COMP1942 Exploring and Visualizing Data (Spring Semester 2020)
Online Midterm Examination (Question Paper)
Date: 20 April, 2020 (Monday)
Time: 12:05pm-1:10pm
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) COMP1942 Virtual Barn Log-in Period
- (a) Your allocated period of logging in "COMP1942 Virtual Barn" could be found in the Canvas webpage.
- (b) In the whole exam period, you could use XLMiner installed at your laptop/desktop/browser **at any time** but, you could use XLMiner installed in "COMP1942 Virtual Barn" **in the allocated period only**. If you log in "COMP1942 Virtual Barn" outside the allocated period in the exam period, 20 (out of 100) will be deducted from your exam score.

(3) Question
- (a) Please answer **all** questions.
- (b) There are 2 parts in this exam, Part A (Short/Long Question) and Part B (Multiple-Choice Question).

(4) 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.
- (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.

(5) 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.

(6) File Submission
- (a) We allow a 10-minute buffer for your PDF file upload. Remember to upload your file at around 1:10pm. We allow your file uploading time at most 10 minutes. Canvas will terminate any file uploading process at 1:20pm if your file is still being uploaded at 1:20pm.

(7) 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, your exam score will be set to 0.
- (c) We only mark your latest PDF file uploaded by 1:20pm. Your exam score will be set to 0 if we could not see any PDF file uploaded by 1:20pm (even though you do the question paper or you "could" upload your PDF file after 1:20pm).

XLMiner Question

In this exam, we have the following questions which need you to use XLMiner.

- Part A – Q1(a)
- Part B – Q5

Part A (Short/Long Question)

In this part, there are 4 short/long questions, namely Q1, Q2, Q3 and Q4. The total scores in this part are 60 scores. Each question weights 15 scores.

Q1 (15 Marks)

(a) [You could leave 0.5 A4-sized page for your answer sheet if you need to do this part later.]

You are given an Excel file called "Q1a.xlsx". This Excel file stores a number of records about students' exam scores in 5 different subjects, namely "Computer", "History", "English", "Chinese" and "Mathematics". Each score is in the range from 0 to 100. The following shows the header row in this Excel file.

Computer	History	English	Chinese	Mathematics
…	…	…	…	…

You should use XLMiner to find 3 clusters. Please use "No. of iterations = 50, random starts = 10, seed = 1942 without normalizing the input data" in XLMiner. You are required to write down the information about each final cluster (including the mean of the cluster and the total number of data points in this cluster) as an output of the algorithm. When you write down the mean of each cluster, please write it in the format of "(Computer=X_1, History=X_2, English=X_3, Chinese=X_4, Mathematics=X_5)" where X_i is a real number nearest to 3 decimal places for i = 1, 2, 3, 4, 5.

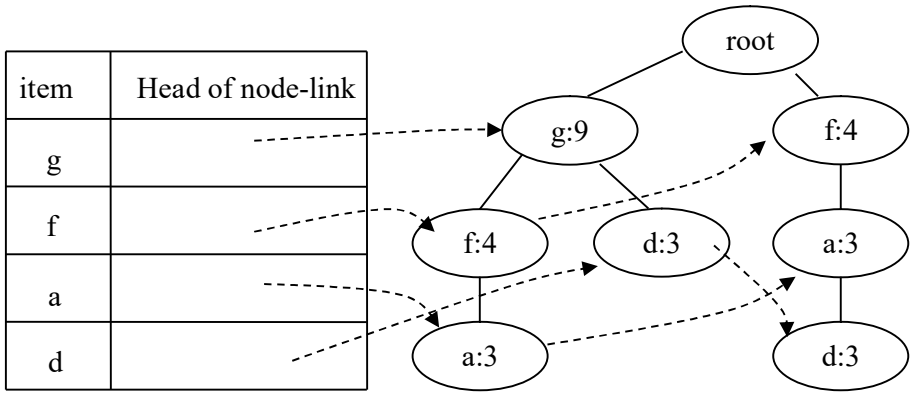
(b) We are given the following table. We have 8 students with 4 attributes, namely "Gender", "Major", "Mock Exam" and "Midterm". "Mock Exam" denotes whether the student attended the mock exam and "Midterm" denotes whether the student obtained a midterm score above the mean of the whole class or a midterm score below the mean of the whole class. (For simplicity, we ignore the case when the midterm score is exactly equal to the mean of the whole class. Thus, if a midterm score is NOT above mean, this denotes that it is below mean, and vice versa.).

Gender	Major	Mock Exam	Midterm
Male	COMP	Yes	AboveMean
Male	QSA	Yes	BelowMean
Female	QSA	Yes	AboveMean
Female	QSA	Yes	AboveMean
Male	COMP	No	BelowMean
Male	COMP	No	BelowMean
Female	COMP	No	BelowMean
Female	QSA	Yes	AboveMean

Using Naïve Bayesian Classifier, we want to predict whether a male student with his major "QSA" without taking the mock exam has a midterm score below the mean of the whole class or not. Please show your steps. Besides, please give the exact probability value of this prediction.

Q2 (15 Marks)

We are given a set S of transactions involving only items a, b, c, d, e, f, g. Suppose that the support threshold is set to 3. We use the FP-growth method learnt in the class to generate the following conditional FP-tree on {b} where the support of {b} is 13.



(a) Please use the FP-growth method to continue generate all possible conditional FP-trees starting from the above conditional FP-tree (i.e., all conditional FP-trees with the conditional itemset involving {b}). You just need to write down each conditional FP-tree together with its conditional itemset and the support of this conditional itemset. You don't need to show the steps.

(b) Suppose that we know the following additional information.

- The total number of transactions containing item a is 13.
- The total number of transactions containing item c is 2.
- The total number of transactions containing item d is 14.
- The total number of transactions containing item e is 1.
- The total number of transactions containing item f is 28.
- The total number of transactions containing item g is 19.
- Among all transactions containing item f but not containing item b, we know that **only** one of the following cases occurs:
  - f
  - f, g
  - In other words, we say that "among all transactions containing item f but not containing item b, item f occurs itself only or co-occurs with item g only."
- Among all transactions containing item g but not containing item b, item g occurs itself only or co-occurs with item f.
- Among all transactions containing item a but not containing item b, item a occurs itself only or co-occurs with item d.
- Among all transactions containing item d but not containing item b, item d occurs itself only or co-occurs with item a.
- The total number of transactions containing both item g and item f is 9.
- The total number of transactions containing both item a and item d is 7.

Is it possible that we could construct the original FP-tree (i.e., the first FP-tree generated from the set S)? If yes, please draw this original FP-tree. If not, please give the reasons why we could not construct the original FP-tree, and state all possible information sources needed in order to construct this FP-tree.

Q3 (15 Marks)

We are given the following 7 data points.

x_1: (1, 2), x_2: (2, 1), x_3: (10, 12), x_4: (11, 10), x_5: (8, 11), x_6: (14, 12), x_7: (4, 3)

Consider that the centroid linkage is used as a distance measurement. Please use the divisive approach to find the 2 clusters. Please show each step and the clustering result. In the clustering result, please state the mean and all points for each cluster. Besides, please include the distance between these 2 clusters.

Q4 (15 Marks)

(a) The following shows a history of customers with 3 attributes, namely "Have_NintendoSwitch", "Income" and "Age'. Attribute "Have_NintendoSwitch" indicates whether they have Nintendo Switch. We also indicate whether they will play game "Animal Crossing" or not in the last column. You cannot use XLMiner in this question.

No.	Have_NintendoSwitich	Income	Age	PlayAnimalCrossing
1	yes	high	young	yes
2	yes	high	old	yes
3	no	medium	young	yes
4	no	high	old	yes
5	no	medium	young	no
6	no	medium	young	no
7	no	medium	old	no
8	no	medium	old	no

We want to train a CART decision tree classifier to predict whether a new customer will play game "Animal Crossing" or not. We define the value of attribute PlayAnimalCrossing to be the *label* of a record.

(i) Please find a CART decision tree according to the above example. In the decision tree, whenever a node contains at most 3 records, we do not continue to process this node for splitting. Please show your steps.

(ii) Consider a new young customer whose income is medium and he has a Nintendo Switch. Please predict whether this new customer will play game "Animal Crossing" or not.

(b) Given a dataset with the following transactions in *binary* format, and the support threshold = 2.

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

(i) What is the support of the rule "{R, S} → P"?

(ii) What is the confidence of the rule "{R, S} → P"?

(iii) What is the lift ratio of the rule "{R, S} → P"?

(iv) What are the frequent itemsets? You do not need to give the frequency of each frequent itemset.

Part B (Multiple-Choice Question)

In this part, there are 8 multiple-choice questions, namely Q5, Q6, …, Q12. The total scores in this part are 40 scores. Each question weights 5 scores. In your answer sheet, please write down the following table on your **last** page of your PDF submission. In the corresponding cell, write down the answer for each question.

Note: Please write the letter **clearly** (i.e., A, B, C, D or E) 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 "B" (from some students) is similar to the hand-written letter "E". 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
Q5	
Q6	
Q7	
Q8	
Q9	
Q10	
Q11	
Q12	

Q5. You are given an Excel file called "Q5.xlsx". This Excel file stores a list of customers' purchase records on 8 different movies, namely, "Inception", "Titanic", "Joker", "Frozen", "HarryPotter", "Spiderman", "JurassicWorld" and "Parasite". The following shows the header row in this Excel file.

Inception	Titanic	Joker	Frozen	HarryPotter	Spiderman	JurassicWorld	Parasite
…	…	…	…	…	…	…	…

What is the total number of association rules found by XLMiner where the minimum support threshold is set to 5 and the minimum confidence threshold is set to 50%.

A. 0
B. 1
C. 2
D. 3
E. None of the above choices

Q6. Consider association rule mining on a table with items B, C and D. A user gives a support threshold parameter and a confidence threshold parameter. Suppose that we know that "C→B" is an interesting association rule. Which of the following statements are true?
(1) It is a must that "B→C" is an interesting rule.
(2) It is a must that "{C, D}→B" is an interesting rule.
(3) It is a must that "C→{B, D}" is an interesting rule.

A. Statement (1) only
B. Statement (2) only
C. Statement (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

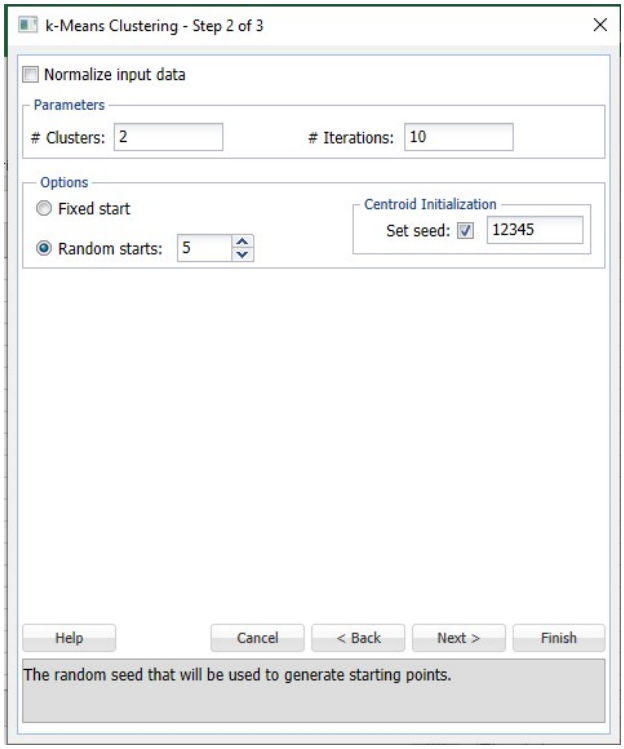
Q7. You learnt some measurements for a decision tree in class. Two of them are represented in the form of charts. They are a lift chart and a decile-wise lift chart. Which of the following statements are correct?

(1) It is always true that the greatest possible value in the y-axis in the lift chart is equal to the greatest possible value in the y-axis in the decile-wise lift chart.
(2) It is always true that we can construct the decile-wise lift chart according to the lift chart without the original training dataset.
(3) It is always true that we can construct the lift chart according to the decile-wise lift chart without the original training dataset.

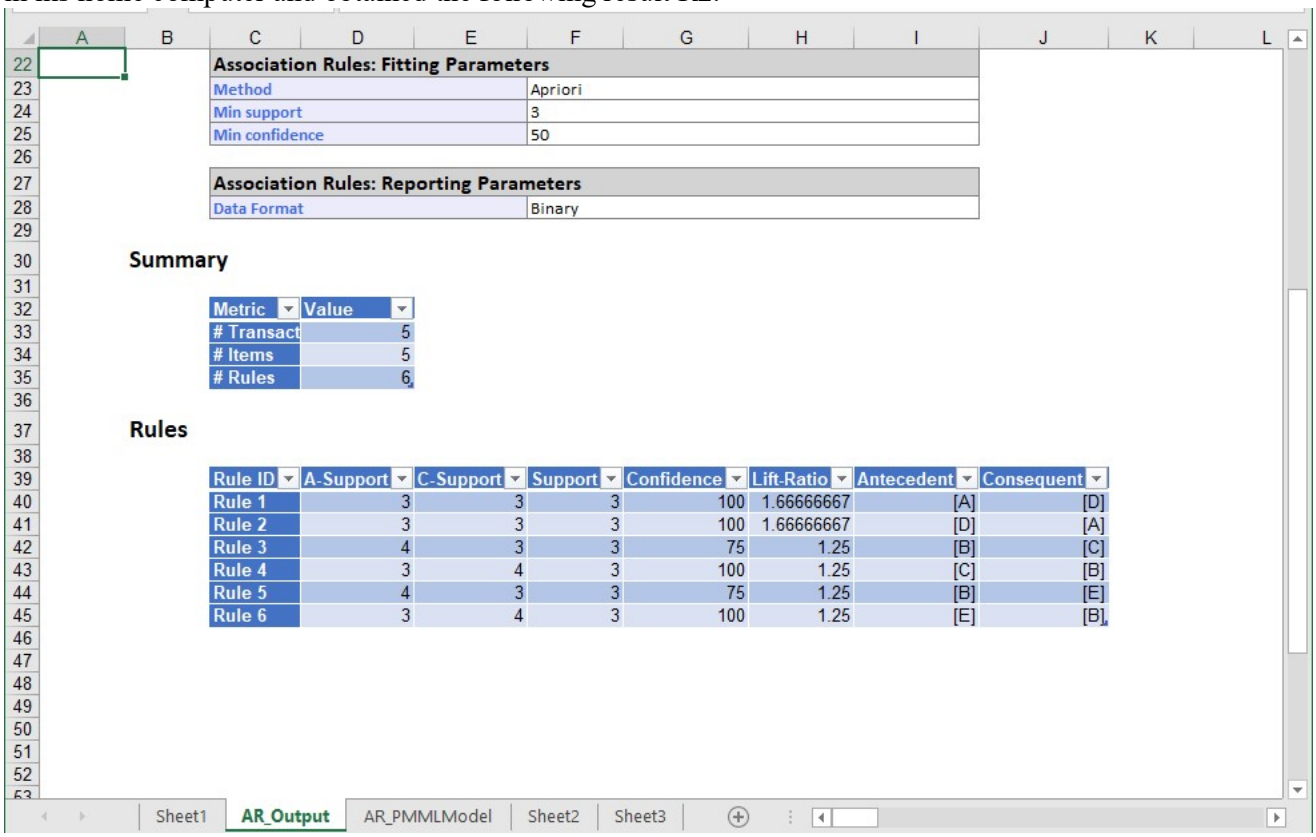
A. Statement (1) only
B. Statement (2) only
C. Statement (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q8. Consider the following two scenarios.

Scenario 1: Suppose that a user set some parameters for k-means clustering in XLMiner in his home computer according to the following k-means clustering dialog box. He obtained a clustering result R1.



Scenario 2: Suppose that the same user set some parameters for association rule mining in XLMiner in his home computer and obtained the following result R2.



Association Rules: Fitting Parameters	
Method	Apriori
Min support	3
Min confidence	50

Association Rules: Reporting Parameters	
Data Format	Binary

Summary

Metric	Value
# Transact	5
# Items	5
# Rules	6

Rules

Rule ID	A-Support	C-Support	Support	Confidence	Lift-Ratio	Antecedent	Consequent
Rule 1	3	3	3	100	1.66666667	[A]	[D]
Rule 2	3	3	3	100	1.66666667	[D]	[A]
Rule 3	4	3	3	75	1.25	[B]	[C]
Rule 4	3	4	3	100	1.25	[C]	[B]
Rule 5	4	3	3	75	1.25	[B]	[E]
Rule 6	3	4	3	100	1.25	[E]	[B]

Which of the following statements are true?

(1) In Scenario 1, if the user sets the same parameters for k-means clustering in XLMiner in his friend's computer (with the same operating system and the same XLMiner and Excel version) according to the above k-means clustering dialog box, it is a must that he will obtain the same clustering result as R1.

(2) In Scenario 2, if the user changes the minimum support threshold of a rule to 4 only, it is a must that he will obtain the same association rule result as R2.

(3) In Scenario 2, if the user changes the minimum confidence threshold of a rule to 70% only, it is a must that he will obtain the same association rule result as R2.

A. Statements (1) and (2) only
B. Statements (1) and (3) only
C. Statements (2) and (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q9. Which of the following statements are true?

(1) Consider frequent pattern mining. If we set the support threshold smaller, then the number of frequent itemsets is also smaller.
(2) Consider a lift chart. It is always true that the y-axis value of the lift curve in the chart does not decrease when the x-axis value of the lift curve increases.
(3) Consider a decile-wise lift chart. It is always true that the y-axis value of the bar in the chart does not increase when the x-axis value of the bar increases.

A. Statement (1) only
B. Statement (2) only
C. Statement (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q10. Which of the following statements are true?

(1) Under the Chi-square measure, if the Chi-square measure value for a pair of attributes is larger, the correlation between these 2 attributes is larger.
(2) It is always true that the clustering result obtained by the agglomerative approach using the median linkage is equal to the clustering result obtained by the agglomerative approach using the centroid linkage.
(3) It is possible that the clustering result obtained by the "sequential k-means" algorithm (which performs clustering one example by one example) is equal to the clustering result obtained by the "traditional k-means" algorithm (which performs clustering all examples in a batch).

A. Statements (1) and (2) only
B. Statements (1) and (3) only
C. Statements (2) and (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q11. Which of the following statements are true?

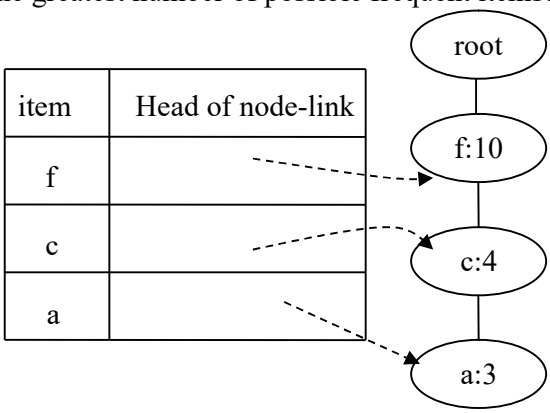
(1) Consider the Apriori algorithm. It is always true that C_i is larger than or equal to L_i for each i = 1, 2, 3, …

(2) Consider the Apriori algorithm. It is always true that C_{i+1} is larger than or equal to L_i for each i = 1, 2, 3, …

(3) Consider the distance measurement between two clusters. It is possible that the distance between two clusters computed based on the single linkage is equal to the distance between these two clusters computed based on the complete linkage.

A. Statement (1) only
B. Statement (2) only
C. Statement (3) only
D. Statements (1) and (2) only
E. None of the above choices

Q12. Suppose that we are given the following conditional FP-tree on item "b". We are also given a support threshold = 2 and the support of "b" is 15. According to all the information we have, what is the greatest number of possible frequent itemsets that we can generate?



A. 3
B. 4
C. 5
D. 7
E. None of the above choices

End of Paper

COMP1942 Exploring and Visualizing Data (Spring Semester 2020)
Online Midterm Examination (Suggested Solution)
Date: 20 April, 2020 (Monday)
Time: 12:05pm-1:10pm
Duration: 1 hour 5 minutes

Part A

Q1

(a)

Version A

Cluster 1:
Mean: (Computer=20.208, History=84.892, English=79.864, Chinese=87.416, Mathematics=25.576)
No. of points = 250

Cluster 2:
Mean: (Computer=85.014, History=87.404, English=84.906, Chinese=82.936, Mathematics=87.418)
No. of points = 500

Cluster 3:
Mean: (Computer=50.124, History=50.328, English=30.576, Chinese=30.496, Mathematics=49.86)
No. of points = 250

Version B

Cluster 1:
Mean: (Computer=95.014, History=97.404, English=94.906, Chinese=92.936, Mathematics=97.418)
No. of points = 500

Cluster 2:
Mean: (Computer=60.124, History=60.328, English=40.576, Chinese=40.496, Mathematics=59.86)
No. of points = 250

Cluster 3:
Mean: (Computer=30.208, History=94.892, English=89.864, Chinese=97.416, Mathematics=35.576)
No. of points = 250

Version C

Cluster 1:
Mean: (Computer=40.124, History=40.328, English=20.576, Chinese=20.496, Mathematics=39.86)
No. of points = 250

Cluster 2:
Mean: (Computer=10.208, History=74.892, English=69.864, Chinese=77.416, Mathematics=15.576)
No. of points = 250

Cluster 3:
Mean: (Computer=75.014, History=77.404, English=74.906, Chinese=72.936, Mathematics=77.418)
No. of points = 500

(b)

P(BelowMean | Gender=Male, Major=QSA, MockExam=No)
= [P(Gender=Male, Major=QSA, MockExam=No | Below Mean) P(BelowMean)]/
P(Gender=Male, Major=QSA, MockExam=No)
= [P(Gender=Male | BelowMean) P(Major=QSA | BelowMean) P(MockExam=No | BelowMean)
P(BelowMean)]/ P(Gender=Male, Major=QSA, MockExam=No)
= [¾ x ¼ x ¾ x ½]/ P(Gender=Male, Major=QSA, MockExam=No)
= 0.0703125 / P(Gender=Male, Major=QSA, MockExam=No)

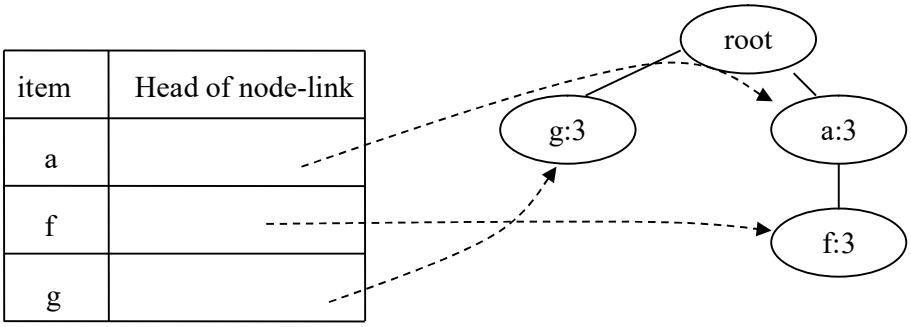
P(AboveMean | Gender=Male, Major=QSA, MockExam=No)
= [P(Gender=Male, Major=QSA, MockExam=No | Below Mean) P(AboveMean)]/
P(Gender=Male, Major=QSA, MockExam=No)
= [P(Gender=Male | AboveMean) P(Major=QSA | AboveMean) P(MockExam=No | AboveMean)
P(AboveMean)]/ P(Gender=Male, Major=QSA, MockExam=No)
= [¼ x ¾ x 0 x ½]/ P(Gender=Male, Major=QSA, MockExam=No)
= 0 / P(Gender=Male, Major=QSA, MockExam=No)
= 0

We deduce that this student will has a midterm score below the mean of the whole class.

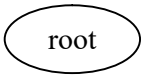
Since the sum of the probabilities is equal to 1,
P(BelowMean | Gender=Male, Major=QSA, MockExam=No)
+ P(AboveMean | Gender=Male, Major=QSA, MockExam=No) = 1
0.0703125 / P(Gender=Male, Major=QSA, MockExam=No) + 0 = 1
P(Gender=Male, Major=QSA, MockExam=No) = 0.0703125

Thus, we deduce that
P(BelowMean | Gender=Male, Major=QSA, MockExam=No) = 0.0703125/0.0703125 = 1
(which corresponds to the exact probability of this prediction).

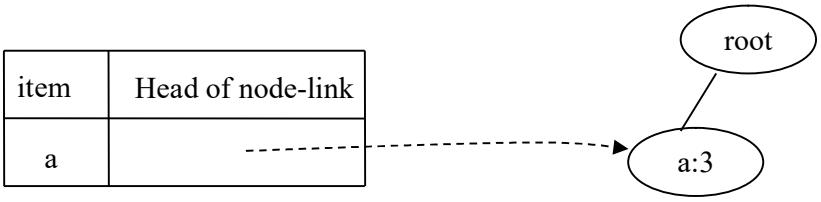
Q2
(a)
Conditional FP-tree on {b, d} where support({b, d}) = 6



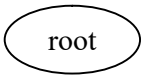
Conditional FP-tree on {b, d, g} where support({b, d, g}) = 3



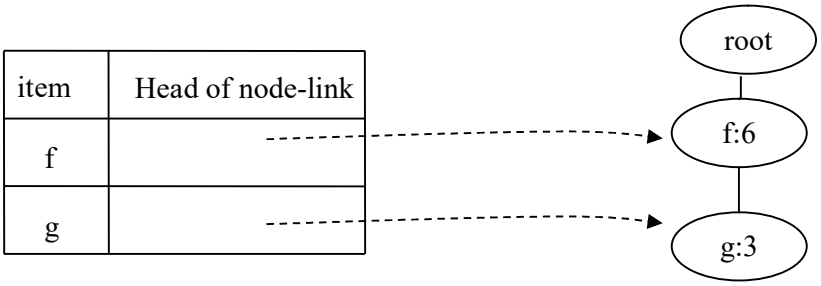
Conditional FP-tree on {b, d, f} where support({b, d, f}) = 3



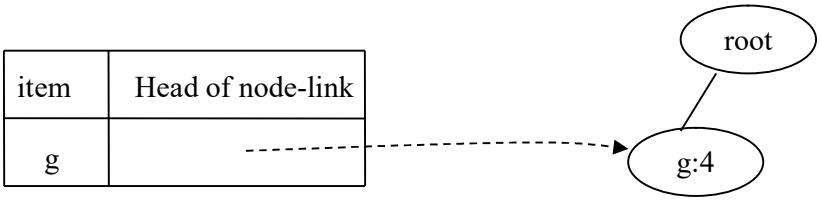
Conditional FP-tree on {b, d, a} where support({b, d, a}) = 3



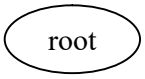
Conditional FP-tree on {b, a} where support({b, a}) = 6



Conditional FP-tree on {b, f} where support({b, f}) = 8

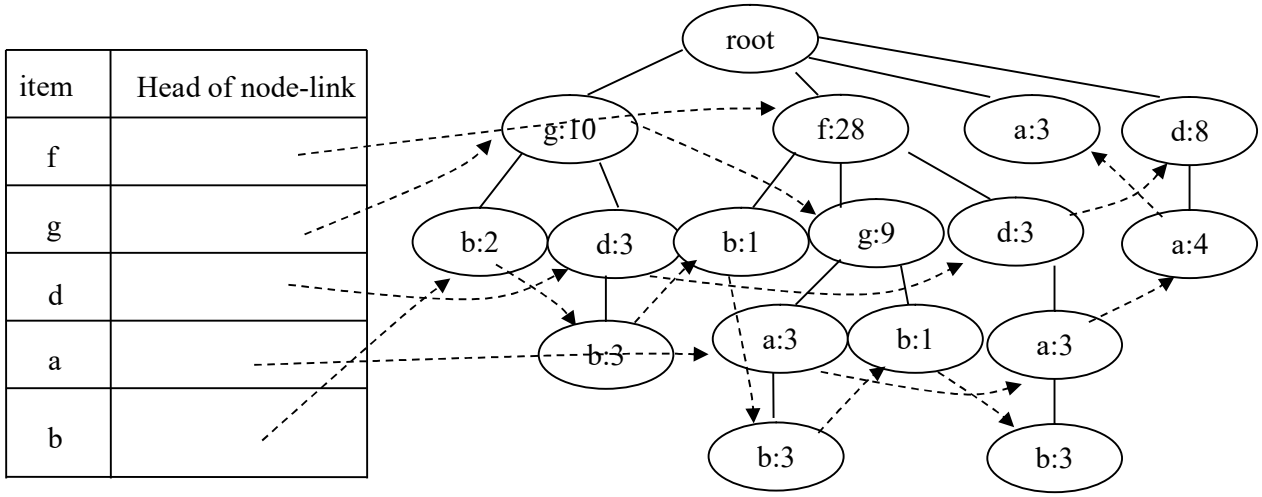


Conditional FP-tree on {b, g} where support({b, g}) = 9



(b)

Yes. It is possible. The FP-tree is shown as follows.



Q3
(a)

D(1,*)=9.45
D(2,*)=9.48
D(3,*)=6.43
D(4,*)=5.50
D(5,*)=4.45
D(6,*)=9.71
D(7,*)=6.20

A={6}
B={1,2,3,4,5,7}

D(1,A)=16.40 D(1,B)=8.07 Δ_1=-8.33
D(2,A)=16.28 D(2,B)=8.16 Δ_2=-8.12
D(3,A)=4 D(3,B) = 8.16 Δ_3=4.16
D(4,A)=3.61 D(4,B)=7.32 Δ_4=3.71
D(5,A)=6.08 D(5,B)=5.90 Δ_5=-0.17
D(7,A)=13.45 D(7,B)=4.84 Δ_7=-8.62

A={3,6}
B={1,2,4,5,7}

D(1,A)=14.87 D(1,B)=6.75 Δ_1=-8.11
D(2,A)=14.87 D(2,B)=6.80 Δ_2= -8.07
D(4,A)=2.24 D(4,B)=9.25 Δ_4= 7.02
D(5,A)=4.12 D(5,B)=7.83 Δ_5=3.70
D(7,A)=12.04 D(7,B)=3.35 Δ_7= -8.69

A={3,4,6}
B={1,2,5,7}

D(1,A)=14.17 D(1,B)=4.74 Δ_1=-9.44
D(2,A)=14.15 D(2,B)=4.92 Δ_2= -9.23
D(5,A)=3.68 D(5,B)=10.64 Δ_5= 6.95
D(7,A)=11.32 D(7,B)=1.70 Δ_7= -9.62

A={3,4,5,6}
B={1,2,7}

D(1,A)=13.44 D(1,B)=2 Δ_1=-11.44
D(2,A)=13.48 D(2,B)=1.58 Δ_2= -11.90
D(7,A)=10.66 D(7,B)=2.92 Δ_7= -7.74

Stop!
A={3,4,5,6}
B={1,2,7}

So, there are two clusters:
Cluster 1: data points 3, 4, 5, 6
Cluster 2: data points 1, 2, 7

The distance between 2 clusters is 12.51.

Q4

(a) (i)

Info(T) = $1-0.5^2-0.5^2=0.5$

For attribute Have-NintendoSwitch,

Info(T_{yes})=$1-1^2-0^2=0$

Info(T_{no})=$1-(\frac{1}{3})^2-(\frac{2}{3})^2=0.4444$

Info(Have-NintendoSwitch, T)= $\frac{1}{4}Info(T_{yes})+\frac{3}{4}Info(T_{no})=0.3333$

Gain(Have-NintendoSwitch, T)= Info(T)-Info(Have-NintendoSwitch, T)= 0.1667

For attribute Income,

Info(T_{high})= $1-1^2-0^2=0$

Info(T_{medium})= $1-(\frac{1}{5})^2-(\frac{4}{5})^2=0.32$

Info(Income, T)= $\frac{3}{8}Info(T_{high})+\frac{5}{8}Info(T_{medium})=0.2$

Gain(Income, T)= Info(T)-Info(Income, T)=0.3

For attribute Age,

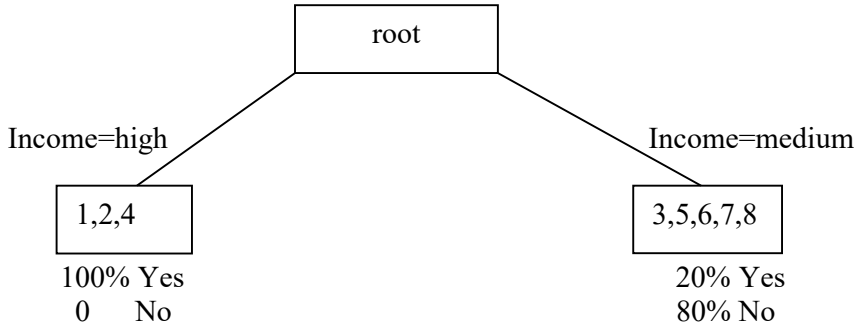
Info(T_{young})=$1-0.5^2-0.5^2=0.5$

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

We choose attribute Income for Splitting:



Consider the node for "Income=medium"

Info(T)= $1-(\frac{1}{5})^2-(\frac{4}{5})^2=0.32$

For attribute Have-NintendoSwitch,

Info(T_{yes})=undefined

Info(T_{no})=$1-(\frac{1}{5})^2-(\frac{4}{5})^2=0.32$

Info(Have-NintendoSwitch, T)= $0\times Info(T_{yes})+1\times Info(T_{no})=0.32$

Gain(Have-NintendoSwitch, T)= Info(T)-Info(Have-NintendoSwitch, T)= 0

For attribute Age,

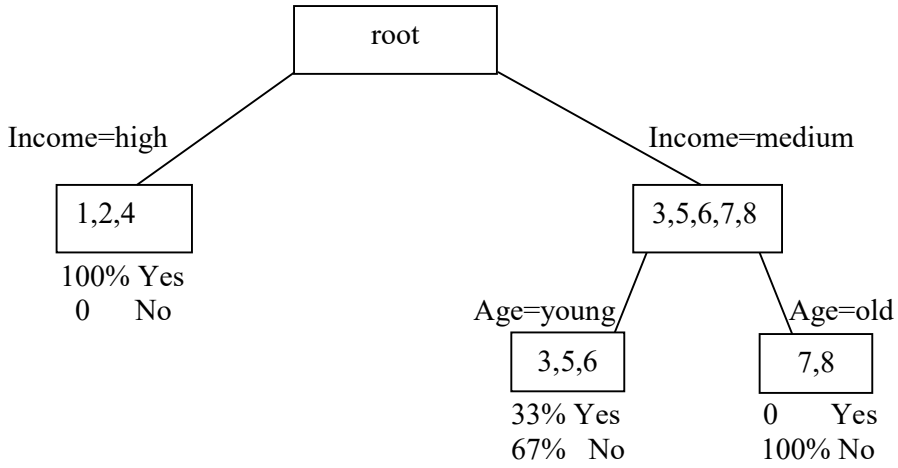
Info(T_{young})=$1-(\frac{1}{3})^2-(\frac{2}{3})^2=0.4444$

Info(T_{old})=$1-1^2-0^2=0$

Info(Age, T)= $\frac{3}{5}Info(T_{young})+\frac{2}{5}Info(T_{old})=0.26664$

Gain(Age, T)= Info(T)-Info(Age, T)= 0.05336

We choose attribute Age for Splitting:



(ii) It is likely that he will not play game "Animal Crossing".

(b) (i) support = 2

(ii) confidence = 2/3 = 66.7%

(iii) expected confidence of the consequent of the rule = 3/5
lift ratio of the rule = 66.7/60 = 1.11

(iv) freq. itemsets
= { {P}, {Q}, {R}, {S},
{P, R}, {P, S}, {Q, R},{Q, S}, {R, S},
{P, R, S}, {Q, R, S} }

Part B

Version 1

Question	Your Answer
Q5	C
Q6	E
Q7	B
Q8	B
Q9	B
Q10	B
Q11	E
Q12	E

Version 2

Question	Your Answer
Q5	C
Q6	B
Q7	B
Q8	E
Q9	E
Q10	E
Q11	B
Q12	B

Version 3

Question	Your Answer
Q5	C
Q6	B
Q7	E
Q8	B
Q9	B
Q10	E
Q11	B
Q12	E

End of Paper