COMP1942 Exploring and Visualizing Data (Spring Semester 2013)
Final Examination (Answer Sheet)
Date: 21 May, 2013 (Tue)
Time: 12:30-15:30
Duration: 3 hours

Student ID:_____ Student Name:_____

Seat No. :_____

Instructions:
(1) Please answer **all** questions in **Part A** and **Part B** in this paper.
(2) You can **optionally** answer the bonus question in **Part C** in this paper. You can obtain additional marks for the bonus question if you answer it correctly.
(3) The total marks in Part A and part B are 200.
(4) The total marks in Part C are 20.
(5) The total marks you can obtain in this exam are 200 only.
If you answer the bonus question in Part C correctly, you can obtain additional marks.
But, if the total marks you obtain from Part A, Part B and Part C are over 200, your marks will be truncated to 200 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	
	Q6	20	
	Q7	20	
	Q8	20	
	Q9	20	
B	Q10-Q13	20	
Total (Parts A and B)		200	
C	Q14 (OPTIONAL)	20	
Total (Parts A, B and C)		200	

Part A (Compulsory Short Questions)

Q1 (20 Marks)

(a)

P	Q	R	S	T
1	1	1	0	0
1	1	1	1	0
1	1	0	1	0

(b) (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.

(c) (i)

{a}:4, {b}:4, {d}:3, {e}:3, {f}:3, {g}:3,
{a, b}:3, {a, g}:3, {b, d}:3, {b, e}:3

(ii)

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 (20 Marks)

(a)

The scenario is when we are more interested in more recent data points compared with less recent data points.

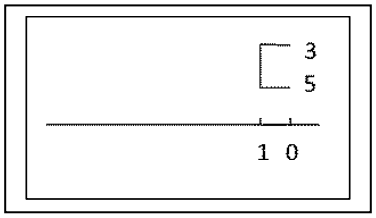
(b)

	1	2	3	4	5	6	7	8
1	0							
2	11	0						
3	5	13	0					
4	12	2	14	0				
5	7	17	1	18	0			
6	13	4	15	5	20	0		
7	9	15	12	16	15	19	0	
8	11	20	12	21	17	22	30	0

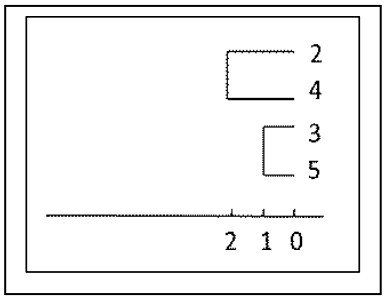
Distance

	1	2	(35)	4	6	7	8
1	0						
2	11	0					
(35)	6	15	0				
4	12	2	16	0			
6	13	4	17.5	5	0		
7	9	15	13.5	16	19	0	
8	11	20	14.5	21	22	30	0

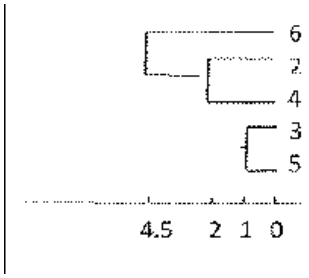
Dendrogram



	1	(24)	(35)	6	7	8
1	0					
(24)	11.5	0				
(35)	6	15.5	0			
6	13	4.5	17.5	0		
7	9	15.5	13.5	19	0	
8	11	20.5	14.5	22	30	0

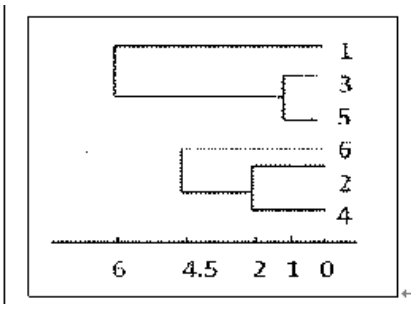


	1	(246)	(35)	7	8
1	0				
(246)	12	0			
(35)	6	16.17	0		
7	9	16.67	13.5	0	
8	11	21	14.5	30	0

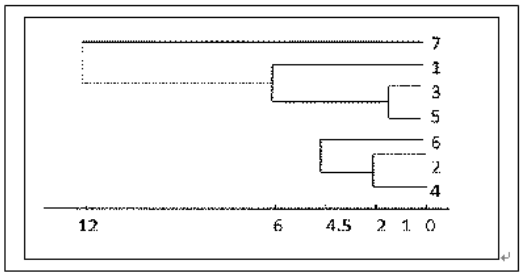


Q2 (Continued)

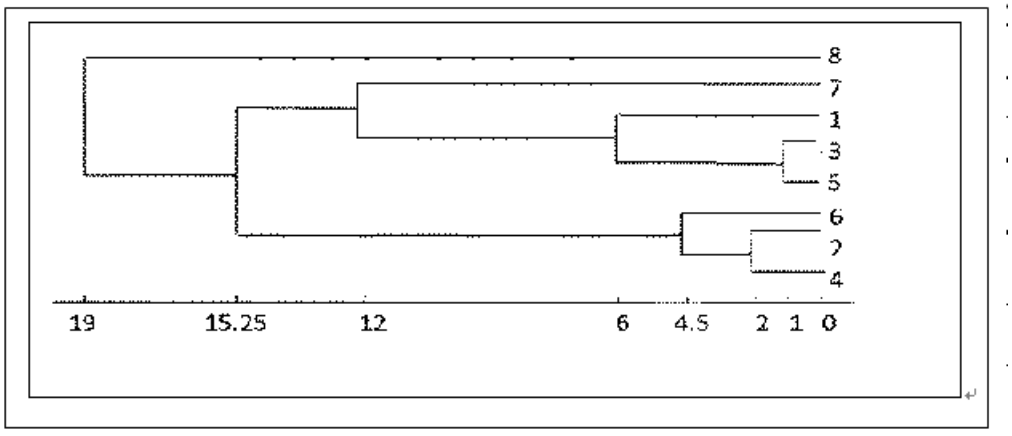
$$\begin{array}{c} \\ (135) \\ (246) \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccc} (135) & (246) & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & & & \\ 14.78 & 0 & & \\ 12 & 16.67 & 0 & \\ 13.33 & 21 & 30 & 0 \end{pmatrix} \end{array}$$



$$\begin{array}{c} \\ (1357) \\ (246) \\ 8 \end{array} \begin{array}{c} \begin{array}{ccc} (1357) & (246) & 8 \end{array} \\ \begin{pmatrix} 0 & & \\ 15.25 & 0 & \\ 17.5 & 21 & 0 \end{pmatrix} \end{array}$$



$$\begin{array}{c} \\ (1234567) \\ 8 \end{array} \begin{array}{c} \begin{array}{cc} (12345467) & 8 \end{array} \\ \begin{pmatrix} 0 & \\ 19 & 0 \end{pmatrix} \end{array}$$



Q2 (Continued)

Q2 (Continued)

Q3 (20 Marks)

(a)

$$P(X, Y \mid Z)$$

$$= \frac{P(X,Y,Z)}{P(Z)}$$

$$= \frac{P(X,Y,Z)}{P(Y,Z)} \times \frac{P(Y,Z)}{P(Z)}$$

$$= P(X|Y, Z) \times P(Y|Z)$$

$$= P(X|Z) \times P(Y|Z)$$

(b)

The curse of dimensionality can be described as follows.
When the number of dimensions increases, the distance between any two points is nearly the same.

(c)(i)

Neural Network

(ii)

(1) Parallel processing – each neuron operates individually
(2) Fault tolerance – if a small number of neurons break down, the whole system is still able to operate with slight degradation in performance
(3) A universal approximator – it can model all types of function y = f(x)

(Any two of the above advantages are OK.)

Q4 (20 Marks)

(a)

$$\text{mean vector} = \begin{pmatrix} \frac{6+8+5+9}{4} \\ \frac{6+8+9+5}{4} \end{pmatrix} = \begin{pmatrix} 7 \\ 7 \end{pmatrix}$$

For data (6, 6), difference from mean vector $= \begin{pmatrix} 6-7 \\ 6-7 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$

For data (8, 8), difference from mean vector $= \begin{pmatrix} 8-7 \\ 8-7 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

For data (5, 9), difference from mean vector $= \begin{pmatrix} 5-7 \\ 9-7 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$

For data (9, 5), difference from mean vector $= \begin{pmatrix} 9-7 \\ 5-7 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$

$$Y = \begin{pmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{pmatrix}$$

$$\Sigma = \frac{1}{4} YY^T = \frac{1}{4} \begin{pmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & 1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix}$$

$$= \frac{1}{4} \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{5}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2} \end{pmatrix}$$

$$\begin{vmatrix} \frac{5}{2}-\lambda & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-\lambda \end{vmatrix} = 0 \quad \Rightarrow \quad (\frac{5}{2}-\lambda)^2 - (-\frac{3}{2})^2 = 0 \quad \Rightarrow \quad \lambda = 4 \quad or \quad \lambda = 1$$

when $\lambda = 4$,

$$\begin{pmatrix} \frac{5}{2}-4 & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} -\frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & -\frac{3}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow x_1 + x_2 = 0$$

Q4 (Continued)

We choose the eigenvector of unit length: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$.

When $\lambda = 1$,

$$\begin{pmatrix} \frac{5}{2}-1 & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{3}{2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow x_1 - x_2 = 0$$

We choose the eigenvector of unit length: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$.

Thus, $\Phi = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$, $Y = \Phi^T X = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} X$.

For data (6, 6), $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 6 \\ 6 \end{pmatrix} = \begin{pmatrix} 0 \\ 8.49 \end{pmatrix}$

For data (8, 8), $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 8 \\ 8 \end{pmatrix} = \begin{pmatrix} 0 \\ 11.31 \end{pmatrix}$

For data (5, 9), $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 5 \\ 9 \end{pmatrix} = \begin{pmatrix} -2.83 \\ 9.90 \end{pmatrix}$

For data (9, 5), $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 9 \\ 5 \end{pmatrix} = \begin{pmatrix} 2.83 \\ 9.90 \end{pmatrix}$

The mean vector of the above transformed data points is $\begin{pmatrix} \frac{0+0+(-2.83)+2.83}{4} \\ \frac{8.49+11.31+9.90+9.90}{4} \end{pmatrix} = \begin{pmatrix} 0 \\ 9.90 \end{pmatrix}$

The final transformed data points are:

Q4 (Continued)

For data (6, 6), final transformed vector = $\begin{pmatrix} 0-0 \\ 8.49-9.90 \end{pmatrix} = \begin{pmatrix} 0 \\ -1.41 \end{pmatrix}$

For data (8, 8), final transformed vector = $\begin{pmatrix} 0-0 \\ 11.31-9.90 \end{pmatrix} = \begin{pmatrix} 0 \\ 1.41 \end{pmatrix}$

For data (5, 9), final transformed vector = $\begin{pmatrix} -2.83-0 \\ 9.90-9.90 \end{pmatrix} = \begin{pmatrix} -2.83 \\ 0 \end{pmatrix}$

For data (9, 5), final transformed vector = $\begin{pmatrix} 2.83-0 \\ 9.90-9.90 \end{pmatrix} = \begin{pmatrix} 2.83 \\ 0 \end{pmatrix}$

Thus, (6, 6) is reduced to (0);
(8, 8) is reduced to (0);
(5, 9) is reduced to (-2.83);
(9, 5) is reduced to (2.83).

(Note: Another possible answer is
(6, 6) is reduced to (0);
(8, 8) is reduced to (0);
(5, 9) is reduced to (2.83);
(9, 5) is reduced to (-2.83).
This is because the eigenvectors used in this case are:

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$ and $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$.)

Q4 (Continued)

Q4 (Continued)

Q4 (Continued)

(b)

(5, 5) is reduced to (0);
(7, 7) is reduced to (0);
(4, 8) is reduced to (-2.83);
(8, 4) is reduced to (2.83).

(Note: Another possible answer is
(5, 5) is reduced to (0);
(7, 7) is reduced to (0);
(4, 8) is reduced to (2.83);
(8, 4) is reduced to (-2.83).)

(c)

(18, 18) is reduced to (0);
(24, 24) is reduced to (0);
(15, 27) is reduced to (-8.49);
(27, 15) is reduced to (8.49).

(Note: Another possible answer is
(18, 18) is reduced to (0);
(24, 24) is reduced to (0);
(15, 27) is reduced to (8.49);
(27, 15) is reduced to (-8.49).)

Q5 (20 Marks)
(a)

We need to transform the first three categorical attributes to the other corresponding three numeric attributes.

(b) (i)

Yes. The number is 3.

(ii)

Yes. The number is 7.

(iii)

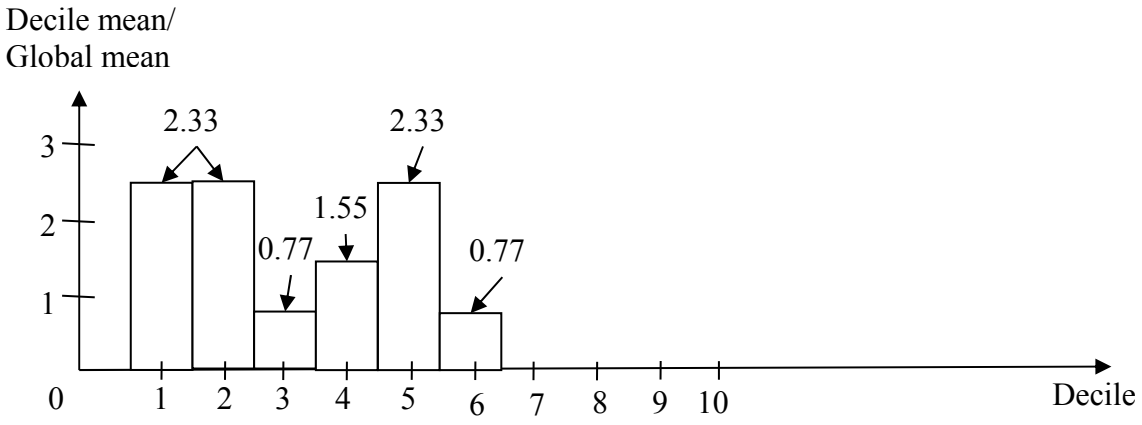
Yes. The number is 6.

(iv)

Yes. The number is 14.

(v)

Yes. The chart is shown as follows.



Q6 (20 Marks)

(a) $P(LC=Yes) = \sum_{x\in\{Yes,No\}}\sum_{y\in\{Yes,No\}} P(LC=Yes \mid FH=x, S=y)P(FH=x, S=y)$

$= 0.7\times0.3\times0.6 + 0.45\times0.3\times0.4 + 0.55\times0.7\times0.6 + 0.2\times0.7\times0.4$

$= 0.467$

$P(LC=Yes \mid FH=Yes, Smoker=Yes, PR=No)$

$= \frac{P(PR=No \mid FH=Yes, Smoker=Yes, LC=Yes)}{P(PR=No \mid FH=Yes, Smoker=Yes)} P(LC=Yes \mid FH=Yes, Smoker=Yes)$

$= \frac{P(PR=No \mid LC=Yes)\times P(LC=Yes \mid FH=Yes, Smoker=Yes)}{\sum_{x\in\{Yes,No\}} P(PR=No \mid LC=x)P(LC=x \mid FH=Yes, Smoker=Yes)}$

$= \frac{0.15\times0.7}{0.15\times0.7 + 0.55\times0.3}$

$= 0.389$

$P(LC=No \mid FH=Yes, Smoker=Yes, PR=No) = 1 - 0.389 = 0.611$

$P(LC=Yes \mid FH=Yes, Smoker=Yes, PR=No) < P(LC=No \mid FH=Yes, Smoker=Yes, PR=No)$

Thus, it is less likely that the person has Lung Cancer.

Q6 (Continued)

(b) Disadvantages:
The Bayesian Belief network classifier requires a predefined knowledge about the network.
The Bayesian Belief Network classifier cannot work directly when the network contains cycles.

Q7 (20 Marks)
(a)

Classifier 1: No
Classifier 2: Yes
Classifier 3: Yes

The overall predicted results is “Yes” (since the majority of the results is “Yes”).

(b)

The target attribute of this new record is “Yes”.
The 3 nearest neighbors are 8, 9 and 13.

(c)

No.

P(Insurance = Yes) = 1/2
P(Insurance = No) = 1/2

P(Insurance = Yes | A = 0) = 3/4
P(Insurance = No | A = 0) = 1/4

P(Insurance = Yes | A = 1) = 3/4
P(Insurance = No | A = 1) = 1/4

P(A = 0) = 1/2
P(A = 1) = 1/2

P(Insurance = Yes | B = 0) = 1
P(Insurance = No | B = 0) = 0

P(Insurance = Yes | B = 1) = 1/3
P(Insurance = No | B = 1) = 2/3

P(B = 0) = 1/8
P(B = 1) = 7/8

Q7 (Continued)

Consider ID3.
Info(T) = - 1/2 log 1/2 - 1/2 log 1/2 = 1

Consider attribute A.
Info(T_0) = -3/4 log 3/4 - 1/4 log 1/4 = 0.8113
Info(T_1) = -3/4 log 3/4 - 1/4 log 1/4 = 0.8113
Info(A, T) = 1/2 Info(T_0) + 1/2 Info(T_1) = 0.8113
Gain(A, T) = Info(T) - Info(A, T) = 1 - 0.8113 = 0.1887

Consider attribute B.
Info(T_0) = -1 log 1 - 0 log 0 = 0
Info(T_1) = -1/3 log 1/3 - 2/3 log 2/3 = 0.9183
Info(B, T) = 1/8 Info(T_0) + 7/8 Info(T_1) = 0.8035
Gain(b, T) = Info(T) - Info(B, T) = 1 - 0.8035 = 0.1965

Here, Gain(A, T) < Gain(B, T)

Under ID3, Imp-ID3(A) = Gain(A, T)
and Imp-ID3(B) = Gain(B, T).
Thus, we have "Imp-ID3(A) < Imp-ID3(B)".

Consider CART.
Info(T) = 1 - $(1/2)^2$ - $(1/2)^2$ = 1/2

Consider attribute A.
Info(T_0) = 1 - $(3/4)^2$ - $(1/4)^2$ = 0.375
Info(T_1) = 1 - $(3/4)^2$ - $(1/4)^2$ = 0.375
Info(A, T) = 1/2 Info(T_0) + 1/2 Info(T_1) = 0.375
Gain(A, T) = Info(T) - Info(A, T) = 1/2 - 0.375 = 0.125

Consider attirbute B.
Info(T_0) = 1 - 1^2 - 0^2 = 0
Info(T_1) = 1 - $(1/3)^2$ - $(2/3)^2$ = 0.444
Info(B, T) = 1/8 Info(T_0) + 7/8 Info(T_1) = 0.3885
Gain(B, T) = Info(T) - Info(B, T) = 1/2 - 0.3885 = 0.1115

Here, Gain(A, T) > Gain(B, T).

Under CART, Imp-CART(A) = Gain(A, T)
and Imp-CART(B) = Gain(B, T).
Thus, we have "Imp-CART(A) > Imp-CART(B)".

In conclusion, it is possible that
"Imp-CART(A) > Imp-CART(B)" but "Imp-ID3(A) < Imp-ID3(B)".

Q7 (Continued)

Q8 (20 Marks)

(a) (i)

Yes.

For each part p, we count the total number of records in the answer of Q3 and insert a record (p, C) into the answer of Q4.

(ii)

No

We need one additional kind of information, the answer of Q3, in addition to the answer of Q5.

The total access cost is 2GB only (the minimum access cost).

For each part p, we do the following.

we initialize the SUM variable to be 0.

we initialize the TOTAL variable to be 0.

For each combination of part and customer c where the part is equal to p,

we obtain the average price A for (c, p) from the answer of Q5

and the total number of records in T for (c, p) from the answer of Q3.

we increment SUM by A x C.

we increment TOTAL by C.

we construct a record (p, AVG) where AVG = SUM/TOTAL

we insert this record into the answer of Q6.

(b)(i)

drilldown

(ii)

rollup

Q8 (Continued)

(c) (i)

Yes.

No. of records with the actual target attribute value “Yes” and the predicted target attribute value “Yes”
= 20 x 60/100
= 12

Recall
= 12/30
= 40%

(ii)

Yes.

No. of records with the actual target attribute value “Yes” and the predicted target attribute value “Yes”
= 30 x 50/100
= 15

No. of records with the actual target attribute value “No” and the predicted target attribute value “Yes”
= 20 -15
= 5

No. of records with the actual target attribute value “No” and the predicted target attribute value “No”
= 100 – 5 – 30
= 65

Accuracy
= (15+65)/100
= 80%

Q9 (20 Marks)

(a)

$$r_n = M\, r_0$$

$$\begin{pmatrix} r_{n,1} \\ r_{n,2} \\ r_{n,3} \end{pmatrix} = \begin{pmatrix} r_{0,1} \\ r_{0,2} \\ r_{0,3} \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

$$= \begin{pmatrix} m_{11}r_{0,1} + m_{12}r_{0,2} + m_{13}r_{0,3} \\ m_{21}r_{0,1} + m_{22}r_{0,2} + m_{23}r_{0,3} \\ m_{31}r_{0,1} + m_{32}r_{0,2} + m_{33}r_{0,3} \end{pmatrix}$$

Sum of the values in r_n

$= m_{11}r_{0,1} + m_{12}r_{0,2} + m_{13}r_{0,3} + m_{21}r_{0,1} + m_{22}r_{0,2} + m_{23}r_{0,3} + m_{31}r_{0,1} + m_{32}r_{0,2} + m_{33}r_{0,3}$

$= (m_{11} + m_{21} + m_{31})r_{0,1} + (m_{12} + m_{22} + m_{32})r_{0,2} + (m_{13} + m_{23} + m_{33})r_{0,3}$

$= 1 \cdot r_{0,1} + 1 \cdot r_{0,2} + 1 \cdot r_{0,3}$

$= r_{0,1} + r_{0,2} + r_{0,3}$

= sum of the values in r_0

(b) (i)

$$X = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

(ii)

$$Y = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Part B (Compulsory Multiple-Choice (MC) Questions)

Note: For each question in this part, you just need to **circle** one of the five possible choices (i.e., A, B, C, D or E). The total scores in this part are 20 scores. Each question in this part weighs 5 scores.

Question	Answer
Q10	A / B / C / D / E
Q11	A / B / C / D / E
Q12	A / B / C / D / E
Q13	A / B / C / D / E

[There are no answers for this part.]

Part C (Bonus Question)

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

Q14 (20 Additional Marks)

(a)

Yes. The Apriori Algorithm can be adapted.
The idea is similar to the original Apriori Algorithm learnt in class.
However, we use the profit of an itemset in the adapted algorithm instead of the total number of rows containing the itemset in the dataset.
The reason why we can follow the framework if the Apriori Algorithm is that we have he following property.

- if an itemset S has profit >= 50, then any proper subset of S has profit >=50. (Or if an itemset S does not have profit >= 50, then any proper superset of S must not have profit >=50.)

Algorithm:

1. $L_1 \leftarrow$ All items with profit >= 50
2. $k \leftarrow 2$
3. while $L_{k-1} \neq \varnothing$
   - $C_k \leftarrow$ Generate candidates from L_1 by Join Step and Prune Step discussed in class
   - Perform a counting step on C_k and obtain L_k
4. Output $\bigcup_i L_i$

For example,

	frequency	profit
A	0+3+0+1+6=10	10*10=100
B	0+4+0+0+0=4	4*10=40
C	3+0+1+3+0=7	7*10=70
D	2+0+3+5+0=10	10*10=100

So, L_1={A,C,D}, C_2={AC,AD,CD}

	frequency	profit
AC	0+0+0+1+0=1	1*10=10
AD	0+0+0+1+0=1	1*10=10
CD	2+0+1+3+0=6	6*10=60

L_2 = {CD}
Output = {A, C, D, CD}

Q14 (Continued)

(b)

No. The Apriori Algorithm cannot be adapted. In this problem, the following Apriori property cannot be satisfied.

- If an itemset S has profit >=50, then any proper subset of S has profit >=50. (Or if an itemset S does not have profit >=50, then any proper superset of S must not have profit >=50.)

The following shows an example that this property cannot be satisfied.

	frequency	profit
CD	2+0+1+3+0=6	6*(6+4)=60
C	3+0+1+3+0=7	7*6=42

In the above example, CD has profit >=50, but a proper subset of CD (e.g., C) has profit < 42.

Here, we have another algorithm for reference.

1. $O \leftarrow \varnothing$
2. For each possible itemset S with frequency >=1,
   - Find the profit of S
   - If the profit of S >=50,

     $O \leftarrow O \cup \{S\}$
3. Return O.

Q14 (Continued)

End of Answer Sheet