COMP1942 Exploring and Visualizing Data (Spring Semester 2017)
Midterm Examination (Answer Sheet)
Date: 17 March, 2017 (Fri)
Time: 9:00-10:15
Duration: 1 hour 15 minutes

Student ID:__________________ Student Name:__________________________________

Seat No. :__________________

Instructions:
(1) Please answer **all** questions in **Part A** in this paper.
(2) You can **optionally** answer the bonus question in **Part B** in this paper. You can obtain additional marks for the bonus question if you answer it correctly.
(3) The total marks in Part A are 100.
(4) The total marks in Part B are 10.
(5) The total marks you can obtain in this exam are 100 only.
If you answer the bonus question in Part B correctly, you can obtain additional marks.
But, if the total marks you obtain from Part A and Part B are over 100, your marks will be truncated to 100 only.
(6) You can use a calculator.

Answer Sheet

Part	Question	Full Mark	Mark
A	Q1	20	
	Q2	20	
	Q3	20	
	Q4	20	
	Q5	20	
Total (Part A)		100	
B	Q6 (OPTIONAL)	10	
Total (Parts A and B)		100	

Part A (Compulsory Short Questions)

Q1 (20 Marks)
(a) (i)

No.

Consider the following example.

B	C
1	1
1	1

In Step 1, {B, C} and {B} are in S_1
since supp({B, C}) ≥ 2 and supp({B})≥2

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})
= 2
< 3

In conclusion, B→C is in S_2 but supp(B→C) < 3

Q1 (continued)
(a) (ii)

Yes.

Since $B \rightarrow C$ is in S_0,
 $conf(B \rightarrow C) \geq 50\%$
 $supp(\{B,C\})/supp(\{B\}) \geq 50\%$
Since $B \rightarrow C$ is in S_0,
 $supp(B \rightarrow C) \geq 3$
Since $supp(\{B,C\}) = supp(B \rightarrow C)$,
 $supp(\{B,C\}) \geq 3$
Thus, $\{B, C\}$ is in S_1.
Since $supp(\{B, C\}) \geq 3$,
 $supp(\{B\}) \geq 3$
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 $B \rightarrow C$ (since $supp(\{B,C\})/supp(\{B\}) \geq 50\%$)
$B \rightarrow C$ is in S_2.

Q1 (continued)
(b)(i)

Yes.

Since "$B \rightarrow C$" 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
$supp(\{B, C\}) \geq 4$ and
$supp(\{B\}) \geq 4$

Since $supp(B \rightarrow C) = supp(\{B, C\})$,
$supp(B \rightarrow C) \geq 4$
Thus,
$supp(B \rightarrow C) \geq 3$

Q1 (continued)
(b) (ii)

No.

Consider the following example.

B	C
1	1
1	1
1	1

B$\rightarrow$C is in S_0
(since conf(B$\rightarrow$C) = 100% and supp(B$\rightarrow$C) = 3)

Since supp({B, C}) = 3,
in Step 1, {B, C} is not in S_1.
In order to generate B$\rightarrow$C in the output set S_2 in Step 2,
both {B, C} and {B} must be in S_1.
We deduce that B$\rightarrow$C is not in S_2.

Q2 (20 Marks)

Item	Freq
a	3
b	3
c	5
d	5
e	1
f	6
g	1
h	1
i	1
j	1
k	1
l	1
m	1
n	1
o	1
p	1
q	1

Freq items:

Item	Freq
a	3
b	3
c	5
d	5
f	6

Sorted Freq items:

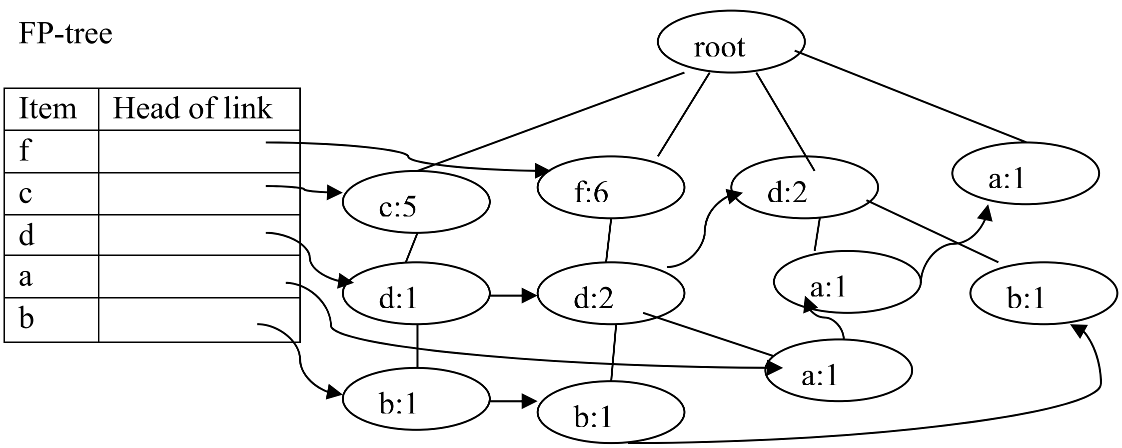
Item	Freq
f	6
c	5
d	5
a	3
b	3

Q2 (continued)

Ordered freq items

TID	Items bought	(ordered) freq items
1	b,c,d,p	c,d,b
2	f,j,q	f
3	c,i	c
4	a,d	d,a
5	c,m	c
6	b,d,f	f,d,b
7	a,d,f	f,d,a
8	e,f	f
9	f,h	f
10	b,d	d,b
11	a,l	a
12	c,g	c
13	c,k	c
14	f,n,o	f

FP-tree



Q2 (continued)

Conditional FP-tree on "b" **count (b)=3**

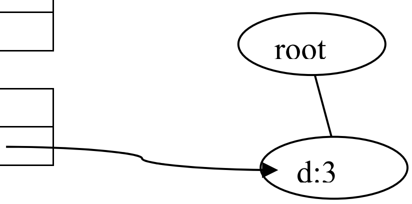
(c:1,d:1,b:1) (b:1,d:1)
(f:1,d:1,b:1) ⇒ (b:1,d:1)
(d:1,b:1) (b:1,d:1)

Item	Freq
f	1
c	1
d	3
a	0
b	3

↓

Item	Freq
b	3
d	3

Item	Head
d	



{b,d}:3

Conditional FP-tree on "a" **count (a)=3**

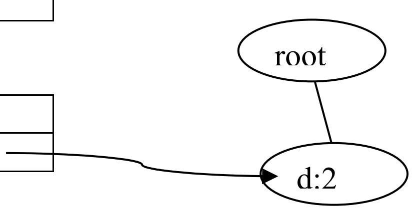
(f:1,d:1,a:1) (a:1,d:1)
(d:1,a:1) ⇒ (a:1,d:1)
(a:1) (a:1)

Item	Freq
f	1
c	0
d	2
a	3
b	0

↓

Item	Freq
a	3
d	2

Item	Head
d	



{a,d}:2

Q2 (continued)

Conditional FP-tree on "d" **count (d)=5**

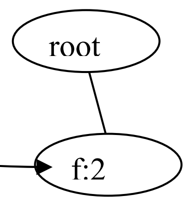
(c:1,d:1) ⇒ (d:1)
(f:2,d:2) ⇒ (d:2,f:2)
(d:2) ⇒ (d:2)

Item	Freq
f	2
c	1
d	5
a	0
b	0

↓

Item	Freq
d	5
f	2

Item	Head
f	



{d,f}:2

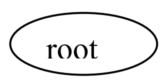
Conditional FP-tree on "c" **count (c)=5**

(c:5) ⇒ (c:5)

Item	freq
c	5

↓

Item	freq
c	5



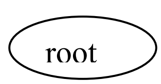
Conditional FP-tree on "f" **count (f)=6**

(f:6) ⇒ (f:6)

Item	freq
f	6

↓

Item	freq
f	6



Freq itemsets

={{b},{b,d},
{a},{a,d},
{d},{d,f},
{c},
{f}}

Q2 (continued)

Q2 (continued)

Q3 (20 Marks)

(a) (i)

Make initial guesses of the means $m_1, m_2, \ldots, m_k$
Set the counts $n_1, n_2, \ldots, n_k$ to zero
Until interrupted
  Acquire the next example x
  If m_i is closest to x
    Increment n_i
    Replace m_i by $m_i + 1/n_i (x - m_i)$

Q3 (continued)

(a) (ii)

x_j : the j-th example in cluster i

$m_i(t)$: the mean vector of cluster i containing t examples

Consider that x is the t-th example in cluster i

$$m_i(t-1) = \frac{x_1 + x_2 + \ldots + x_{t-1}}{t-1}$$

$$m_i(t) = \frac{x_1 + x_2 + \ldots + x_{t-1} + x_t}{t}$$

$$= \frac{m_i(t-1) \times (t-1) + x_t}{t}$$

$$= \frac{t \times m_i(t-1) + x_t - m_i(t-1)}{t}$$

$$= m_i(t-1) + \frac{1}{t}(x_t - m_i(t-1))$$

Q3 (continued)

(b) (i)

2

(b) (ii)

Cluster 1:

Final mean: (22, 95.6)
Initial mean: (20, 95)

Cluster 2:

Final mean: (91.8, 42.2)
Initial mean: (89, 42)

Q4 (20 Marks)

(a)

$$\begin{array}{c|c} & \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} & \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 4 & 3 & 0 & \\ 5.10 & 4.12 & 1.41 & 0 \end{pmatrix} \end{array}$$

e.g., Value 1 above (i.e., the entry for (x1, x2)) is equal to $\sqrt{(2-3)^2+(3-3)^2}$

(b)

$$\begin{array}{c|c} & \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} & \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 4 & 3 & 0 & \\ 5.10 & 4.12 & 1.41 & 0 \end{pmatrix} \end{array}$$

x_1: (2, 3)
x_2: (3, 3)
x_3: (6, 3)
x_4: (7, 2)

$$\begin{array}{c|c} & \begin{array}{ccc} (x_1x_2) & x_3 & x_4 \end{array} \\ \begin{array}{c} (x_1x_2) \\ x_3 \\ x_4 \end{array} & \begin{pmatrix} 0 & & \\ 3.5 & 0 & \\ 4.61 & 1.41 & 0 \end{pmatrix} \end{array}$$

(x_1x_2): (2.5, 3)
x_3: (6, 3)
x_4: (7, 2)

Note: Value 3.5 above is calculated by $\sqrt{(2.5-6)^2+(3-3)^2}$

$$\begin{array}{c|c} & \begin{array}{ccc} (x_1x_2) & x_3 & x_4 \end{array} \\ \begin{array}{c} (x_1x_2) \\ x_3 \\ x_4 \end{array} & \begin{pmatrix} 0 & & \\ 3.5 & 0 & \\ 4.61 & 1.41 & 0 \end{pmatrix} \end{array}$$

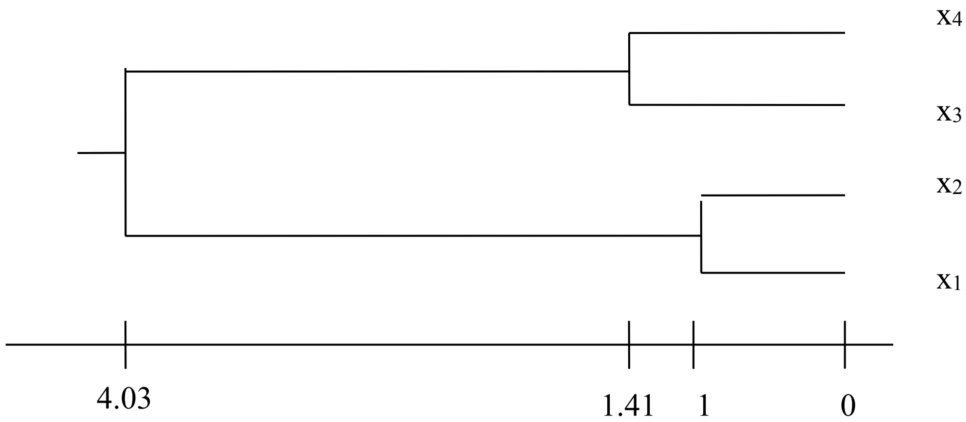
(x_1x_2): (2.5, 3)
x_3: (6, 3)
x_4: (7, 2)

$$\begin{array}{c|c} & \begin{array}{cc} (x_1x_2) & (x_3x_4) \end{array} \\ \begin{array}{c} (x_1x_2) \\ (x_3x_4) \end{array} & \begin{pmatrix} 0 & \\ 4.03 & 0 \end{pmatrix} \end{array}$$

(x_1x_2): (2.5, 3)
(x_3x_4): (6.5, 2.5)

Q4 (continued)

Dendrogram



Q5 (20 Marks)

(a)(i)

Info(T) = $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

For attribute Age,

Info(T_{young})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{old})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

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

SplitInfo(Age)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Gain(Age, T)=$\frac{1-1}{1}=0$

For attribute Income,

Info(T_{high})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{medium})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{low})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(Income, T)= $\frac{1}{2}Info(T_{high})+\frac{1}{4}Info(T_{medium})+\frac{1}{4}Info(T_{low})=1$

SplitInfo(Income)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{4}\log\frac{1}{4}=1.5$

Gain(Income, T)= $\frac{1-1}{1.5}=0$

For attribute Credit-Rating,

Info(T_{high})= $-1\log 1-0\log 0=0$

Info(T_{fair})=$-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}=0.8113$

Info(T_{low})=$-0\log 0-1\log 1=0$

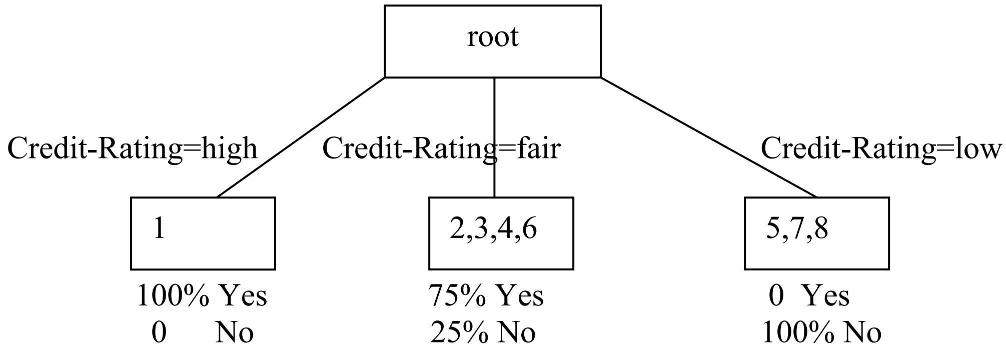
Info(Credit-Rating)= $\frac{1}{8}Info(T_{high})+\frac{1}{2}Info(T_{fair})+\frac{3}{8}Info(T_{low})=0.405$

SplitInfo(Credit-Rating)= $-\frac{1}{8}\log\frac{1}{8}-\frac{1}{2}\log\frac{1}{2}-\frac{3}{8}\log\frac{3}{8}=1.4056$

Gain(Credit-Rating, T)= $\frac{1-0.405}{1.4056}=0.4233$

Q5 (Continued)

We choose attribute Credit-Rating for Splitting:



Q5 (Continued)

(a)(ii) It is very likely that this customer will buy an apple watch.

(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.

Part B (Bonus Question)

Note: The following bonus question is an **OPTIONAL** question. You can decide whether you will answer it or not.

Q6 (10 Additional Marks)

The Apriori property is:
If a set Y of dimensions has good clustering, then any proper subset of S must have good clustering.
(i.e., For any two sets of dimensions, namely Y and Z, where $Z \subset Y$, if $H(Y) < \omega$, $H(Z) < \omega$.)

Next, we show the correctness of this property.
We want to show that
"for any two sets of dimensions, namely Y and Z, where $Z \subset Y$, if $H(Y) < \omega$, $H(Z) < \omega$."

Consider two cases.
Case 1: $|Y - Z| = 1$
In this case, let $X_i = Y\text{-}Z$ (where X_i is a dimension). Note that $Y = Z \cup \{X_i\}$.

Given $c' \in A(Z)$ and $x \in [1, 4]$, we define $\alpha(c', x)$ to be the set of all grids in c' together with the grid "X_i: $[10(x-1)+1, 10(x-1)+10]$".

We know that $A(Z \cup \{X_i\}) = \{\alpha(c', x) \mid c' \in A(Z) \text{ and } x \in [1, 4]\}$

Consider

$$
\begin{aligned}
& H(Y) \\
&= H(Z \cup \{X_i\}) \\
&= -\sum_{c \in A(Z \cup \{X_i\})} d(c) \log d(c) \\
&= -\sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log d(\alpha(c', x)) \\
&= -\sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log[d(c') \cdot \frac{d(\alpha(c', x))}{d(c')}] \\
&= -\sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log d(c') - \sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log \frac{d(\alpha(c', x))}{d(c')} \\
&= -\sum_{c' \in A(Z)} d(c') \log d(c') - \sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log \frac{d(\alpha(c', x))}{d(c')} \\
&= H(Z) - \sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log \frac{d(\alpha(c', x))}{d(c')} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(*)
\end{aligned}
$$

We know that $d(\alpha(c', x)) \le d(c')$ and thus $\log \frac{d(\alpha(c', x))}{d(c')} \le 0$.

Since $d(\alpha(c', x)) \ge 0$, we know that $\sum_{c' \in A(Z)} \sum_{x \in [1,4]} d(\alpha(c', x)) \log \frac{d(\alpha(c', x))}{d(c')} \le 0$

Q6 (Continued)

Thus, from (*), we deduce that $H(Y) \geq H(Z)$.

We know that $H(Y) < \omega$. We deduce that $H(Z) < \omega$.

Case 2: $|Y - Z| > 1$

In Case 1, we know that $H(Z \cup \{X_i\}) \geq H(Z)$ (since $Y = Z \cup \{X_i\}$).
By similar derivations, we could deduce that for any set X of dimensions, $H(Z \cup X) \geq H(Z)$.
Thus, we know that if $H(Y) < \omega$, then $H(Z) < \omega$.

Algorithm:

The idea is similar to the original Apriori Algorithm learnt in class.
Algorithm:

1. $L_1 \leftarrow$ a set of dimensions where each dimension has good clustering
2. $k \leftarrow 2$
3. while $L_{k-1} \neq \emptyset$
    - $C_k \leftarrow$ Generate candidates from L_{k-1} by Join Step and Prune Step discussed in class
    - Perform a counting step on C_k (i.e., computing the H value of each element in C_k) and obtain L_k (i.e., keeping each element in C_k with the H value smaller than ω)
4. Output $\bigcup_i L_i$

Q6 (Continued)

Q6(Continued)

End of Answer Sheet