COMP1942 Exploring and Visualizing Data (Spring Semester 2018)
Midterm Examination (Answer Sheet)
Date: 16 March, 2018 (Fri)
Time: 3:15pm-4:15pm
Duration: 1 hour

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) confidence = 2/3 = 66.7%

(ii) freq. itemsets
= { {A}, {C}, {D}, {E},
{A, C}, {A, D}, {A, E}, {C, E}, {D, E},
{A, C, E}, {A, D, E} }

(b)
Yes.

Confidence of "{A, B} → C" = support({A, B, C})/support({A, B})
= support({A, B, C})/4

Expected Confidence of the Consequence of "{A, B} → C" = support({C})/total number of transactions
= 6/10

Lift ratio = Confidence of "{A, B} → C"/ Expected Confidence of the Consequence of "{A, B} → C"
1.25 = (support({A, B, C})/4)/(6/10)
support({A, B, C}) = 3

The support of "{A, B}→C" = support({A, B, C})
= 3

Q2 (20 Marks)

(a)

The corresponding transactions are:

TID	Items
1	a, b, c
2	a, b
3	a, c
4	a, c
5	a
6	a

(b)

No. This is because whenever we construct a conditional FP-tree on an item A from the FP-tree, some itemsets including item A may become infrequent and will be discarded. Thus, the information about this infrequent itemsets cannot be found in the conditional FP-tree. We could not construct the FP-tree from all conditional FP-trees constructed.

Q2 (Continued)

(c) (i)
Yes. This is because L_2 is exactly equal to the set of itemsets in C_2 which frequency is at least a given support threshold.

(ii)

No. Suppose that $L_1 = \{\{A\}, \{B\}\}$. Then, $C_2 = \{\{A, B\}\}$. In this case, the number of itemsets in C_2 is smaller than the number of itemsets in L_1.

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 $m_i + 1/n_i\ (x - m_i)$

(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)

Yes. This is because the user sets the same seed as “12345”. No matter which machine he used, the same clustering result will be generated.

(ii)

No. This is because the supports of all rules in Scenario 2 are equal to 3. If we change the threshold to 4, those rules will not be in the output.

(iii)

Yes. This is because the confidences of all rules in Scenario 2 are at least 75%. If we change the threshold to 70%, those rules will still be in the output.

Q4 (20 Marks)

(a) (i)

the distance between the first cluster and the second cluster
= [dist(x1, x4) + dist(x1, x5) + dist(x2, x4) + dist(x2, x5) + dist(x3, x4) + dist(x3, x5)]/6
= (12.73+15.56+8.49+11.31+11.40+14.21)/6
= 12.28

(ii)

Mean of the first cluster = ((1+4+3)/3, (2+5+2)/3) = (2.67, 3)

Mean of the second cluster = ((10+12)/2, (11+13)/2) = (11, 12)

the distance between the first cluster and the second cluster
= the distance between the mean of the first cluster and the mean of the second cluster
= 12.26 (or 12.27)

(iii)

Center of the first cluster = ([(1+3)/2+4]/2, [(2+2)/2+5]/2) = (3, 3.5)

Center of the second cluster = ((10+12)/2, (11+13)/2) = (11, 12)

the distance between the first cluster and the second cluster
= the distance between the center of the first cluster and the center of the second cluster
= 11.67

Q4 (Continued)

(b)

No. This is because we do not know the distance between cluster (a, b) and cluster (c d) and the distance between (a b) and e.

Q5 (20 Marks)

(a)(i)

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

For attribute Gender,

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

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

Info(Gender, T)=0.5 Info(T_{male}) + 0.5 Info(T_{female}) = 1

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

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

For attribute Year,

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

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

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

Info(Year, T)= 1/2 x Info(T_{three}) + 1/4 x Info(T_{two}) + 1/4 x Info(T_{one}) = 1

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

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

For attribute School,

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

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

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

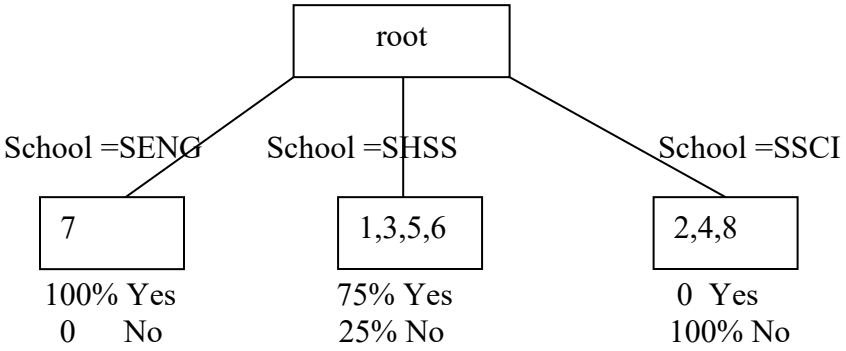
Info(School)=1/8 x Info(T_{SENG}) + 1/2 x Info(T_{SHSS}) + 3/8 x Info(T_{SSCI}) = 0.405

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

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

Q5 (Continued)

We choose attribute School for Splitting:



Q5 (Continued)

Q5 (Continued)

(a) (ii)

It is very likely that this user will study COMP1942.

(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.

End of Answer Sheet