1,b)= (1-3)^2+(3-3)^2=4$; $d^2(m2, b)= (1-3)^2+(2-3)^2=5$, b ->C1

$d^2(m1,c)= (1-4)^2+(3-3)^2=9$; $d^2(m2, c)= (1-4)^2+(2-3)^2=10$, c ->C1

$d^2(m1,d)= (1-5)^2+(3-3)^2=16$; $d^2(m2, d)= (1-5)^2+(2-3)^2=17$, d ->C1

$d^2(m1,f)= (1-4)^2+(3-2)^2=10$; $d^2(m2, f)= (1-4)^2+(2-2)^2=9$, f ->C2

$d^2(m1,g)= (1-1)^2+(3-1)^2=4$; $d^2(m2, g)= (1-1)^2+(2-1)^2=1$, g ->C2

$d^2(m1,h)= (1-2)^2+(3-1)^2=5$; $d^2(m2, h)= (1-2)^2+(2-1)^2=2$, h ->C2

C1={a,b,c,d}, C2={e,f,g,h}

Calcul des nouveaux centres: m1=((1+3+4+5)/4),(3+3+3+3)/4)=(3.25,3)

m2=((1+4+1+2)/4),(2+2+1+1)/4)=(2,1.5)

It2 : calcul des distances et affectation aux clusters C1, C2

$d^2(m1,a)= (3.25-1)^2+(3-3)^2=5.06$; $d^2(m2, a)= (2-1)^2+(1.5-3)^2=3.25$, a ->C2

$d^2(m1,b)= (3.25-3)^2+(3-3)^2=0.06$; $d^2(m2, b)= (2-3)^2+(1.5-3)^2=3.25$, b ->C1

$d^2(m1,c)= (3.25-4)^2+(3-3)^2=0.56$; $d^2(m2, c)= (2-4)^2+(1.5-3)^2=6.25$, c ->C1

$d^2(m1,d)= (3.25-5)^2+(3-3)^2=3.06$; $d^2(m2, d)= (2-5)^2+(1.5-3)^2=11.25$, d ->C1

$d^2(m1,e)= (3.25-1)^2+(3-2)^2=6.06$; $d^2(m2, e)= (2-1)^2+(1.5-2)^2=1.25$, e ->C2

$d^2(m1,f)= (3.25-4)^2+(3-2)^2=1.56$; $d^2(m2, f)= (2-4)^2+(1.5-2)^2=4.25$, f ->C1

$d^2(m1,g)= (3.25-1)^2+(3-1)^2=9.06$; $d^2(m2, g)= (2-1)^2+(1.5-1)^2=1.25$, g ->C2

$d^2(m1,h)= (3.25-2)^2+(3-1)^2=4.56$; $d^2(m2, h)= (2-2)^2+(1.5-1)^2=0.25$, h ->C2

C1={b,c,d,f}, C2={a,e,g,h}

Calcul des nouveaux centres: m1=((3+4+5+4)/4),(3+3+3+2)/4)=(4,2.75)

m2=((1+1+1+2)/4),(3+2+1+1)/4)=(1.25,1.75)

It3 : calcul des distances et affectation aux clusters C1, C2

$d^2(m1,a)= (4-1)^2+(2.75-3)^2=9.06$; $d^2(m2, a)= (1.25-1)^2+(1.75-3)^2=2.12$, a ->C2

$d^2(m1,b)= (4-3)^2+(2.75-3)^2=1.06$; $d^2(m2, b)= (1.25-3)^2+(1.75-3)^2=4.62$, b ->C1

$d^2(m1,c)= (4-4)^2+(2.75-3)^2=0.06$; $d^2(m2, c)= (1.25-4)^2+(1.75-3)^2=9.12$, c ->C1

$d^2(m1,d)= (4-5)^2+(2.75-3)^2=1.06$; $d^2(m2, d)= (1.25-5)^2+(1.75-3)^2=15.62$, d ->C1

$d^2(m1,e)= (4-1)^2+(2.75-2)^2=9.56$; $d^2(m2, e)= (1.25-1)^2+(1.75-2)^2=0.12$, e ->C2

$d^2(m1,f)= (4-4)^2+(2.75-2)^2=0.56$; $d^2(m2, f)= (1.25-4)^2+(1.75-2)^2=7.62$, f ->C1

$d^2(m1,g)= (4-1)^2+(2.75-1)^2=12.06$; $d^2(m2, g)= (1.25-1)^2+(1.75-1)^2=0.62$, g ->C2

$d^2(m1,h)= (4-2)^2+(2.75-1)^2=5.06$; $d^2(m2, h)= (1.25-2)^2+(1.75-1)^2=1.12$, h ->C2

C1={b,c,d,f}, C2={a,e,g,h}

Stabilisation des clusters. Fin
Résultats : C1={b,c,d,f}, C2={a,e,g,h}

b) Vérifier que la méthode k-means tend à réduire la somme des moindres carrés suivante : **(2pts)**

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

Où m_i est le centre du cluster C_i et p les points de chaque cluster.

Résultat de l'itération 1 : C1={a,b,c,d}, C2={e,f,g,h} avec m1= (3.25,3), m2=(2,1.5)
E1=(|1-3.25|²+|3-3.25|²+|4-3.25|²+|5-3.25|²+|3-3|²+|3-3|²+|3-3|²+|3-3|²) +
(|1-2|²+|4-2|²+|1-2|²+|2-2|²+|2-1.5|²+|2-1.5|²+|1-1.5|²+|1-1.5|²)
E1 = (5.06+0.06+0.56+3.06+0+0+0+0)+(1+4+1+0+0.25+0.25+0.25+0.25)=15.74

Résultat de l'itération 2 : C1={b,c,d,f}, C2={a,e,g,h} avec m1= (4,2.75), m2=(1.25,1.75)
E2=(|3-4|²+|4-4|²+|5-4|²+|4-4|²+|3-2.75|²+|3-2.75|²+|3-2.75|²+|2-2.75|²) +
(|1-1.25|²+|1-1.25|²+|1-1.25|²+|2-1.25|²+|3-1.75|²+|2-1.75|²+|1-1.75|²+|1-1.75|²)
E2 = (1+0+1+0+0.06+0.06+0.06+0.56)+(0.06+0.06+0.06+0.56+1.56+0.06+0.56+0.56)=6.22

Conclusion : E2<E1 d'où la formule est vérifiée.

Exo3 : (08 points)

1- Le dendrogramme d'un clustering n'est pas unique ; il dépend de la stratégie de regroupement : lien minimum, maximum ou moyen. En plus, si la distance minimale choisie n'est pas unique, le choix aléatoire diversifie le dendrogramme.

2- On détermine le nombre de clusters à partir d'un dendrogramme, en déterminant le nombre de points d'intersection entre la droite *y=d* et le dendrogramme, *d* étant la distance choisie. Cette méthode est exacte. Le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité se fait généralement à l'aide de l'avis d'un expert dans le domaine.

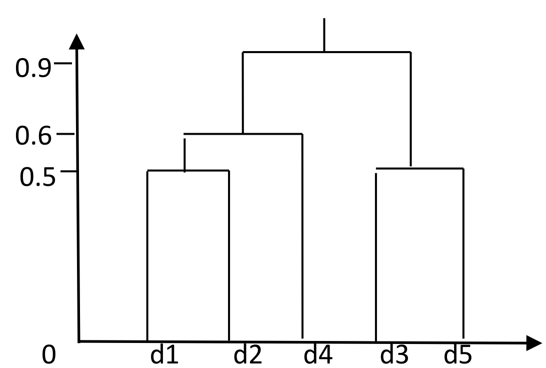
3- Le dendrogramme résultant de l'application du text clustering hiérarchique ascendant de cinq documents d1, d2, d3, d4 et d5, en utilisant le « **lien maximum** ».

	d1	**d2**	**d3**	**d4**	**d5**
d1	0				
d2	**0.5**	0			
d3	0.5	0.7	0		
d4	0.6	0.6	0.6	0	
d5	0.8	0.5	0.5	0.9	0

grouper (d1, d2)

	d1d2	**d3**	**d4**	**d5**
d1d2	0			
d3	0.7	0		
d4	0.6	0.6	0	
d5	0.8	**0.5**	0.9	0

grouper (d3, d5)



Dendrogramme

	d1d2	**d3d5**	**d4**
d1d2	0		
d3d5	0.8	0	
d4	**0.6**	0.9	0

grouper (d1d2, d4)

	d1d2d4	**d3d5**
d1d2d4	0	
d3d5	**0.9**	0

grouper(d1d2d4,d3d5)