COMP1942 Exploring and Visualizing Data (Spring Semester 2021)
Online Midterm Examination (Question Paper)
Date: 7 April, 2021 (Wednesday)
Time: 10:35am-11:40am
Duration: 1 hour 5 minutes

Instructions:

(1) Guideline

(a) Please follow **all** instructions about the exam guideline (e.g., your face video capturing) stated in the Canvas website.

(b) For the sake of space, we do not write them again.

(2) Question

(a) There are 3 parts in this exam, Part A (Short/Long Question), Part B (Multiple-Choice Question) and Part C (Bonus Question).

(b) Please answer **all** questions in Part A and Part B. The total scores in this exam are 100.
Part C is optional and the question in Part C has 10 additional scores.

(c) If the sum of all scores is greater than 100, the final score will be truncated to 100.

(3) Answer Sheet

(a) Please submit your answers in PDF to the Canvas website.

(b) Please use the cover page stated in the Canvas website as the **first** page of your PDF file. This cover page includes your information and an agreement with your signature.

(c) Please start to write your answers starting on the **second** page of your PDF file.

(d) The PDF file should “clearly” show your answers without any blurred images. No marks will be given to any “blurred” parts in the PDF file. Please make sure that the PDF file shows your answers clearly.

(4) Online Exam

(a) This is an online exam where you could access all online materials.
However, it is **not** allowed to communicate with other people (except the instructor and the tutors in this course) in any form (including but not limited to orally, electronically and in writing) during the entire exam period.

(5) File Submission

(a) We allow a 15-minute buffer for your PDF file upload. Remember to upload your file at around 11:40am. We allow your file uploading time at most 15 minutes. Canvas will terminate any file uploading process at 11:55am if your file is still being uploaded at 11:55am.

(6) Zero-Score Regulation

(a) If your face could not be shown in your video for at least 10 seconds in the exam period, your exam score will be set to 0 (even though you submit your PDF file in Canvas).

(b) If you do not submit the first cover page which is filled and signed completely, your exam score will be set to 0.

(c) We only mark your latest PDF file uploaded by 11:55am. Your exam score will be set to 0 if we could not see any PDF file uploaded by 11:55am (even though you do the question paper or you “could” upload your PDF file after 11:55am).

Part A (Short/Long Question)

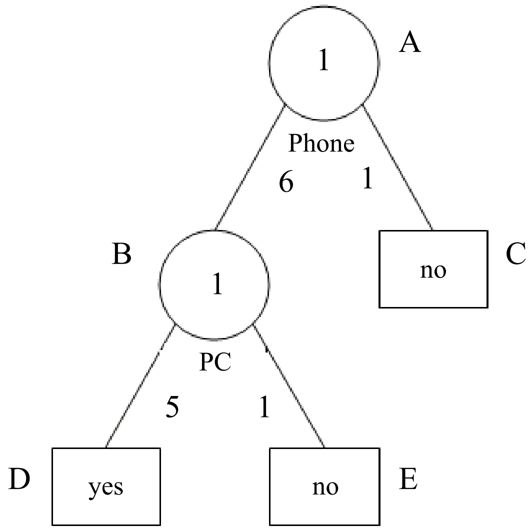
In this part, there are 3 short/long questions, namely Q1, Q2 and Q3. The total scores in this part are 50 scores (out of 100).

Q1 (15 Marks)

(a) Consider a set of transactions containing 5 items, namely V, W, X, Y and Z.
Suppose that the support threshold is set to 2 and the confidence threshold is set to 75% .
We found that the set of large itemsets is
{ {V}:4, {X}:3, {Y}:3, {Z}:3,
{V, X}:2, {V, Y}:3, {V, Z}:3, {X, Z}:2, {Y, Z}:2,
{V, X, Z}:2, {V, Y, Z}:2 }

Please write down a set of all possible interesting association rules each in the form of "X → Y" where X is an itemset and Y is an item. You are required to write down the confidence of each association rule in the set.

(b) This part is not related to part (a). We are given the following decision tree generated by XLMiner. There are five nodes in this tree, namely A, B, C, D and E. In the diagram below, "Phone" corresponds to the number of phones and "PC" corresponds to the number of PCs.



(i) Which nodes are terminal nodes?
(ii) Which nodes are decision nodes?
(iii) There is a number, "6", next to the line between node A and node B. What is the physical meaning of this number?
(iv) Suppose that there is a new record with "Phone" = 1 and "PC"=1. What is the predicted value of the target attribute of this new record according to this decision tree?

Q2 (20 Marks)

Given the following transactions and the support threshold = 2.

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

Follow the steps of the FP-growth algoritm to find all frequent itemsets. Please show the steps and list all frequent itemsets. Note that you are required to draw the original FP-tree and all conditional FP-tress involved.

Q3 (15 Marks)

(a) (i) Please give the formula for the Bayes Rule.
(ii) Please give the reason why we need to use Bayes Rule.

(b) Consider Algorithm sequential k-means clustering.
When it reads a data point x, it will update the mean m of a cluster with the following operation.

$$m \leftarrow m + 1/n\ (x - m)$$

where n is the size of the cluster including the new data point x.

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

(ii) Please prove that, with the above operation, the mean m is calculated correctly. That is, the mean m calculated is equal to the expected vector among all data points in the cluster.
(Hints: Let x_j be the j-th example in cluster i and $m_i(t)$ be the mean vector of cluster i containing t examples. Consider that x is the t-th example in cluster i. Note that $m_i(t) = \frac{x_1 + x_2 + ... + x_{t-1} + x_t}{t}$.)

Part B (Multiple-Choice Question)

In this part, there are 10 multiple-choice questions, namely Q4-Q13. The total scores in this part are 50 scores (out of 100). Each question weighs 5 scores (out of 100). In your answer sheet, please write down the following table on one of the pages of your PDF submission. In the corresponding cell, write down the answer for each question.

Note: Please write the letter **clearly** (i.e., P, Q, R, S or T) for each answer so that it could be distinguished from other letters **easily**. In the past, some students wrote the letter unclearly which look like two possible letters. One example is that the hand-written letter "P" (from some students) is similar to the hand-written letter "T". There are more examples which are not included here. In any case, if your letter is judged by us that it is unclear, even though you "thought" that your answer is correct, 0 score will be given to you for that question.

Part B

Question	**Your Answer**
Q4	
Q5	
Q6	
Q7	
Q8	
Q9	
Q10	
Q11	
Q12	
Q13	

Q4. Which of the following statement(s) is/are true? Suppose that the support threshold is set to 10.

(1) In general, the total number of nodes in the FP-tree constructed according to the default item ordering (i.e., the item ordering in the descending order of item frequencies) is smaller than that according to the reverse item ordering (i.e., the item ordering in the ascending order of item frequencies).

(2) The height of an FP-tree is exactly equal to the maximum number of frequent items in a transaction.

(3) We must construct the entire dataset according to the FP-tree.

P. Statements (1) and (2) only
Q. Statements (1) and (3) only
R. Statements (2) and (3) only
S. Statements (1), (2) and (3)
T. None of the above choices

Q5. Which of the following statement(s) is/are true? Consider the Apriori algorithm.

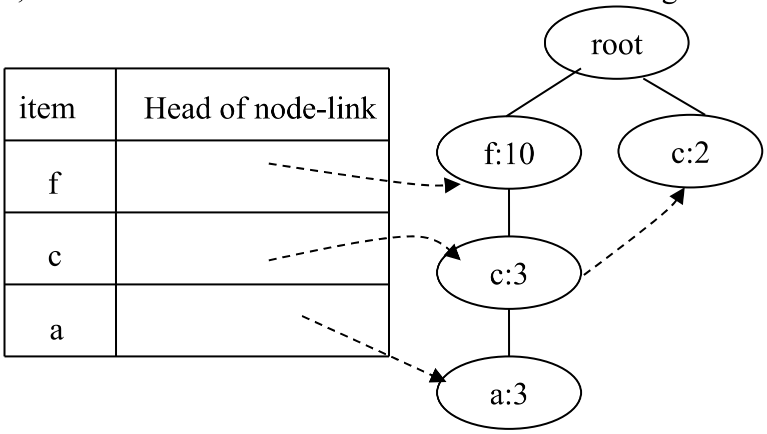
(1) It is possible that an itemset in the candidate set generated just after the prune step of the algorithm is not large.

(2) There is no need to read the original dataset in the prune step of the algorithm.

(3) There is no need to read the original dataset in the counting step of the algorithm.

P. Statements (1) and (2) only
Q. Statements (1) and (3) only
R. Statements (2) and (3) only
S. Statements (1), (2) and (3)
T. None of the above choices

Q6. Suppose that each transaction contains at most one occurrence of any item. We set the support threshold to 1. We are given the following FP-tree generated from a dataset. According to all the information we have, what is the total number of transactions in the original dataset?



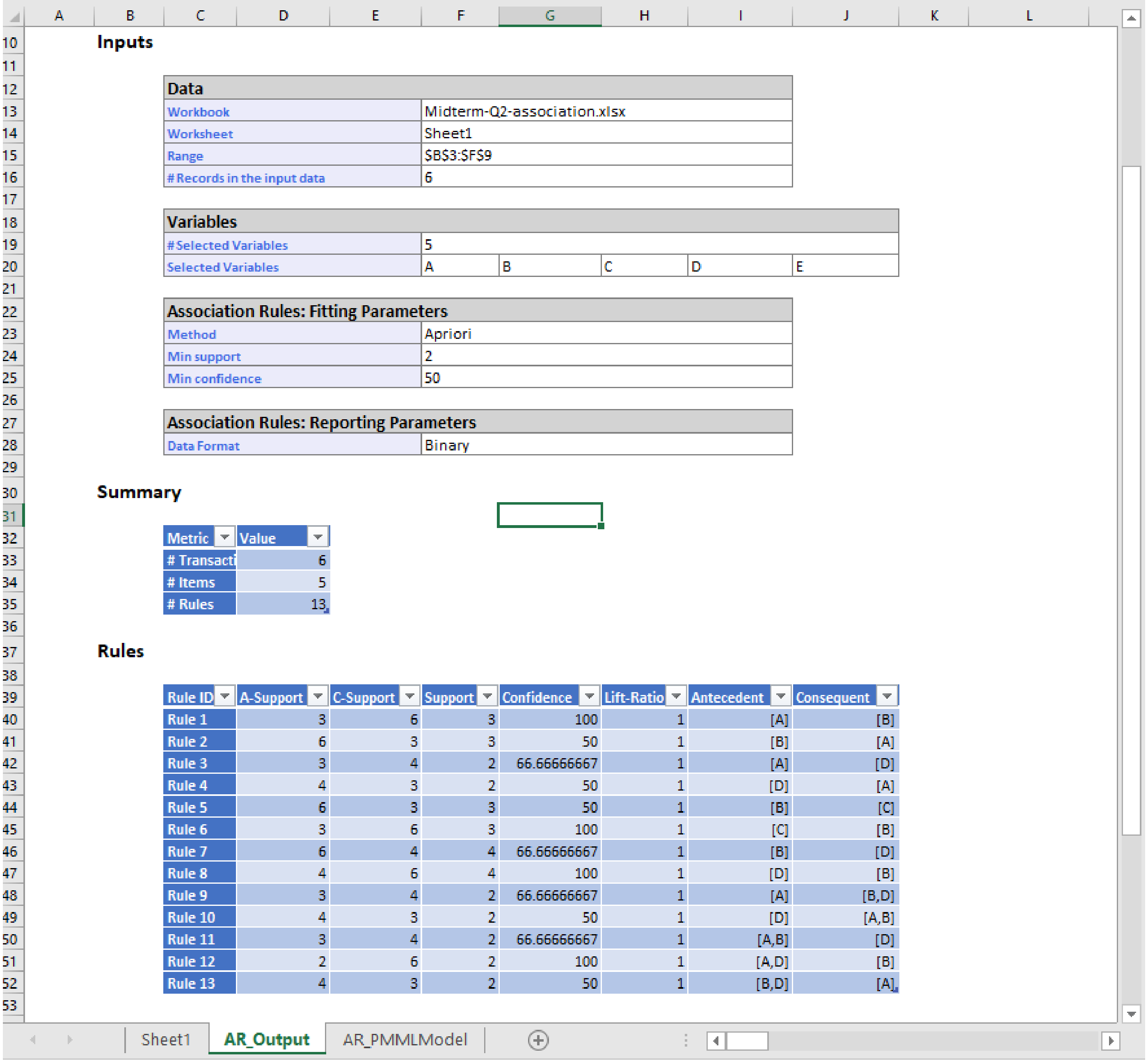
P. 5
Q. 12
R. 15
S. 18
T. None of the above choices

Q7. Which of the following statement(s) is/are true?

(1) If we want to find 2 clusters, dendrogram could be used to find all data points in each of 2 clusters.
(2) The Chi-square measure is used to measure the correlation between 2 attributes.
(3) Consider a given probability distribution A on two possible values (e.g., the values in the target attribute) and another given probability distribution B on the same values. If the entropy of distribution A (used in ID3) is greater than the entropy of distribution B, the Gini index of distribution A (used in CART) is greater than the Gini index of distribution B.

P. Statement (1) only
Q. Statement (3) only
R. Statements (1) and (2) only
S. Statements (1), (2) and (3).
T. None of the above choices

Q8. The following shows the output of the XLMiner result for association rule mining.



Inputs

Data	
Workbook	Midterm-Q2-association.xlsx
Worksheet	Sheet1
Range	$B$3:$F$9
# Records in the input data	6

Variables					
# Selected Variables	5				
Selected Variables	A	B	C	D	E

Association Rules: Fitting Parameters	
Method	Apriori
Min support	2
Min confidence	50

Association Rules: Reporting Parameters	
Data Format	Binary

Summary

Metric	Value
# Transacti	6
# Items	5
# Rules	13

Rules

Rule ID	A-Support	C-Support	Support	Confidence	Lift-Ratio	Antecedent	Consequent
Rule 1	3	6	3	100	1	[A]	[B]
Rule 2	6	3	3	50	1	[B]	[A]
Rule 3	3	4	2	66.66666667	1	[A]	[D]
Rule 4	4	3	2	50	1	[D]	[A]
Rule 5	6	3	3	50	1	[B]	[C]
Rule 6	3	6	3	100	1	[C]	[B]
Rule 7	6	4	4	66.66666667	1	[B]	[D]
Rule 8	4	6	4	100	1	[D]	[B]
Rule 9	3	4	2	66.66666667	1	[A]	[B,D]
Rule 10	4	3	2	50	1	[D]	[A,B]
Rule 11	3	4	2	66.66666667	1	[A,B]	[D]
Rule 12	2	6	2	100	1	[A,D]	[B]
Rule 13	4	3	2	50	1	[B,D]	[A]

Which of the following statement(s) is/are true?

(1) If the support threshold is set to 2 and the confidence threshold is set to 70, the total number of interesting association rules is 4.

(2) The support of item E must be equal to 1 or 0 in this dataset.

(3) If the support threshold is set to 3 and the confidence threshold is set to 50, the total number of interesting association rules is 6.

P. Statement (1) only

Q. Statement (3) only

R. Statements (1) and (3) only

S. Statements (1), (2) and (3).

T. None of the above choices

Q9. Which of the following statement(s) is/are true?

(1) In any dataset, the confidence of an association rule in the form of "A→B" (where A and B are items) is exactly equal to the confidence of an association rule in the form of "B→A".

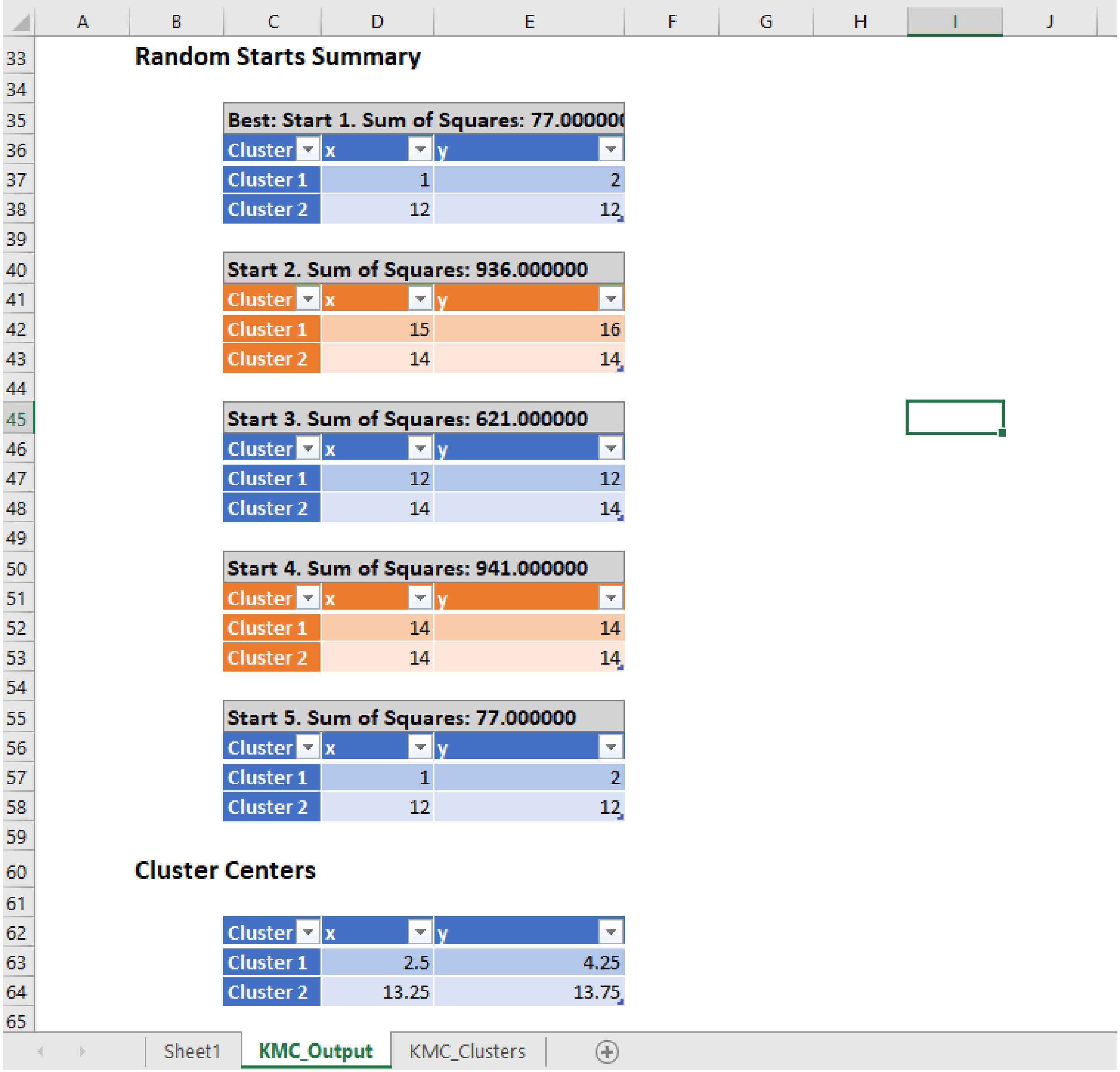
(2)In any dataset, the lift ratio of an association rule in the form of "A→B" (where A and B are items) is exactly equal to the lift ratio of an association rule in the form of "B→A".

(3)In any dataset, the support of an association rule in the form of "A→B" (where A and B are items) is exactly equal to the support of an association rule in the form of "B→A".

P. Statement (1) only
Q. Statement (2) only
R. Statement (3) only
S. Statements (2) and (3) only
T. None of the above choices

Q10. Which of the following statement(s) is/are true?

(1) In the k-nearest neighbor classifier, it is better to set the parameter k to an odd number in any dataset.
(2) In the k-means clustering, it is better to set the parameter k to an odd number in any dataset.
(3) According to the following XLMiner result of a k-means clustering method, we know that the final mean of the first cluster is (1, 2) and the final mean of the second cluster is (12, 12).



Random Starts Summary

Best: Start 1. Sum of Squares: 77.00000

Cluster	x	y
Cluster 1	1	2
Cluster 2	12	12

Start 2. Sum of Squares: 936.000000

Cluster	x	y
Cluster 1	15	16
Cluster 2	14	14

Start 3. Sum of Squares: 621.000000

Cluster	x	y
Cluster 1	12	12
Cluster 2	14	14

Start 4. Sum of Squares: 941.000000

Cluster	x	y
Cluster 1	14	14
Cluster 2	14	14

Start 5. Sum of Squares: 77.000000

Cluster	x	y
Cluster 1	1	2
Cluster 2	12	12

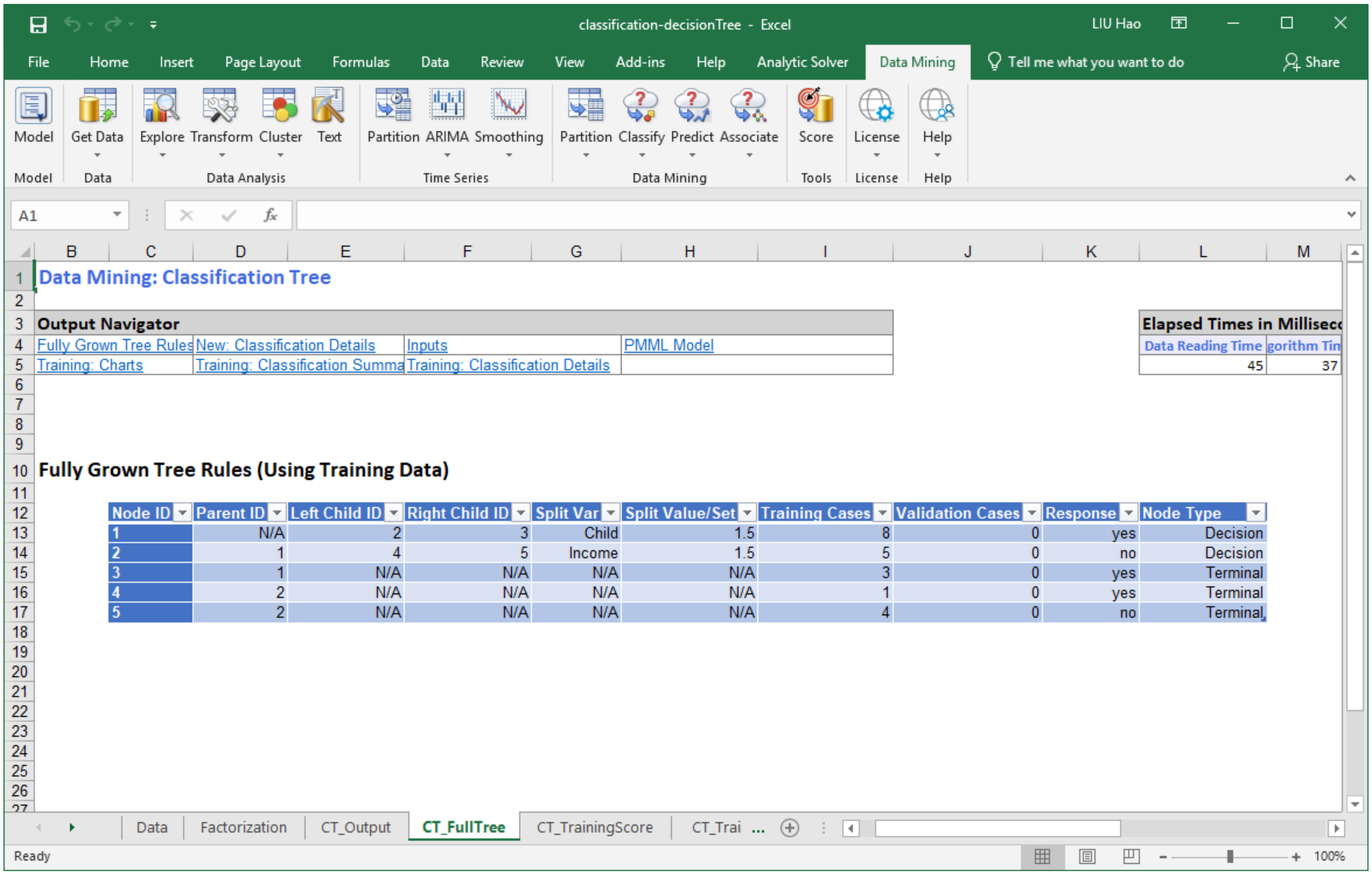
Cluster Centers

Cluster	x	y
Cluster 1	2.5	4.25
Cluster 2	13.25	13.75

P. Statement (2) only
Q. Statement (3) only
R. Statements (1) and (3) only
S. Statements (1), (2) and (3) only
T. None of the above choices

Q11. Consider the decision tree. Which of the following statement(s) is/are true?

(1) The entropy value must be greater than 0.
(2) The information gain of an attribute must be greater than 0.
(3) The following shows the XLMiner result of a decision tree based on a training set with all known input (numeric) attributes and the known target attribute. It is possible to draw the decision tree according to this XLMiner result such that the decision tree could be used to predict the target attribute of a record with all known input attributes.



Data Mining: Classification Tree

Output Navigator			
Fully Grown Tree Rules	New: Classification Details	Inputs	PMML Model
Training: Charts	Training: Classification Summa	Training: Classification Details	

Elapsed Times in Millisecc	
Data Reading Time	gorithm Tin
45	37

Fully Grown Tree Rules (Using Training Data)

Node ID	Parent ID	Left Child ID	Right Child ID	Split Var	Split Value/Set	Training Cases	Validation Cases	Response	Node Type
1	N/A	2	3	Child	1.5	8	0	yes	Decision
2	1	4	5	Income	1.5	5	0	no	Decision
3	1	N/A	N/A	N/A	N/A	3	0	yes	Terminal
4	2	N/A	N/A	N/A	N/A	1	0	yes	Terminal
5	2	N/A	N/A	N/A	N/A	4	0	no	Terminal

P. Statement (3) only
Q. Statements (1) and (2) only
R. Statements (2) and (3) only
S. Statements (1), (2) and (3) only
T. None of the above choices

Q12. The following shows a history of persons with attributes "Study_Medicine" (i.e., Studying Medicine), "Age" and "Income". We also indicate whether they get vaccinated against COVID19 or not in the last column.

Study_Medicine	Age	Income	Vaccinated
no	young	fair	yes
no	young	high	yes
yes	old	fair	yes
yes	middle	fair	yes
no	young	fair	no
no	middle	low	no
yes	old	low	no
yes	young	low	no

We want to train a C4.5 decision tree classifier to predict whether a new person will get vaccinated or not. We define the value of attribute "Vaccinated" to be the *label* of a record. In the decision tree, whenever we process (1) a node containing at least 80% records with the same label or (2) a node containing at most 2 records, we stop to process this node for splitting.

Which of the following statement(s) is/are true?

(1) In this decision tree, we first split all records according to attribute "Income" into a number of nodes and then split all records in one of the split nodes according to attribute "Age".
(2) Consider a new young person studying medicine whose income is fair. This decision tree predicts that this person will get vaccinated.
(3) There are 4 terminal nodes in this decision tree.

P. Statement (1) only
Q. Statement (2) only
R. Statement (3) only
S. Statements (2) and (3) only
T. None of the above choices

Q13. Consider eight data points.
The following matrix shows the pairwise distances between any two points.

$$\begin{array}{ccccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
1 & 0 & & & & & & & \\
2 & 11 & 0 & & & & & & \\
3 & 5 & 13 & 0 & & & & & \\
4 & 12 & 2 & 14 & 0 & & & & \\
5 & 7 & 17 & 1 & 18 & 0 & & & \\
6 & 13 & 4 & 15 & 5 & 20 & 0 & & \\
7 & 9 & 15 & 12 & 16 & 15 & 19 & 0 & \\
8 & 11 & 20 & 12 & 21 & 17 & 22 & 30 & 0
\end{array}$$

The agglomerative approach is used to group these points using the group average linkage as the distance measurement between 2 clusters. Which of the following statement(s) is/are true?

(1) Suppose that we want to obtain 4 clusters. Each of the two clusters has 1 data point only and each of the other two clusters has 3 data points.

(2) Suppose that we want to obtain 2 clusters. One cluster has 2 data points and the other has 6 data points.

(3) Suppose that we want to obtain 2 clusters. The distance between these 2 clusters is 19.

P. Statement (1) only
Q. Statement (2) only
R. Statement (3) only
S. Statements (1) and (3) only
T. None of the above choices

Part C (Bonus Question)

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

Q14 (10 Additional Marks)

In our lecture about frequent itemset mining, we learnt a property called the Apriori property: “Given an itemset I, if I is frequent, then any proper subset of I is frequent”. Then, we could use the Apriori algorithm

Let us consider another scenario. We are given n data points with d attributes, namely A_1, A_2, A_3, ..., A_d. Suppose that each attribute A_i is divided into a number of units/intervals, namely [1, 10], [11, 20], [21, 30], ... , [91, 100] for each i in [1, d]. Each data point has its value of attribute A_i to be within one of these intervals/units for each i in [1, d].

Given a set X of attributes, the grid structure with respect to X is defined to be a set of all units involved in each attribute in X. For example, if X = {A_2, A_4}, then the grid structure with respect to X is equal to {"A_2: [1, 10]", "A_2: [11, 20]", "A_2: [21, 30]", ..., "A_2: [91, 100]", "A_4: [1, 10]", "A_4: [11, 20]", "A_4: [21, 30]", ... "A_4: [91, 100]"}. A cell with respect to X is defined to be a set of |X| units such that for each attribute in X, there is exactly one unit involved for this attribute. (Note: |X| means the total number of elements in X.) For example, if X = {A_2, A_4}, one example of a cell with respect to X could be {"A_2: [11, 20]", "A_4: [61, 70]"}. The density of a cell with respect to X is defined to be P divided by n where P is the total number of data points such that for each attribute A_i in X, the value of A_i of each of these data points is within the interval/unit specified in the cell. Let Y be the set of all possible cells with respect to X. Let Z be the collection of the densities of all cells with respect to X in Y. X is said to have good information if the entropy of Z is at most 2.0.

Please show whether an Apriori algorithm could be adapted for finding a set of attribute sets each of which have good information. No scores will be given to those answers without justifications.

End of Paper