COMP1942 Exploring and Visualizing Data (Spring Semester 2019)
Midterm Examination (Answer Sheet)
Date: 8 April, 2019 (Monday)
Time: 10:40am-11:50am
Duration: 1 hour 10 minutes

Student ID:_____ Student Name:_____

Seat No. :_____

Instructions:
(1) Please answer **all** questions in this paper.
(2) The total marks are 100.
(3) You can use a calculator.

Answer Sheet

Question	Full Mark	Mark
Q1	20	
Q2	20	
Q3	20	
Q4	20	
Q5	20	
Total	100	

Q1 (20 Marks)
(a) (i)

No.

Consider the following example.

B	C
1	1
1	1
1	1

In Step 1, {B, C} and {B} are in S_1
since supp({B, C}) ≥ 3 and supp({B})≥3

In Step 2, B→C is generated since supp({B,C})/supp({B})=100% ≥50%

Thus, B→C is in S_2

Note that
supp(B→C) =supp({B,C})
= 3
< 4

In conclusion, B→C is in S_2 but supp(B→C) < 4

Q1 (continued)
(a) (ii)

Yes.

Since B→C is in S_0,
conf(B→C) ≥50%
supp({B,C})/supp({B}) ≥50%
Since B→C is in S_0,
supp(B→C) ≥ 4
Since supp({B,C}) = supp(B→C),
supp({B,C})≥4
Thus, {B, C} is in S_1.
Since supp({B, C}) ≥ 4,
supp({B}) ≥ 4
Thus, {B} is in S_1

Since {B} is in S_1,
and {B, C} is in S_1,
Step 2 must consider
{B} and {B, C} together, and
generate B→C (since supp({B,C})/supp({B}) ≥ 50%)
B→C is in S_2.

Q1 (continued)
(b)(i)

Yes.

Since "B→C" is in S_2,
we know that
we have to calculate supp({B,C})/supp({B}) in Step (*)

In other words,
{B, C} and {B} are in S_1
which means that
supp({B, C}) ≥ 4 and
supp({B}) ≥ 4

Since supp(B→C) = supp({B, C}),
supp(B→C) ≥ 4
Thus,
supp(B→C) ≥ 4

Q1 (continued)
(b) (ii)

No.

Consider the following example.

B	C
1	1
1	1
1	1
1	1
1	0
1	0
1	0

B→C is in S_0
(since conf(B→C) = 57.14% and supp(B→C) = 4)

Since supp({B, C}) = 4 and supp({B}) = 7,
both {B, C} and {B} are in S_1.
However, in Step 2, B→C is not generated in the output set S_2 because
supp({B, C})/supp({B}) = 57.14% which is smaller than 60%.

Q2 (20 Marks)
(a)

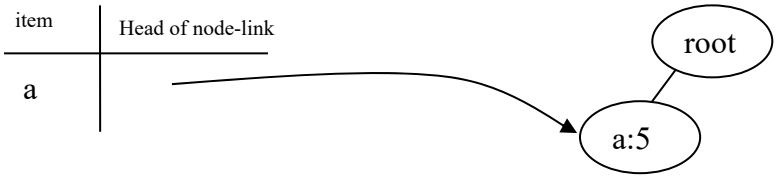
$L_1 = \{\{A\}, \{C\}, \{D\}, \{E\}\}$
Large 2-itemset Generation:
Join Step/Prune Step
$C_2 = \{\{A, C\}, \{A, D\}, \{A, E\}, \{C, D\}, \{C, E\}, \{D, E\}\}$
Counting Step
$L_2 = \{\{A, C\}, \{A, D\}, \{A, E\}, \{C, E\}, \{D, E\}\}$
Large 3-itemset Generation:
Join Step
$C_3 = \{\{A, C, D\}, \{A, C, E\}, \{A, D, E\}\}$
Prune Step
$C_3 = \{\{A, C, E\}, \{A, D, E\}\}$
Counting Step
$L_3 = \{\{A, C, E\}, \{A, D, E\}\}$
Large 4-item set Generation:
Join Step
$C_4 = \{\}$

Large itemsets = $L_1 \cup L_2 \cup L_3$
$= \{\{A\}, \{C\}, \{D\}, \{E\}, \{A, C\}, \{A, D\}, \{A, E\}, \{C, E\}, \{D, E\}, \{A, C, E\}, \{A, D, E\}\}$

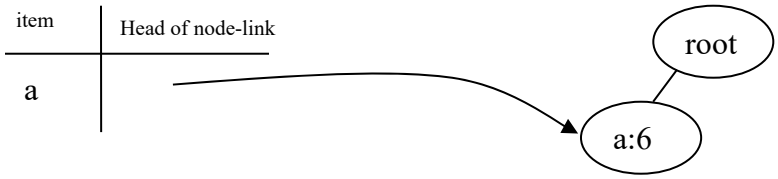
Q2 (Continued)

Q2 (Continued)

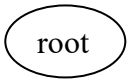
(b)(i)



(ii)



(iii)



(iv)

{a}, {b}, {c}, {a, c}, {a, b}

Q3 [20 Marks]

(a)(i)

Info(T) = $1-0.5^2-0.5^2=0.5$

For attribute Age,

Info(T_{young})=$1-0.5^2-0.5^2=0.5$

Info(T_{old})=$1-0.5^2-0.5^2=0.5$

Info(Age, T)=$\frac{1}{2}Info(T_{young})+\frac{1}{2}Info(T_{old})=0.5$

Gain(Age, T)= Info(T)-Info(Age, T)=0

For attribute Gender,

Info(T_{male})=$1-1^2-0^2=0$

Info(T_{female})=$1-(\frac{1}{3})^2-(\frac{2}{3})^2=0.4444$

Info(Gender, T)= 1/4 Info(T_{male}) + 3/4 Info(T_{female}) = 0.3333

Gain(Gender, T)= Info(T)-Info(Gender, T)= 0.1667

For attribute MMR_Vaccine,

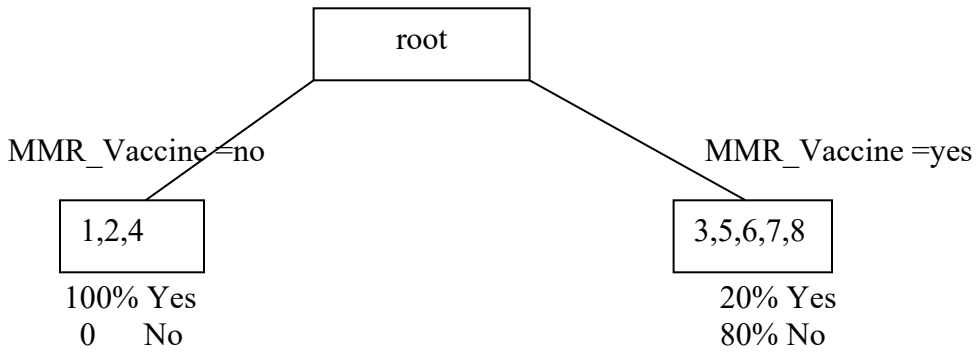
Info(T_{no})= $1-1^2-0^2=0$

Info(T_{yes})= $1-(\frac{1}{5})^2-(\frac{4}{5})^2=0.32$

Info(MMR_Vaccine, T)= 3/8 Info(T_{no}) + 5/8 Info(T_{yes}) = 0.2

Gain(MMR_Vaccine, T)= Info(T)-Info(MMR_Vaccine, T)=0.3

We choose attribute MMR_Vaccine for Splitting:



Q3 (Continued)

Q3 (Continued)
(a) (ii)

It is likely that he will not have measles.

(b)
Differences:
The definition of the gain used in C4.5 is different from that used in ID3.
The gain used in C4.5 is equal to the gain used in ID3 divided by SplitInfo.

The reason why there is a difference is described as follows.
In ID3, there is a higher tendency to choose an attribute containing more values (e.g., attribute identifier and attribute HKID). Thus, splitInfo in C4.5 is used to penalize an attribute containing more values. If this value is larger, the penalty is larger.

Q4 [20 Marks]
(a)

Yes.
Cluster 1: $\{x_1, x_2, x_4, x_5, x_6\}$
Cluster 2: $\{x_3, x_7, x_8, x_9, x_{10}\}$

(b) (i)

Yes.
Cluster 1: $\{x_2, x_5, x_6\}$
Cluster 2: $\{x_3, x_7, x_8, x_9, x_{10}\}$
Cluster 3: $\{x_4\}$
Cluster 4: $\{x_1\}$

(ii)

Yes.
Cluster 1: $\{x_1, x_2, x_4, x_5, x_6\}$
Cluster 2: $\{x_3, x_7, x_8, x_9, x_{10}\}$

Q4 (Continued)

(b) (iii)

No.
This is because in the dendrogram, we have to specify the distance between 2 clusters.
However, when we use the centroid linkage as a distance measurement between 2 clusters, we have to know the coordinate of each point (and thus the mean/center of all points in each cluster), which could not be found in the information given.

Q5 [20 Marks]

(a)(i)

- Make initial guesses for the means $m_1, m_2, \ldots, m_k$
- Until Interrupted
  - Acquire the next example x
  - If m_i is closest to x,
    - replace mi by $m_i + a(x - m_i)$

(ii)

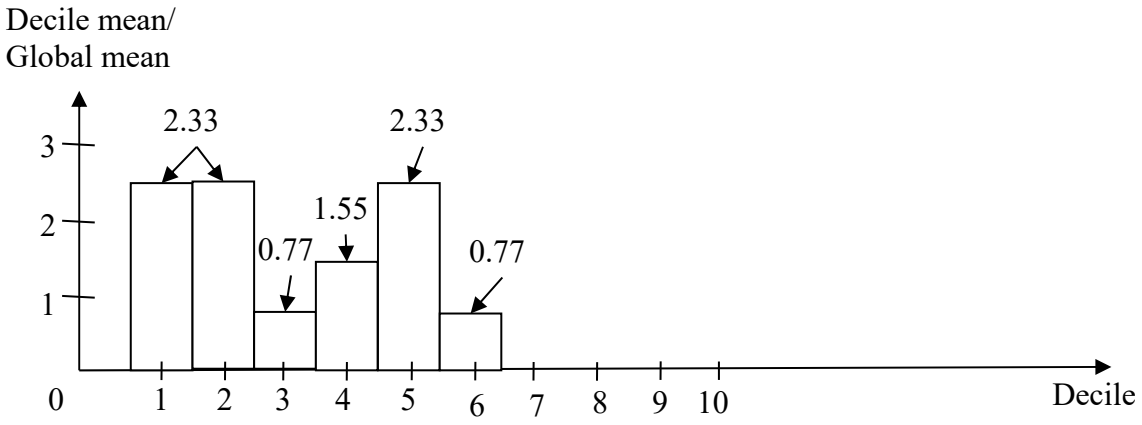
$$
\begin{aligned}
m_n &= m_{n-1} + a(x_n - m_{n-1}) \\
&= (1-a)m_{n-1} + ax_n \\
&= (1-a)[(1-a)m_{n-2} + ax_{n-1}] + ax_n \\
&= (1-a)^2 m_{n-2} + (1-a)ax_{n-1} + ax_n \\
&= (1-a)^2[(1-a)m_{n-3} + ax_{n-2}] + (1-a)ax_{n-1} + ax_n \\
&= (1-a)^3 m_{n-3} + (1-a)^2 ax_{n-2} + (1-a)ax_{n-1} + ax_n \\
&= \ldots \\
&= (1-a)^n m_0 + \sum_{p=1}^{n} (1-a)^{n-p} ax_p
\end{aligned}
$$

$X = (1-a)^n$

$Y = (1-a)^{n-p}a$

Q5 (Continued)
(b)

Yes. The chart is shown as follows.



End of Answer Sheet