COMP1942 Exploring and Visualizing Data (Spring Semester 2019)
Midterm Examination (Question Paper)
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 the **answer sheet**.
(2) You can use a calculator.

Question Paper

Q1 (20 Marks)

We are given a set of transactions with a certain number of items.
Consider association rule mining where the confidence threshold is 50% and the support threshold is 4.
Let S_0 be the set of all possible association rules with their confidence at least 50% and their support at least 4. In other words, S_0 is the set of all association rules we want to find.

In the class, we learnt the following two-step method of generating a set of association rules.
The description of the two-step method is given as follows.

- **Step 1:** To generate a set S_1 of all itemsets with their support at least 4
- **Step 2:** For any two itemsets in S_1, namely X and Y, where $X \subseteq Y$,

  if $\text{supp}(Y)/\text{supp}(X) \geq 50\%$, …………………………………..(*)

  generate an association rule in the form of "X → Y – X"

  Let S_2 be the set of all association rules generated in this step.

In the class, we study the following two claims.

- **Claim 1**: Each association rule in S_2 has its support at least 4.
- **Claim 2**: For each association rule in S_0, it is in S_2.

(a) Consider that the original Step 1 of the two-step method is changed to

"To generate a set S_1 of all itemsets with their support at least 3"

(i.e., number "4" is changed to number "3" in Step 1).

(i) Is it always true that Claim 1 is correct? If the answer is "yes", please show the correctness of the following "simplified" form of Claim 1 where B and C are two items (similar to the form shown in the class):

If "B→C" is in S_2, the support of "B→C" is at least 4.

If the answer is "no", please give a concrete example containing the *smallest* possible number of transactions and illustrate with this example that "B→C" is in S_2 but the support of "B→C" is smaller than 4.

(ii) Is it always true that Claim 2 is correct? If the answer is "yes", please show the correctness of the following "simplified" form of Claim 2 where B and C are two items (similar to the form shown in the class):

If "B→C" is in S_0, it is in S_2.

If the answer is "no", please give a concrete example containing the *smallest* possible number of transactions and illustrate with this example that "B→C" is in S_0 but "B→C" is not in S_2.

(b) This part is independent of Part (a) (i.e., in this part, we did not change any component in Step 1). Consider that Condition (*) in the original Step 2 of the two-step method is changed to

"if $\text{supp}(Y)/\text{supp}(X) \geq 60\%$"

(i.e., number "50%" is changed to number "60%" in Condition (*) in Step 2).

(i) Is it always true that Claim 1 is correct? Please elaborate it following the instruction in (a)(i).

(ii) Is it always true that Claim 2 is correct? Please elaborate it following the instruction in (a)(ii).

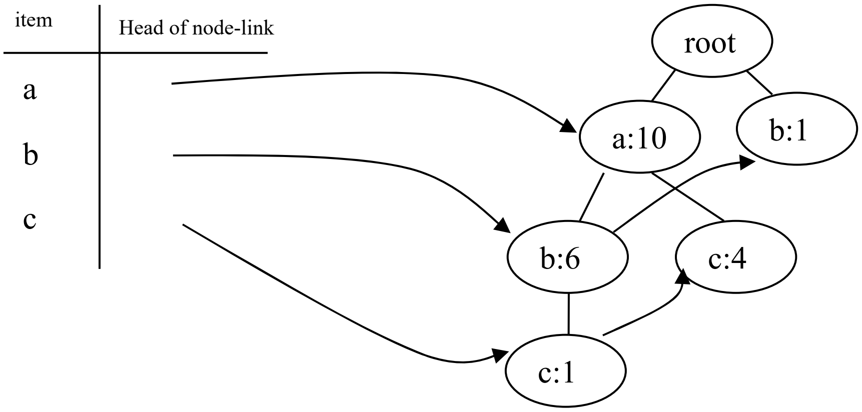
Q2 (20 Marks)

(a) Given the following transactions, and the support threshold = 2. We want to find all large itemsets.

A	B	C	D	E
1	0	0	1	0
1	0	0	1	1
0	0	1	0	0
1	0	1	1	1
1	0	1	0	1

Follow the steps of the Apriori Algorithm and deduce L_1, C_2, L_2, C_3, L_3, C_4, … until all the large itemsets are discovered. Finally, show all large itemsets.

(b) The following shows an FP-tree which is constructed from a set of transactions. Let the support threshold be 4.



(i) Please draw the conditional FP-tree on c.
(ii) Please draw the conditional FP-tree on b.
(iii)Please draw the conditional FP-tree on a.
(iv)Please list all frequent itemsets. You do not need to give the frequency of each frequent itemset.

Q3 [20 Marks]

The following shows a history of people each with age, gender and an attribute called "MMR_Vaccine" indicating whether this person has been injected with the MMR vaccine before. We also indicate whether they had measles or not in the last column. The first column "No." is just for you to refer the record number only and you do not need to use this column for generating the classifier.

No.	Age	Gender	MMR_Vaccine	Has _Measles
1	young	male	no	yes
2	old	male	no	yes
3	young	female	yes	yes
4	old	female	no	yes
5	young	female	yes	no
6	young	female	yes	no
7	old	female	yes	no
8	old	female	yes	no

(a) We want to train a CART decision tree classifier to predict whether a new person will have measles or not. We define the value of attribute Has_Measles to be the *label* of a record.

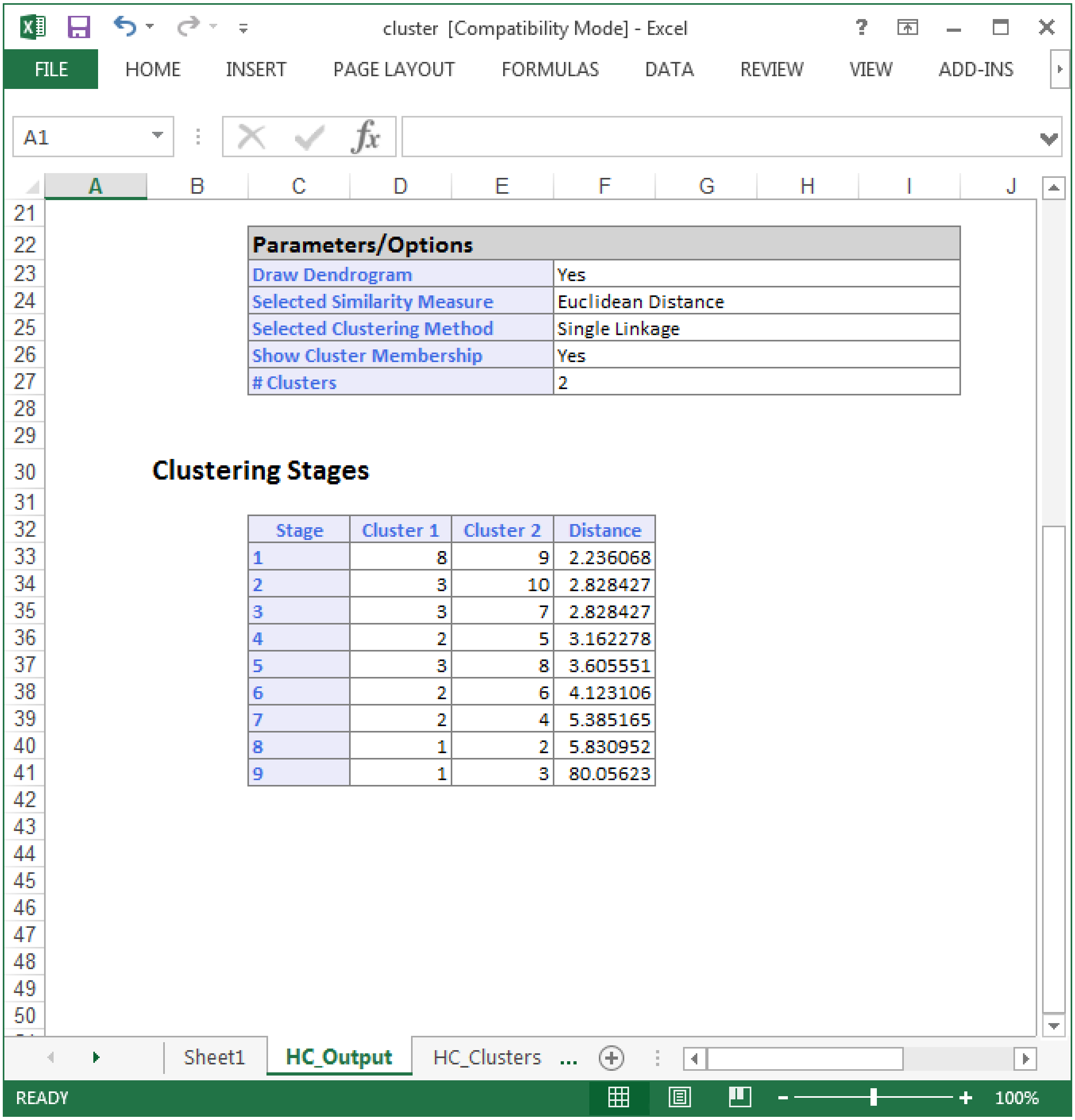
(i) Please find a CART decision tree according to the above example. In the decision tree, whenever we process a node containing at least 75% records with the same label, we stop to process this node for splitting.

(ii) Consider a new young male person who has been injected with the MMR vaccine before. Please predict whether this new person will have measles or not.

(b) What is the difference between the C4.5 decision tree and the ID3 decision tree? Why is there a difference?

Q4 [20 Marks]

(a) Consider a dataset with 2 attributes. This dataset contains 10 data points, namely x_1, x_2, …, x_{10}. Suppose that we obtained the following output for XLMiner for hierarchical clustering using the single linkage as a distance measurement between 2 clusters on this dataset. In XLMiner, we specified "# Clusters" (i.e., the number of clusters to be found) as 2. According to the following output **only**, is it possible to know all data points in each of the 2 clusters? If yes, please write down "Yes" and list out all data points in each of the 2 clusters. Otherwise, please write down "No" and explain why we could not know all data points in each of the 2 clusters.



Parameters/Options	
Draw Dendrogram	Yes
Selected Similarity Measure	Euclidean Distance
Selected Clustering Method	Single Linkage
Show Cluster Membership	Yes
# Clusters	2

Clustering Stages

Stage	Cluster 1	Cluster 2	Distance
1	8	9	2.236068
2	3	10	2.828427
3	3	7	2.828427
4	2	5	3.162278
5	3	8	3.605551
6	2	6	4.123106
7	2	4	5.385165
8	1	2	5.830952
9	1	3	80.05623

(b) Consider the scenario in Part (a). Please answer the following questions.

(i) Consider this dataset only. Suppose that we want to find 4 clusters (instead of 2 clusters specified in the input of XLMiner). According to the above output **only**, is it possible to know all data points in each of the 4 clusters? If yes, please write down "Yes" and list out all data points in each of the 4 clusters. Otherwise, please write down "No" and explain why we could not know all data points in each of the 4 clusters.

(ii) This part is independent of Part (b)(i). Consider this dataset only. According to **only** the above output that was originally generated based on the single linkage as the distance measurement, is it possible to find all points of each of the 2 clusters when the distance measurement used is the complete linkage instead of the single linkage? If yes, please write down "Yes" and list out all data points in each of the 2 clusters. Otherwise, please write down "No" and explain why we could not know all data points in each of the 2 clusters.

(iii) This part is independent of Part (b)(i) and (ii). Consider this dataset only. According to **only** the above output that was originally generated based on the single linkage as the distance measurement, is it possible to draw the dendrogram when the distance measurement between 2 clusters is the cendroid linkage instead of the single linkage? If yes, please write down "Yes" and draw the dendrogram. Otherwise, please write down "No" and give the reason why we could not draw the dendrogram.

Q5 (20 Marks)

(a) Consider Algorithm forgetful sequential k-means clustering. Let a be a constant defined in this algorithm.

(i) Please write down the steps for Algorithm forgetful sequential k-means clustering.

(ii) Consider a cluster found in the algorithm containing n examples where its initial mean is equal to m_0. Let x_j be the first j-th example in this cluster and m_j be the mean vector of this cluster after the first j-th examples are added for j = 1, 2, …, n. We can express m_n in the following form.

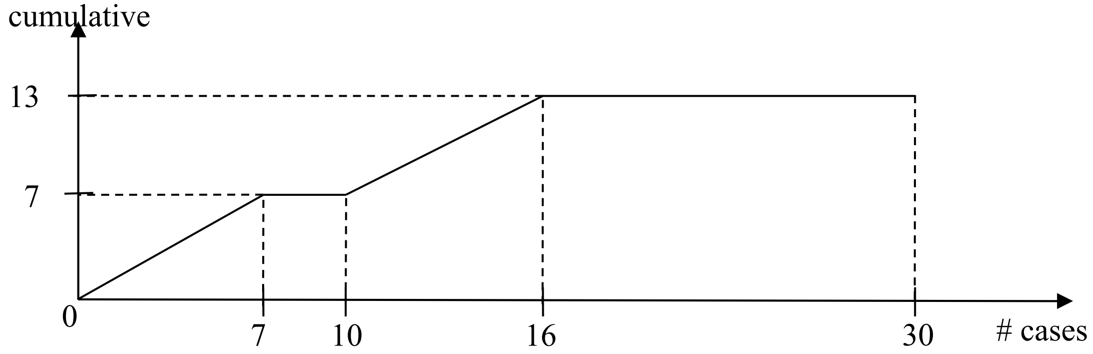
$$m_n = X \cdot m_0 + \sum_{p=1}^{n} Y \cdot x_p$$

where X and Y are some expressions.

Please show that m_n can be expressed in this form. After you show this statement, please also write down what is X and what is Y.

(You are not required to memorize the formula for this question. You just need to show how you obtain the above expression and finally you can obtain X and Y.)

(b) Consider a dataset containing 4 input attributes and 1 target attribute where the target attribute contains 2 possible values. We are given the following lift chart based on a classifier constructed from this dataset.



Is it possible to find the decile-wise lift chart? If yes, please give the chart. Otherwise, please explain it.

End of Paper

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 $supp(\{B, C\}) \geq 3$ and $supp(\{B\})\geq3$

In Step 2, $B\rightarrow C$ is generated since $supp(\{B,C\})/supp(\{B\})=100\% \geq50\%$

Thus, $B\rightarrow C$ is in S_2

Note that

$$\begin{aligned} supp(B\rightarrow C) &=supp(\{B,C\}) \\ &= 3 \\ &< 4 \end{aligned}$$

In conclusion, $B\rightarrow C$ is in S_2 but $supp(B\rightarrow C) < 4$

Q1 (continued)

(a) (ii)

Yes.

Since $B \rightarrow C$ is in S_0,
$\quad$ $\text{conf}(B \rightarrow C) \geq 50\%$
$\quad$ $\text{supp}(\{B,C\})/\text{supp}(\{B\}) \geq 50\%$
Since $B \rightarrow C$ is in S_0,
$\quad$ $\text{supp}(B \rightarrow C) \geq 4$
Since $\text{supp}(\{B,C\}) = \text{supp}(B \rightarrow C)$,
$\quad$ $\text{supp}(\{B,C\}) \geq 4$
Thus, $\{B, C\}$ is in S_1.
Since $\text{supp}(\{B, C\}) \geq 4$,
$\quad$ $\text{supp}(\{B\}) \geq 4$
Thus, $\{B\}$ is in S_1

Since $\{B\}$ is in S_1,
and $\{B, C\}$ is in S_1,
$\quad$ Step 2 must consider
$\quad\quad$ $\{B\}$ and $\{B, C\}$ together, and
$\quad$ generate $B \rightarrow C$ (since $\text{supp}(\{B,C\})/\text{supp}(\{B\}) \geq 50\%$)
$B \rightarrow C$ 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
$\text{supp}(\{B, C\}) \geq 4$ and
$\text{supp}(\{B\}) \geq 4$

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

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)

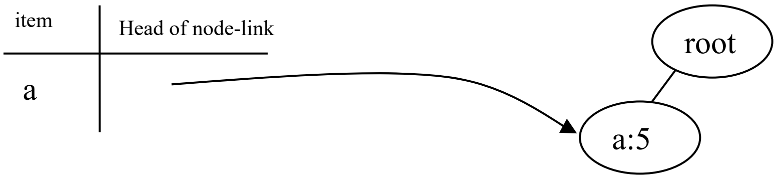
L_1 = {{A}, {C}, {D}, {E}}
Large 2-itemset Generation:
    Join Step/Prune Step
        C_2 = {{A, C}, {A, D}, {A, E}, {C, D}, {C, E}, {D, E}}
    Counting Step
        L_2 = {{A, C}, {A, D}, {A, E}, {C, E}, {D, E}}
Large 3-itemset Generation:
    Join Step
        C_3 = {{A, C, D}, {A, C, E}, {A, D, 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}}

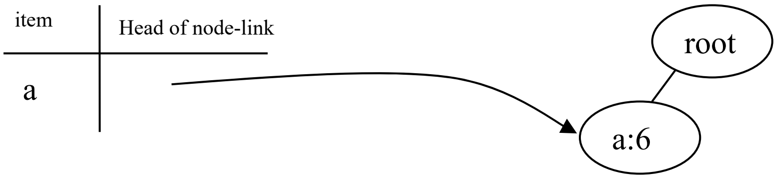
Q2 (Continued)

Q2 (Continued)

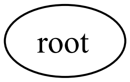
(b)(i)



(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,

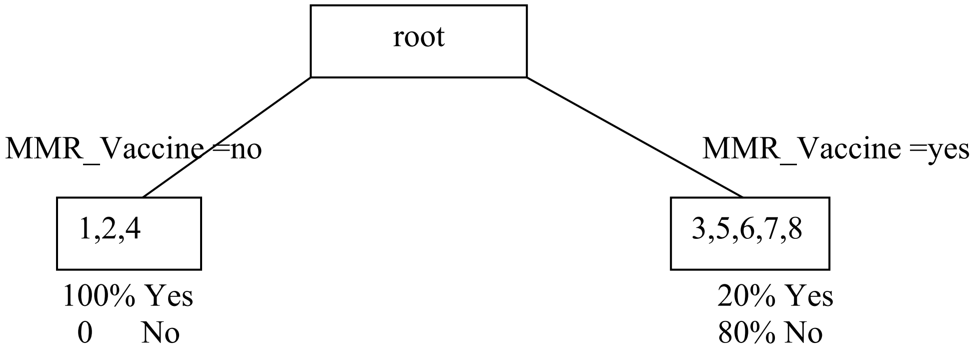
Info(T_{no})= $1-1^2-0^2=0$

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)

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

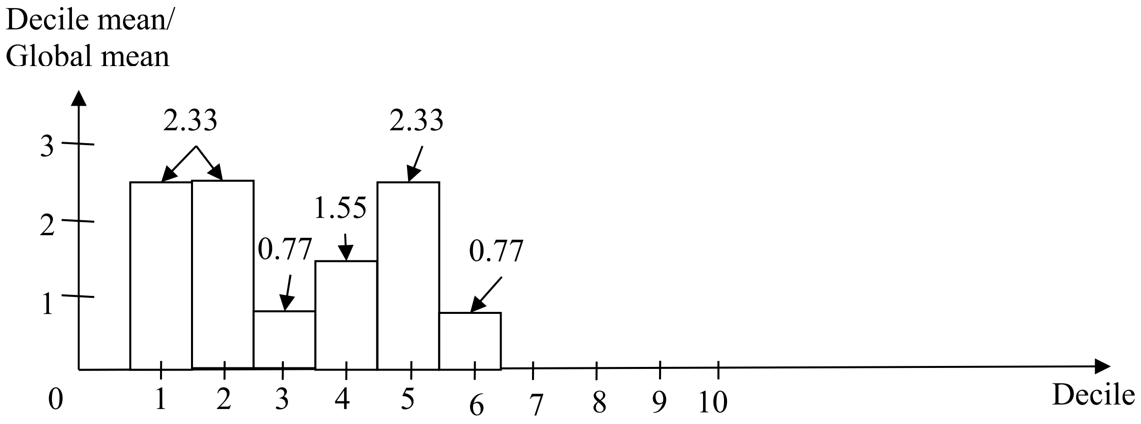
$X = (1-a)^n$

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

Q5 (Continued)

(b)

Yes. The chart is shown as follows.



End of Answer Sheet