USTHB,
FEI, Département Informatique
M2, Master SII

Bab-Ezzouar le 7 janvier 2013

EMD de Data Mining

Exercice1. (12 pts)

Une base de données a cinq transactions. Supposer que *min sup* = 60% et *min conf* = 80%.

TID	Articles achetés
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

(1) Déterminer tous les itemsets fréquents en appliquant :
   a. l'algorithme Apriori
   b. puis l'algorithme FP-growth.
   c. Comparer l'efficacité des deux méthodes.

(2) Lister toutes les règles d'association (avec un support *s* et une confiance *c*) correspondant à la métarègle suivante, où *X* est une variable représentant des clients, et *itemi* dénote les variables représentant des items:
$\forall X \in transaction, achète(X, item1) \wedge achète(X, item2) \Rightarrow achète(X, item3)$ [s, c]

Exercice2. (8pts)

Supposer que le processus de datamining est de clustériser les huit points suivants dans trois clusters:
*A*1(2, 10), *A*2(2, 5), *A*3(8, 4), *B*1(5, 8), *B*2(7, 5), *B*3(6, 4), *C*1(1, 2), *C*2(4, 9):
avec (*x*, *y*) représentant les coordonnées du point et la fonction de distance est la distance euclidienne. Supposer qu'initialement les points *A*1, *B*1, et *C*1 sont les centres des trois clusters respectivement.
Utiliser l'algorithme *k-means* pour déterminer

(1) les centres des trois clusters après l'exécution de chaque itération.
(2) les trois clusters finaux.

Rédiger les deux exercices séparément.

BON COURAGE !

EMD de Data Mining

Exercice1. (12 pts)

Une base de données a cinq transactions. Supposer que *min sup* = 60% et *min conf* = 80%.

TID	Articles achetés
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

(1) Déterminer tous les itemsets fréquents en appliquant :

a. l'algorithme Apriori

support = 60% x 5 transactions = 3 transactions **(0,5 pt)**

C1 **(1pt)**

Itemset	support
{M}	3
{O}	3
[redacted]	
{K}	5
{E}	4
{Y}	3
[redacted]	

N → 2

D → 1
A → 1
U → 1
C → 2
I → 1

L1 **(0,5 pt)**

Itemset	support
{M}	3
{O}	3
{K}	5
{E}	4
{Y}	3

C2 **(1 pt)**

Itemset	support
[redacted]	
{M,K}	3
[redacted]	
{O,K}	3
{O,E}	3
[redacted]	
{K,E}	4
{K,Y}	3
[redacted]	

mo → 1

mE → 2
mY → 2

oy → 2

EY → 2

L2 **(0,5 pt)**

Itemset	support
{M,K}	3
{O,K}	3
{O,E}	3
{K,E}	4
{K,Y}	3

C3 **(1 pt)**

Itemset	support
{O,K,E}	3

L3

Itemset	support
{O,K,E}	3

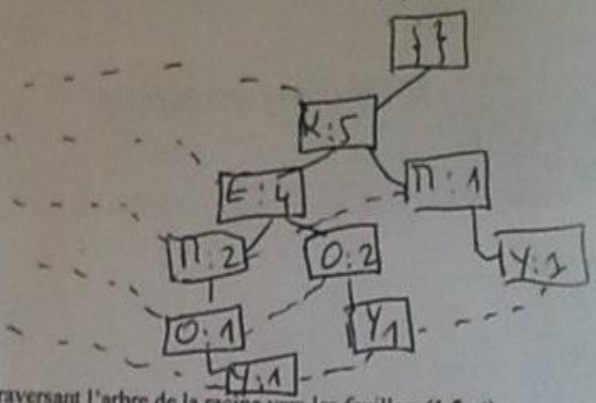
Les itemsets fréquents sont renfermés dans l'ensemble L1UL2UL3 **(0,5 pt)**

b. puis l'algorithme FP-growth.

F-list = K-E-M-O-Y **(1 pt)**

Item	fréquence
K	5
E	4
M	3
O	3
Y	3

FP-Tree **(2 pts)**



Les itemsets fréquents sont obtenus en traversant l'arbre de la racine vers les feuilles: **(1,5 pt)**

- cardinalité égale à 1 : {K}, {E}, {M}, {O} et {Y} (* tous ceux qui existent après la racine avec un support=3 *)
- cardinalité égale à 2 : {K, E}, {K, M}, {K, O}, {K, Y} et {E, O} (* tous ceux qui existent après la racine et le préfixe avec un support=3 *)
- cardinalité égale à 3 : {K, E, O} (* tous ceux qui existent après la racine et les préfixes avec un support=3 *)

c. Comparer l'efficacité des deux méthodes.

FP-growth est plus efficace car, la base de données est consultée une seule fois lors de la construction de FP-tree alors que pour Apriori, la base de données est parcourue à chaque itération de l'algorithme. **(1 pt)**

(2) Lister toutes les règles d'association (avec un support *s* et une confiance *c*) correspondant à la métarègle suivante, où *X* est une variable représentant des clients, et *itemi* dénote les variables représentant des items:
$\forall X \in$ *transaction*, *achète*(*X*, *item*1) ^ *achète*(*X*, *item*2) $\Rightarrow$ *achète*(*X*, *item*3) [*s*, *c*]

achète(*X*, *O*) ^ *achète*(*X*, *K*) $\Rightarrow$ *achète*(*X*, *E*) [60%, 100%] **(0,5 pt)**
achète(*X*, *O*) ^ *achète*(*X*, *E*) $\Rightarrow$ *achète*(*X*, *K*) [60%, 100%] **(0,5 pt)**
achète(*X*, *K*) ^ *achète*(*X*, *E*) $\Rightarrow$ *achète*(*X*, *O*) [60%, 75%] **(0,5 pt) exclue car confiance < 80%**

Exercice2. (8 pts)

Supposer que le processus de datamining est de clustériser les huit points suivants dans trois clusters: $A1(2, 10)$, $A2(2, 5)$, $A3(8, 4)$, $B1(5, 8)$, $B2(7, 5)$, $B3(6, 4)$, $C1(1, 2)$, $C2(4, 9)$.
avec (x, y) représentant les coordonnées du point et la fonction de distance est la distance euclidienne. Supposer qu'initialement les points $A1$, $B1$, et $C1$ sont les centres des clusters respectivement.
Utiliser l'algorithme *k-means* pour déterminer

(1) les centres des trois clusters après l'exécution de chaque itération

Première itération (1,5 pts)

cluster1 = {A1}, cluster2 = {B1}, cluster3 = {C1}

$d(A2,A1) = \sqrt{(2-2)^2 + (5-10)^2} = 5$
$d(A2,B1) = \sqrt{(2-5)^2 + (5-8)^2} = \sqrt{9+9} = \sqrt{18} = 4,24$
$d(A2,C1) = \sqrt{(2-1)^2 + (5-2)^2} = \sqrt{1+9} = \sqrt{10} = 3,16$
cluster3 = {A2, C1}

$d(A3,A1) = \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{36+36} = \sqrt{72} = 8,48$
$d(A3,B1) = \sqrt{(8-5)^2 + (4-8)^2} = \sqrt{9+16} = \sqrt{25} = 5$
$d(A3,C1) = \sqrt{(8-1)^2 + (4-2)^2} = \sqrt{49+4} = \sqrt{53} = 7,28$
cluster2 = {A3, B1}

$d(B2,A1) = \sqrt{(7-2)^2 + (5-10)^2} = \sqrt{25+25} = \sqrt{50} = 7,07$
$d(B2,B1) = \sqrt{(7-5)^2 + (5-8)^2} = \sqrt{4+9} = \sqrt{13} = 3,60$
$d(B2,C1) = \sqrt{(7-1)^2 + (5-2)^2} = \sqrt{36+9} = \sqrt{45} = 6,70$
cluster2 = {B2, A3, B1}

$d(B3,A1) = \sqrt{(6-2)^2 + (4-10)^2} = \sqrt{16+36} = \sqrt{52} = 7,21$
$d(B3,B1) = \sqrt{(6-5)^2 + (4-8)^2} = \sqrt{1+16} = \sqrt{17} = 4,12$
$d(B3,C1) = \sqrt{(6-1)^2 + (4-2)^2} = \sqrt{25+4} = \sqrt{29} = 5,38$
cluster2 = {B3, B2, A3, B1}

$d(C2,A1) = \sqrt{(4-2)^2 + (9-10)^2} = \sqrt{4+1} = \sqrt{5} = 2,23$
$d(C2,B1) = \sqrt{(4-5)^2 + (9-8)^2} = \sqrt{1+1} = \sqrt{2} = 1,41$
$d(C2,C1) = \sqrt{(4-1)^2 + (9-2)^2} = \sqrt{9+49} = \sqrt{58} = 7,61$
cluster2 = {C2, B3, B2, A3, B1}

cluster1 = {A1}, cluster2 = {C2, B3, B2, A3, B1}, cluster3 = {A2, C1}

cluster	Centre de gravité
cluster1 = {A1}	A1**(2, 10)=G1**
cluster2 = {C2, B3, B2, A3, B1}	((4+6+7+8+5)/5, (9+4+5+4+8)/5)= **(6,6)=G2**
cluster3 = {A2, C1}	((2+1)/2 (5+2)/2)= **(3/2,7/2)=G3**

Deuxième iteration (2 pts)

$d(A2,G1) = \sqrt{(2-2)^2 + (5-10)^2} = 5$

$d(A2,G2) = \sqrt{(2-6)^2 + (5-6)^2} = \sqrt{16+1} = \sqrt{17} = 4,12$
$d(A2,G3) = \sqrt{(2-1,5)^2 + (5-3,5)^2} = \sqrt{0,25+2,25} = \sqrt{2,5} = 1,58$
cluster3 = {A2}

$d(A3,G1) = \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{36+36} = \sqrt{72} = 8,48$
$d(A3,G2) = \sqrt{(8-6)^2 + (4-6)^2} = \sqrt{4+4} = \sqrt{8} = 2,82$
$d(A3,G3) = \sqrt{(8-1,5)^2 + (4-3,5)^2} = \sqrt{30,25+0,25} = \sqrt{30,50} = 5,52$
cluster2 = {A3, G2}

$d(B1,G1) = \sqrt{(5-2)^2 + (8-10)^2} = \sqrt{9+4} = \sqrt{13} = 3,60$
$d(B1,G2) = \sqrt{(5-6)^2 + (8-6)^2} = \sqrt{1+4} = \sqrt{5} = 2,23$
$d(B1,G3) = \sqrt{(5-1,5)^2 + (8-3,5)^2} = \sqrt{12,25+20,25} = \sqrt{32,5} = 5,70$
cluster2 = {B1, A3}

$d(B2,G1) = \sqrt{(7-2)^2 + (5-10)^2} = \sqrt{25+25} = \sqrt{50} = 7,07$
$d(B2,G2) = \sqrt{(7-6)^2 + (5-6)^2} = \sqrt{1+1} = \sqrt{2} = 1,41$
$d(B2,G3) = \sqrt{(7-1,5)^2 + (5-3,5)^2} = \sqrt{30,25+2,25} = \sqrt{32,5} = 5,70$
cluster2 = {B2, B1, A3}

$d(B3,G1) = \sqrt{(6-2)^2 + (4-10)^2} = \sqrt{16+36} = \sqrt{52} = 7,21$
$d(B3,G2) = \sqrt{(6-6)^2 + (4-6)^2} = \sqrt{0+4} = \sqrt{4} = 2$
$d(B3,G3) = \sqrt{(6-1,5)^2 + (4-3,5)^2} = \sqrt{20,25+0,25} = \sqrt{20,5} = 4,52$
cluster2 = {B3, B2, B1, A3}

$d(C1,G1) = \sqrt{(1-2)^2 + (2-10)^2} = \sqrt{1+64} = \sqrt{65} = 8,06$
$d(C1,G2) = \sqrt{(1-6)^2 + (2-6)^2} = \sqrt{25+16} = \sqrt{41} = 6,40$
$d(C1,G3) = \sqrt{(1-1,5)^2 + (2-3,5)^2} = \sqrt{0,25+2,25} = \sqrt{2,5} = 1,58$
cluster3 = {C1, A2}

$d(C2,G1) = \sqrt{(4-2)^2 + (9-10)^2} = \sqrt{4+1} = \sqrt{5} = 2,23$
$d(C2,G2) = \sqrt{(4-6)^2 + (9-6)^2} = \sqrt{4+9} = \sqrt{13} = 3,60$
$d(C2,G3) = \sqrt{(4-1,5)^2 + (9-3,5)^2} = \sqrt{6,25+30,25} = \sqrt{36,5} = 6,04$
cluster1 = {C2}

cluster	Centre de gravité
cluster1 = {C2, A1}	((2+4)/2, (10+9)/2)= **(3,19/2)=G1**
cluster2 = {B3, B2, B1, A3}	((6+7+8+5)/4, (4+5+4+8)/4)= **(13/2,21/4)=G2**
cluster3 = {C1, A2}	((2+1)/2 (5+2)/2)= **(3/2,7/2)=G3**

Troisième iteration (2 pts)

$d(A1,G1) = \sqrt{(2-3)^2 + (10-9,5)^2} = \sqrt{1+0,25} = \sqrt{1,25} = 1,11$
$d(A1,G2) = \sqrt{(2-6,5)^2 + (10-5,25)^2} = \sqrt{20,25+22,56} = 6,54$
$d(A1,G3) = \sqrt{(2-1,5)^2 + (10-3,5)^2} = \sqrt{0,25+42,25} = \sqrt{42,5} = 6,51$
cluster1 = {A1}

$d(A2,G1) = \sqrt{(2-3)^2 + (5-9,5)^2} = \sqrt{1+20,25} = \sqrt{21,25} = 4,60$

$d(A2,G2) = \sqrt{(2-6,5)^2 + (5-5,25)^2} = \sqrt{20,25 + 0,062} = \sqrt{20,31} = 4,50$
$d(A2,G3) = \sqrt{(2-1,5)^2 + (5-3,5)^2} = \sqrt{0,25 + 2,25} = \sqrt{2,5} = 1,58$
cluster3 = {A2}

$d(A3,G1) = \sqrt{(8-3)^2 + (4-9,5)^2} = \sqrt{25 + 30,25} = \sqrt{55,25} = 7,43$
$d(A3,G2) = \sqrt{(8-6,5)^2 + (4-5,25)^2} = \sqrt{2,25 + 1,56} = \sqrt{3,81} = 1,95$
$d(A3,G3) = \sqrt{(8-1,5)^2 + (4-3,5)^2} = \sqrt{30,25 + 0,25} = \sqrt{30,50} = 5,52$
cluster2 = {A3}

$d(B1,G1) = \sqrt{(5-3)^2 + (8-9,5)^2} = \sqrt{4 + 2,25} = \sqrt{6,25} = 2,5$
$d(B1,G2) = \sqrt{(5-6,5)^2 + (8-5,25)^2} = \sqrt{2,25 + 7,56} = \sqrt{9,81} = 3,13$
$d(B1,G3) = \sqrt{(5-1,5)^2 + (8-3,5)^2} = \sqrt{12,25 + 20,25} = \sqrt{32,5} = 5,70$
cluster1 = {B1, A1}

$d(B2,G1) = \sqrt{(7-3)^2 + (5-9,5)^2} = \sqrt{16 + 20,25} = \sqrt{36,25} = 6,02$
$d(B2,G2) = \sqrt{(7-6,5)^2 + (5-5,25)^2} = \sqrt{0,25 + 0,06} = \sqrt{0,31} = 0,55$
$d(B2,G3) = \sqrt{(7-1,5)^2 + (5-3,5)^2} = \sqrt{30,25 + 2,25} = \sqrt{32,5} = 5,70$
cluster2 = {B2, A3}

$d(B3,G1) = \sqrt{(6-3)^2 + (4-9,5)^2} = \sqrt{9 + 30,25} = \sqrt{39,25} = 6,26$
$d(B3,G2) = \sqrt{(6-6,5)^2 + (4-5,25)^2} = \sqrt{0,25 + 1,56} = \sqrt{1,81} = 1,34$
$d(B3,G3) = \sqrt{(6-1,5)^2 + (4-3,5)^2} = \sqrt{20,25 + 0,25} = \sqrt{20,5} = 4,52$
cluster2 = {B3, B2, A3}

$d(C1,G1) = \sqrt{(1-3)^2 + (2-9,5)^2} = \sqrt{4 + 56,25} = \sqrt{60,25} = 7,76$
$d(C1,G2) = \sqrt{(1-6,5)^2 + (2-5,25)^2} = \sqrt{30,25 + 10,56} = \sqrt{40,81} = 6,38$
$d(C1,G3) = \sqrt{(1-1,5)^2 + (2-3,5)^2} = \sqrt{0,25 + 2,25} = \sqrt{2,5} = 1,58$
cluster3 = {C1, A2}

$d(C2,G1) = \sqrt{(4-3)^2 + (9-9,5)^2} = \sqrt{1 + 0,25} = \sqrt{1,25} = 1,11$
$d(C2,G2) = \sqrt{(4-6,5)^2 + (9-5,25)^2} = \sqrt{6,25 + 14,06} = \sqrt{20,31} = 4,50$
$d(C2,G3) = \sqrt{(4-1,5)^2 + (9-3,5)^2} = \sqrt{6,25 + 30,25} = \sqrt{36,5} = 6,04$
cluster1 = {C2, B1, A1}

cluster	Centre de gravité
cluster1 = {C2, B1, A1}	((2+5+4)/3, (10+8+9)/3)= **(11/3,9) = G1**
cluster2 = {B3, B2, A3}	((6+7+8)/3, (4+5+4)/3)= **(7,13/3) = G2**
cluster3 = {C1, A2}	((2+1)/2 (5+2)/2)= **(3/2,7/2) = G3**

Quatrième iteration (2 pts)

$d(A1,G1) = \sqrt{(2-3,66)^2 + (10-9)^2} = \sqrt{2,75 + 1} = \sqrt{3,755} = 1,93$
$d(A1,G2) = \sqrt{(2-7)^2 + (10-4,33)^2} = \sqrt{25 + 32,14} = 7,55$
$d(A1,G3) = \sqrt{(2-1,5)^2 + (10-3,5)^2} = \sqrt{0,25 + 42,25} = \sqrt{42,5} = 6,51$
cluster1 = {A1}

$d(A2,G1) = \sqrt{(2-3,66)^2 + (5-9)^2} = \sqrt{2,75+16} = \sqrt{18,75} = 4,33$
$d(A2,G2) = \sqrt{(2-7)^2 + (5-4,33)^2} = \sqrt{25+0,44} = \sqrt{25,44} = 5,04$
$d(A2,G3) = \sqrt{(2-1,5)^2 + (5-3,5)^2} = \sqrt{0,25+2,25} = \sqrt{2,5} = 1,58$
cluster3 = {A2}

$d(A3,G1) = \sqrt{(8-3,66)^2 + (4-9)^2} = \sqrt{18,83+25} = \sqrt{43,83} = 6,62$
$d(A3,G2) = \sqrt{(8-7)^2 + (4-4,33)^2} = \sqrt{1+0,10} = \sqrt{1,10} = 1,04$
$d(A3,G3) = \sqrt{(8-1,5)^2 + (4-3,5)^2} = \sqrt{30,25+0,25} = \sqrt{30,50} = 5,52$
cluster2 = {A3}

$d(B1,G1) = \sqrt{(5-3,66)^2 + (8-9)^2} = \sqrt{4+2,25} = \sqrt{6,25} = 2,5$
$d(B1,G2) = \sqrt{(5-7)^2 + (8-4,33)^2} = \sqrt{4+13,46} = \sqrt{17,46} = 4,17$
$d(B1,G3) = \sqrt{(5-1,5)^2 + (8-3,5)^2} = \sqrt{12,25+20,25} = \sqrt{32,5} = 5,70$
cluster1 = {B1, A1}

$d(B2,G1) = \sqrt{(7-3,66)^2 + (5-9)^2} = \sqrt{16+20,25} = \sqrt{36,25} = 6,02$
$d(B2,G2) = \sqrt{(7-7)^2 + (5-4,33)^2} = \sqrt{0+0,44} = \sqrt{0,44} = 0,67$
$d(B2,G3) = \sqrt{(7-1,5)^2 + (5-3,5)^2} = \sqrt{30,25+2,25} = \sqrt{32,5} = 5,70$
cluster2 = {B2, A3}

$d(B3,G1) = \sqrt{(6-3,66)^2 + (4-9)^2} = \sqrt{9+30,25} = \sqrt{39,25} = 6,26$
$d(B3,G2) = \sqrt{(6-7)^2 + (4-4,33)^2} = \sqrt{1+0,10} = \sqrt{1,10} = 1,05$
$d(B3,G3) = \sqrt{(6-1,5)^2 + (4-3,5)^2} = \sqrt{20,25+0,25} = \sqrt{20,5} = 4,52$
cluster2 = {B3, B2, A3}

$d(C1,G1) = \sqrt{(1-3,66)^2 + (2-9)^2} = \sqrt{4+56,25} = \sqrt{60,25} = 7,76$
$d(C1,G2) = \sqrt{(1-7)^2 + (2-4,33)^2} = \sqrt{36+5,42} = \sqrt{41,42} = 6,43$
$d(C1,G3) = \sqrt{(1-1,5)^2 + (2-3,5)^2} = \sqrt{0,25+2,25} = \sqrt{2,5} = 1,58$
cluster3 = {C1, A2}

$d(C2,G1) = \sqrt{(4-3,66)^2 + (9-9)^2} = \sqrt{0,11+0} = \sqrt{0,11} = 0,34$
$d(C2,G2) = \sqrt{(4-7)^2 + (9-4,33)^2} = \sqrt{9+21,80} = \sqrt{30,80} = 5,55$
$d(C2,G3) = \sqrt{(4-1,5)^2 + (9-3,5)^2} = \sqrt{6,25+30,25} = \sqrt{36,5} = 6,04$
cluster1 = {C2, B1, A1}

(2) les trois clusters finaux (0,5 pt)

cluster
cluster1 = {C2, B1, A1}
cluster2 = {B3, B2, A3}
cluster3 = {C1, A2}

Rédiger les deux exercices séparément.

BON COURAGE !