USTHB,
FEI, Département Informatique
M2, Master SII

Bab-Ezzouar le 24 mars 2021

Corrigé de l'EMD de Data Mining

Exercice 1 (10 pts)

Considérer les attributs *longueur du sépale en cm* et *longueur de la pétale en cm* du dataset Iris. La table ci-dessous exhibe 12 instances du dataset pour les deux attributs apparaissant respectivement en deuxième et troisième colonnes.

Instance	Sépale	Pétale
1	4.9	1.4
2	5.0	1.4
3	5.4	1.7
4	4.6	1.4
5	5.5	4.0
6	5.1	3.0
7	5.7	4.5
8	5.0	3.3
9	4.9	4.5
10	5.7	5.0
11	5.8	5.1
12	5.6	4.9

1) Dessiner les boites à moustaches pour chacun des attributs Sépale et Pétale. Que pouvez-vous conclure ? **(5 pts)**

Sépale

4.6, 4.9, 4.9, 5.0, 5.0, 5.1, 5.4, 5.5, 5.6, 5.7, 5.7, 5.8

(1 pt)

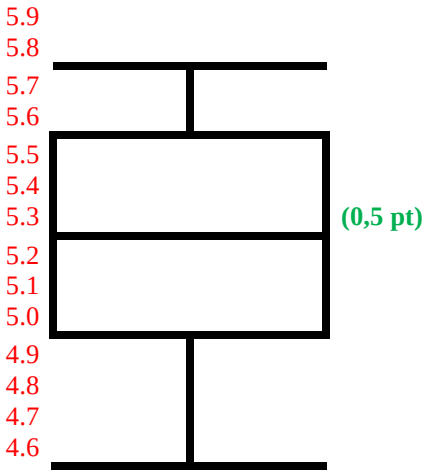
Min = 4.6

Q1 = 4.95

Médiane = 5.25

Q3 = 5.65

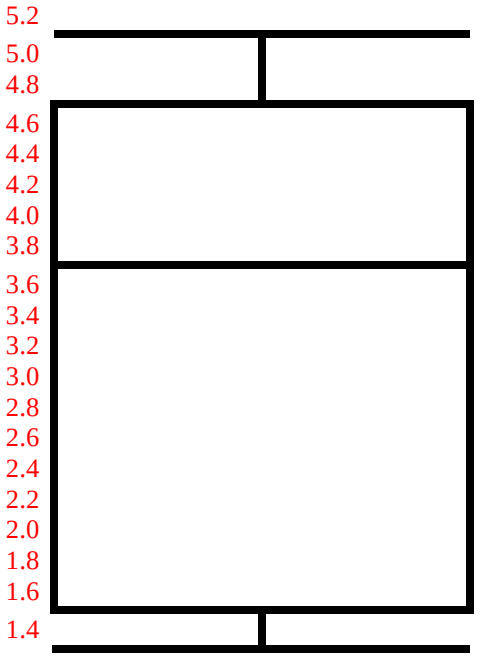
Max = 5.8



(0,5 pt)

Outliers **(1 pt)**
1.5*(Q3-Q1) = 1.5*0.7 = 1.05
Q1 - 1.05 = 3.9 < min
Q3 + 1.05 = 6.7 > max
Donc pas d'outlier
Plus de valeurs supérieures à la médiane

Pétale
1.4, 1.4, 1.4, 1.7, 3.0, 3.3, 4.0, 4.5, 4.5, 4.9, 5.0, 5.1
(1 pt)
Min = 1.4
Q1 = 1.55
Médiane = 3.65
Q3 = 4.7
Max = 5.1



(0,5 pt)

Outliers **(1 pt)**
1.5*(Q3-Q1) = 1.5*3.15 = 4.725
Q1 – 4.725 = -3.175 < min
Q3 + 4.725 = 9.425 > max
Donc pas d'outlier
Plus de valeurs inférieures à la médiane

2) Appliquer l'algorithme *Chimerge* ci-dessous pour discrétiser l'attribut ***Sépale*. (5 pts)**

Algorithme ChiMerge
1. trier les valeurs de l'attribut par ordre croissant.
2. considérer chaque valeur dans un intervalle distinct.
3. calculer la valeur de χ^2 pour tous les intervalles adjacents.
4. fusionner les paires d'intervalles qui ont la plus petite valeur de χ^2.
5. arrêter le processus quand le nombre d'intervalles est égal à 4 sinon aller à (3).

La formule du χ^2 est donnée comme suit:
$\chi^2 = \sum_{i=1}^{i=m} \frac{(R_i - E)^2}{E}$

où:
m est le nombre d'intervalles à comparer (2 dans ce cas),
R_i est le nombre de valeurs de l'intervalle *i*,
E est la fréquence moyenne calculée comme: $E = n/MaxIntervalles$,
n est le nombre total de valeurs,
MaxIntervalles est le nombre maximum d'intervalles.

1) Trier les valeurs de l'attribut Sépale :

4.6
4.9
4.9
5.0
5.0
5.1
5.4
5.5
5.6
5.7
5.7
5.8

Intervalles initiaux :

[4.6 (1)]
[4.9 (2)]
[5.0 (2)]
[5.1 (1)]
[5.4 (1)]
[5.5 (1)]
[5.6 (1)]
[5.7(2)]
[5.8 (1)]

E = 12/4 = 3

1ère itération **(1 pt)**

Intervalle	χ^2
[4.6 (1), 4.9 (2)]	4/3+1/3=5/3
[4.9 (2), 5.0 (2)]	1/3+1/3 =2/3
[5.0 (2), 5.1 (1)]	1/3+4/3 = 5/3
[5.1 (1), 5.4 (1)]	4/3+4/3 = 8/3
[5.4 (1), 5.5 (1)]	4/3+4/3=8/3
[5.5 (1), 5.6 (1)]	8/3
[5.6 (1), 5.7(2)]	5/3
[5.7(2), 5.8 (1)]	5/3

Résultat :

Intervalle
[4.6 (1)]
[4.9 (2), 5.0 (2)]
[5.1 (1)]
[5.4 (1)]
[5.5 (1)]
[5.6 (1)]
[5.7(2)]
[5.8 (1)]

2ème itération **(1 pt)**

Intervalle	χ^2
[[4.6 (1)], [4.9 (2), 5.0 (2)]]	4/3+1/3=5/3
[4.9 (2), 5.0 (2)], [5.1 (1)]	1/3+4/3 =5/3
[5.1 (1)], [5.4 (1)]	4/3+4/3 =8/3
[5.4 (1)], [5.5 (1)]	4/3+4/3=8/3
[5.5 (1)], [5.6 (1)]	8/3
[5.6 (1)], [5.7(2)]	5/3
[5.7(2), 5.8 (1)]	5/3

Résultat :

Intervalle
[4.6 (1), 4.9 (2), 5.0 (2)]
[5.1 (1)]
[5.4 (1)]
[5.5 (1)]
[5.6 (1)]
[5.7(2)]
[5.8 (1)]

3ème itération **(1 pt)**

Intervalle	χ^2
[4.6 (1), 4.9 (2), 5.0 (2)], [5.1 (1)]	4/3+4/3=8/3
[5.1 (1)], [5.4 (1)]	4/3+4/3 =8/3
[5.4 (1)], [5.5 (1)]	4/3+4/3=8/3
[5.5 (1)], [5.6 (1)]	8/3
[5.6 (1)], [5.7(2)]	5/3
[5.7(2)], [5.8 (1)]	5/3

Résultat :

Intervalle
[4.6 (1), 4.9 (2), 5.0 (2)]
[5.1 (1)]
[5.4 (1)]
[5.5 (1)]
[5.6 (1), 5.7(2)]
[5.8(1)]

4ème itération **(1 pt)**

Intervalle	χ^2
[4.6 (1), 4.9 (2), 5.0 (2)], [5.1 (1)]	4/3+4/3=8/3
[5.1 (1)], [5.4 (1)]	4/3+4/3 =8/3
[5.4 (1)], [5.5 (1)]	4/3+4/3=8/3
[5.5 (1)] , [5.6 (1), 5.7(2)]	4/3
[5.6 (1), 5.7(2)], [5.8(1)]	4/3

Résultat :

Intervalle
[4.6 (1), 4.9 (2), 5.0 (2)]
[5.1 (1)]
[5.4 (1)]
[5.5 (1), 5.6 (1), 5.7(2)]
[5.8 (1)]

5ème itération **(1 pt)**

Intervalle	χ^2
[4.6 (1), 4.9 (2), 5.0 (2)], [5.1 (1)]	4/3+4/3=8/3
[5.1 (1)], [5.4 (1)]	4/3+4/3 =8/3
[5.4 (1)], [5.5 (1), 5.6 (1), 5.7(2)]	4/3+1/3=5/3
[5.5 (1), 5.6 (1), 5.7(2)], [5.8 (1)]	5/3

Résultat :

Intervalle
[4.6 (1), 4.9 (2), 5.0 (2)]
[5.1 (1)]
[5.4 (1), 5.5 (1), 5.6 (1), 5.7(2)]
[5.8 (1)]

Exercice2. (10 pts)

Considérer le dataset suivant contenant 10 instances et 5 attributs nommés A, B, C, D et E. On s'intéresse à extraire des motifs fréquents pour déduire des règles d'association entre les attributs. Les instances font office de transactions et les valeurs des attributs d'items.

	A	B	C	D	E
I1	1	4	13	2	3
I2	1	2	12	0	7
I3	1	3	13	2	6
I4	1	4	11	2	7
I5	1	4	14	2	7
I6	0	4	15	2	7
I7	1	1	13	0	3
I8	1	4	14	0	7
I9	1	4	14	2	7
I10	1	4	12	2	7

1) Appliquer l'algorithme A-priori sur le dataset ci-dessus avec un support minimal de 40%. **(5 pts)**

Pour appliquer l'algorithme A-priori sur le dataset, nous devons d'abord coder les valeurs des attributs comme suit :

	A	**B**	**C**	**D**	**E**	**ensemble**
I1	A1	B4	C13	D2	E3	{A1, B4, C13, D2, E3}
I2	A1	B2	C12	D0	E7	{A1, B2, C12, D0, E7}
I3	A1	B3	C13	D2	E6	{A1, B3, C13, D2, E6}
I4	A1	B4	C11	D2	E7	{A1, B4, C11, D2, E7}
I5	A1	B4	C14	D2	E7	{A1, B4, C14, D2, E7}
I6	A0	B4	C15	D2	E7	{A0, B4, C15, D2, E7}
I7	A1	B1	C13	D0	E3	{A1, B1, C13, D0, E3}
I8	A1	B4	C14	D0	E7	{A1, B4, C14, D0, E7}
I9	A1	B4	C14	D2	E7	{A1, B4, C14, D2, E7}
I10	A1	B4	C12	D2	E7	{A1, B4, C12, D2, E7}

La dernière colonne contient les ensembles des éléments de chaque instance.

Première itération : (1,5 pt)

Détermination des candidats C1 : parcours des transactions et comptage des occurrences de chaque item. Ce qui donne : C1

itemset	**support**
{A0}	1
{A1}	9
{B1}	1
{B2}	1
{B3}	1
{B4}	7
{C11}	1
{C12}	2
{C13}	3
{C14}	3
{C15}	1
{D0}	3
{D2}	7
{E3}	2
{E6}	1
{E7}	7

Le support minimum étant égal à 40%, il équivaut à 10 x 40% = 4 transactions. Les motifs fréquents d'ordre 1, appartenant à L1 sont des candidats C1 qui satisfont le support minimum.

C1

itemset	support
{A0}	1
{A1}	9
{B1}	1
{B2}	1
{B3}	1
{B4}	7
{C11}	1
{C12}	2
{C13}	3
{C14}	3
{C15}	1
{D0}	3
{D2}	7
{E3}	2
{E6}	1
{E7}	7

L1

itemset	support
{A1}	9
{B4}	7
{D2}	7
{E7}	7

Deuxième itération : (1,5 pt)

Détermination de C2, des itemsets candidats contenant au plus deux items :

C2

itemset	support
{A1, B4}	6
{A1, D2}	6
{A1, E7}	6
{B4, D2}	6
{B4, E7}	6
{D2, E7}	5

Les itemsets retenus sont donc :

L2

itemset	support
{A1, B4}	6
{A1, D2}	6
{A1, E7}	6
{B4, D2}	6
{B4, E7}	6
{D2, E7}	5

Les items d'un même attribut ne peuvent pas apparaître dans un itemset car ils représentent le même attribut avec des valeurs différentes.

Troisième itération : (1,5 pt)

C3

itemset	support
{A1, B4, D2}	5
{A1, B4, E7}	5
{A1, B4, D2, E7}	4
{A1, D2, E7}	4
{B4, D2, E7}	5

L3

itemset	support
{A1, B4, D2}	5
{A1, B4, E7}	5
{A1, B4, D2, E7}	4
{A1, D2, E7}	4
{B4, D2, E7}	5

Quatrième itération : (0,5 pt)

C4 = ∅, l'ensemble vide. Le processus s'arrête. L'ensemble des motifs fréquents est donc comme suit :
$L = L1 \cup L2 \cup L3$.

2) Appliquer l'algorithme k-means sur les 6 premières instances du dataset pour k = 2 et en démarrant avec les instances I2 et I4 comme centroides initiaux. Considérer tous les types des attributs comme des entiers. **(5 pts)**

Initialisations :

C1 = {I2} C2 = {I4}

Première itération :

Calcul des distances entre les instances et I2 et I4 :

I2	1	2	12	0	7
I4	1	4	11	2	7
I1	1	4	13	2	3

Distance (I1, I2) = $|1-1| + |4-2| + |13-12| + |2-0| + |3-7| = 0 + 2 + 1 + 2 + 4 = 9$
Distance (I1, I4) = $|1-1| + |4-4| + |13-11| + |2-2| + |3-7| = 0 + 0 + 2 + 0 + 4 = 6$
C2 = {I4, I1} **(0,5 pt)**

I3	1	3	13	2	6

Distance (I3, I2) = $|1-1| + |3-2| + |13-12| + |2-0| + |6-7| = 0 + 1 + 1 + 2 + 1 = 5$
Distance (I3, I4) = $|1-1| + |3-4| + |13-13| + |2-2| + |6-3| = 0 + 1 + 0 + 0 + 3 = 4$
C2 = {I4, I1, I3} **(0,5 pt)**

I5	1	4	14	2	7

Distance (I5, I2) = $|1-1| + |4-2| + |14-12| + |2-0| + |7-7| = 0 + 2 + 2 + 2 + 0 = 6$
Distance (I5, I4) = $|1-1| + |4-4| + |14-11| + |2-2| + |7-7| = 0 + 0 + 3 + 0 + 0 = 3$
C2 = {I4, I1, I3, I5} **(0,5 pt)**

I6	0	4	15	2	7

Distance (I6, I2) = $|0-1| + |4-2| + |15-12| + |2-0| + |7-7| = 1 + 2 + 3 + 2 + 0 = 8$
Distance (I6, I4) = $|0-1| + |4-4| + |15-13| + |2-2| + |7-3| = 0 + 3 + 0 + 2 + 0 = 5$
C2 = {I4, I1, I3, I5, I6} **(0,5 pt)**

Mise à jour des centroides:

I2	1	2	12	0	7
c1	1	2	12	0	7

I1	1	4	13	2	3
I3	1	3	13	2	6
I4	1	4	11	2	7
I5	1	4	14	2	7
I6	0	4	15	2	7
c2	0.8	3.8	13.2	2	6

Deuxième itération :

Calcul des distances entre les instances et c1 et c2 :

c1	1	2	12	0	7

c2	0.8	3.8	13.2	2	6

I1	1	4	13	2	3

Distance (I1, c1) = $|1-1|+|4-2|+|13-12|+|2-0|+|3-7| = 0+2+1+2+4 = 9$
Distance (I1, c2) = $|1-0.8|+|4-3.8|+|13-13.2|+|2-2|+|3-6| = 0.2+0.2+0.2+0+3 = 3.6$
C2 = {I1} **(0,5 pt)**

I2	1	2	12	0	7

Distance (I2, c1) = $|1-1|+|2-2|+|12-12|+|0-0|+|7-7| = 0$
Distance (I2, c2) = $|1-0.8|+|2-3.8|+|12-13.2|+|0-2|+|7-6| = 0.2+1.8+1.2+2+1 = 6.2$
C1 = {I2} **(0,5 pt)**

I3	1	3	13	2	6

Distance (I3, c1) = $|1-1|+|3-2|+|13-12|+|2-0|+|6-7| = 0+1+1+2+1 = 5$
Distance (I3, c2) = $|1-0.8|+|3-3.8|+|13-13.2|+|2-2|+|6-6| = 0.2+0.8+0.2+0+0 = 1.2$
C2 = {I1, I3} **(0,5 pt)**

I4	1	4	11	2	7

Distance (I4, c1) = $|1-1|+|4-2|+|11-12|+|2-0|+|7-7| = 0+2+1+2+0 = 5$
Distance (I4, c2) = $|1-0.8|+|4-3.8|+|11-13.2|+|2-2|+|7-6| = 0.2+0.2+2.2+0+1 = 3.6$
C2 = {I1, I3, I4} **(0,5 pt)**

I5	1	4	14	2	7

Distance (I5, c1) = $|1-1|+|4-2|+|14-12|+|2-0|+|7-7| = 0+2+2+2+0 = 6$
Distance (I5, c2) = $|1-0.8|+|4-3.8|+|14-13.2|+|2-2|+|7-6| = 0.2+0.2+0.8+0+1 = 2.2$
C2 = {I1, I3, I4, I5} **(0,5 pt)**

I6	0	4	15	2	7

Distance (I6, c1) = $|0-1|+|4-2|+|15-12|+|2-0|+|7-7| = 1+2+3+2+0 = 8$
Distance (I6, c2) = $|0-0.8|+|4-3.8|+|15-13.2|+|2-2|+|7-6| = 0.8+0.2+1.8+0+1 = 3.8$
C2 = {I1, I3, I4, I5, I6} **(0,5 pt)**

Le processus s'arrête car les contenus des clusters ne changent pas. Le résultat est donc :
C1 = {I2}
C2 = {I1, I3, I4, I5, I6}