COMP1942 Exploring and Visualizing Data (Spring Semester 2022)
Online Final Examination (Question Paper)
Date: 19 May, 2022 (Thursday)
Time: 4:30pm-6:30pm
Duration: 2 hours

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) Please answer **all** questions. The total scores in this exam are 200.
- (b) There are 2 parts in this exam, Part A (Short/Long Question) and Part B (Multiple-Choice Question).

(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 together with the pre-15-minute period and the post-15-minute period.

(5) File Submission
- (a) We allow a 15-minute buffer for your PDF file upload. Remember to upload your file at around 6:30pm. We allow your file uploading time at most 15 minutes. Canvas will terminate any file uploading process at 6:45pm if your file is still being uploaded at 6:45pm.

(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, your exam score will be set to 0.
- (c) We only mark your latest PDF file uploaded by 6:45pm. Your exam score will be set to 0 if we could not see any PDF file uploaded by 6:45pm (even though you do the question paper or you "could" upload your PDF file after 6:45pm).

Part A (Short/Long Question)

In this part, there are 8 short/long questions, namely Q1, Q2, Q3, Q4, Q5, Q6, Q7 and Q8. The total scores in this part are 160 scores (out of 200). Each question weights 20 scores (out of 200).

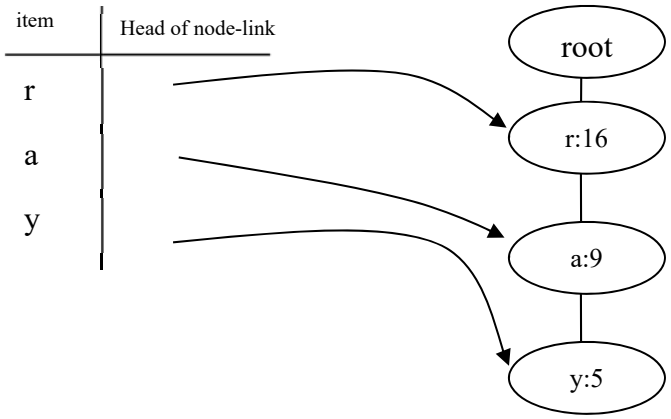
Q1 (20 Marks)

(a) We are given a dataset with the following 5 transactions. The support threshold is set to 2.

TID	Items
1	Q, T
2	H, Q, T
3	S
4	H, Q, S, T
5	H, Q, S

After we perform the join step and the prune step in the Apriori algorithm, we obtain a set C of itemsets. Then, we need to do the counting step for C (i.e., we need to find the frequency of each itemset in C). Finally, we output all itemsets in C with frequency at least a given support threshold as a part of the final output. Why do we need to do the counting step? That is, why can't we simply output C as a part of the final output? Please elaborate it. You can use the above dataset for illustration.

(b) The following shows an FP-tree. Let the support threshold be 3.
Please list all frequent itemsets with their correspondence frequency counts.



Q2 (20 Marks)

(a) Consider eight two-dimensional data points, namely x_1, x_2, x_3, …, x_8. We only know the coordinates of the first three data points:

x_1: (2, 3), x_2: (5, 2) , x_3: (12, 15)

We also know the pairwise distance matrix among these eight data points according to the Euclidean distance.

	1	2	3	4	5	6	7	8
1	0							
2	$\sqrt{10}$	0						
3	$\sqrt{244}$	$\sqrt{218}$	0					
4	$\sqrt{490}$	$\sqrt{388}$	$\sqrt{146}$	0				
5	$\sqrt{325}$	$\sqrt{265}$	$\sqrt{29}$	$\sqrt{45}$	0			
6	2	$\sqrt{26}$	$\sqrt{288}$	$\sqrt{578}$	$\sqrt{389}$	0		
7	$\sqrt{5}$	1	$\sqrt{233}$	$\sqrt{425}$	$\sqrt{290}$	$\sqrt{17}$	0	
8	$\sqrt{580}$	$\sqrt{530}$	$\sqrt{72}$	$\sqrt{146}$	$\sqrt{65}$	$\sqrt{648}$	$\sqrt{557}$	0

(i) Is it possible to find the coordinates of all other five data points (i.e., x_4, x_5,…, x_8)? If yes, please write down the coordinates of all of these data points. Otherwise, please write down the coordinates of some of these data points (if any) and explain why the coordinates of the remaining data points could not be found. Please show all fractional numbers up to 4 decimal places.

(ii) If the answer of (a) (i) is "yes", please use the agglomerative approach to group these points with the centroid linkage. Draw the corresponding dendrogram for the clustering. Please show all of the steps. You are required to specify the distance metric in the dendrogram. Please show all fractional numbers up to 4 decimal places.

If the answer of (a) (i) is "no", please use the agglomerative approach to group these points with the group average linkage. Draw the corresponding dendrogram for the clustering. Please show all of the steps. You are required to specify the distance metric in the dendrogram. Please show all fractional numbers up to 4 decimal places.

(b) Consider a classification problem for the table with two input attributes, namely A_1 and A_2, and one target attribute Y, containing 200 records.

(i) In the support vector machine, we learnt that we want to maximize the margin in a classification problem. We learnt that the margin is equal to

$$\frac{2}{\sqrt{w_1^2 + w_2^2}}$$

where w_1 and w_2 are two variables to be found. In class, we learnt that we need to re-write the objective function as $w_1^2 + w_2^2$ and then we want to minimize this objective function. Why do we need to re-write this objective function?

(ii) In the support vector machine, how many constraints are there in form of $Y(w_1A_1 + w_2A_2 + b) \geq 1$ where w_1, w_2 and b are three variables to be found?

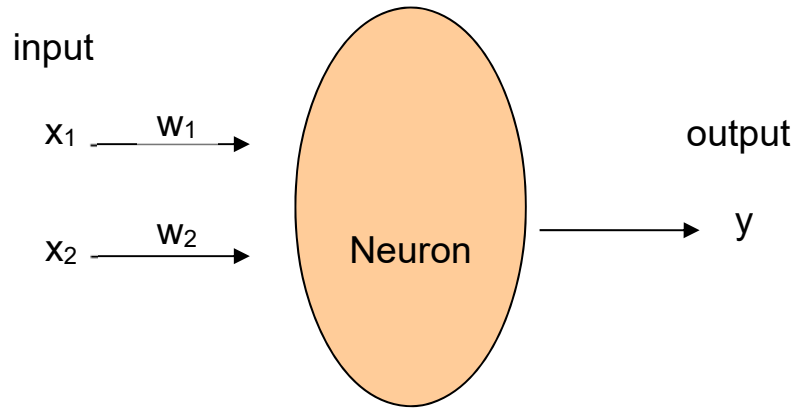
Q3 (20 Marks)

Consider the following table with three attributes where "Height" and "Weight" are input attributes and "HeartDisease" is the target attribute. Each tuple corresponds to an individual. An attribute "Record ID" denotes the ID of each record.

Record ID	Height (cm)	Weight (kg)	HeartDisease
1	170	90	Yes
2	190	95	No
3	160	50	No
4	180	70	No

(a) Rewrite the above table such that values "Yes" and "No" in attribute "HeartDisease" are mapped to values 1 and -1, respectively.

(b) Consider a neural network containing a single neuron where x_1 = "Height", x_2 = "Weight" and y = "HeartDisease".



Initially, we set the values of w_1, w_2 and b to be 0.7 where b is a bias value in the neuron.

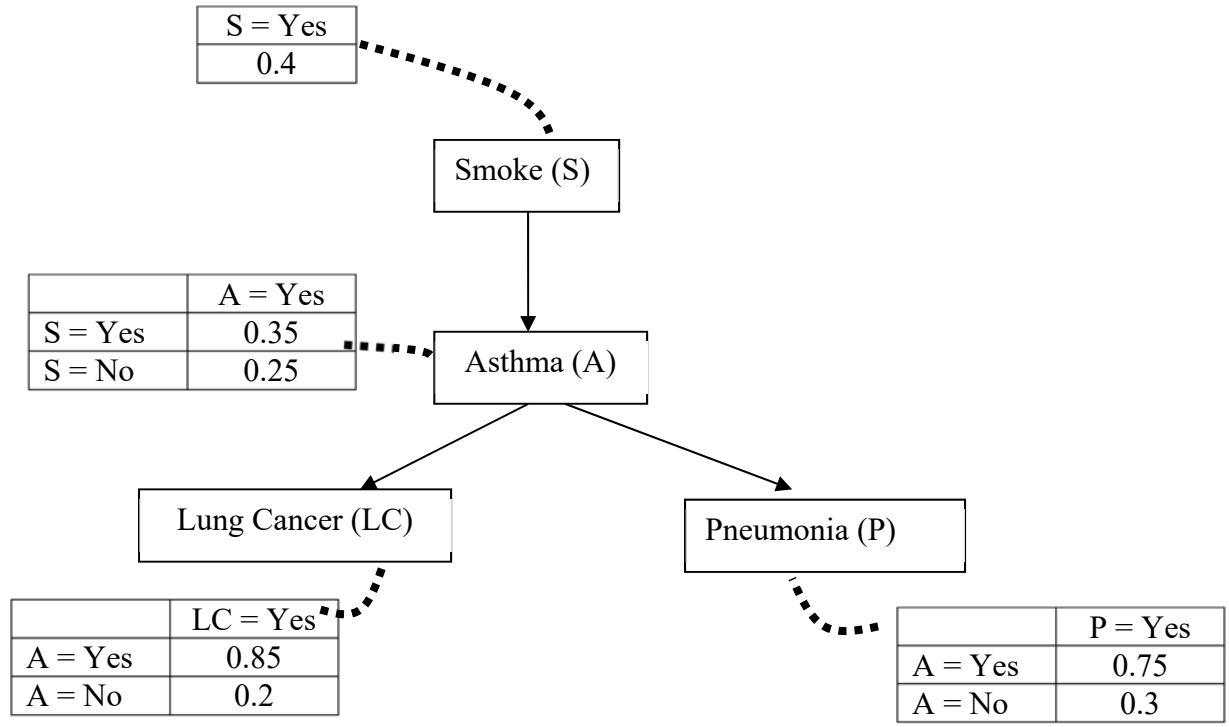
Suppose the learning rate is denoted by α. Let $\alpha = 0.3$.
Suppose we adopt the hyperbolic tangent function as an activation function.
In this example, we do not need to normalize the input attributes.
Please try to train the neural network with five instances by the following inputs in the given sequence and the answer in Part (a).

1. $(x_1, x_2) = (170, 90)$
2. $(x_1, x_2) = (190, 95)$
3. $(x_1, x_2) = (160, 50)$
4. $(x_1, x_2) = (180, 70)$
5. $(x_1, x_2) = (170, 90)$

What are the final values of w_1, w_2 and b after these five instances?
In this question, please show all of your steps and show all numbers up to 4 decimal places.

Q4 (20 Marks)

We have the following Bayesian Belief Network.



Suppose that there is a new person. We know that

(1) he has lung cancer

(2) he has no pneumonia

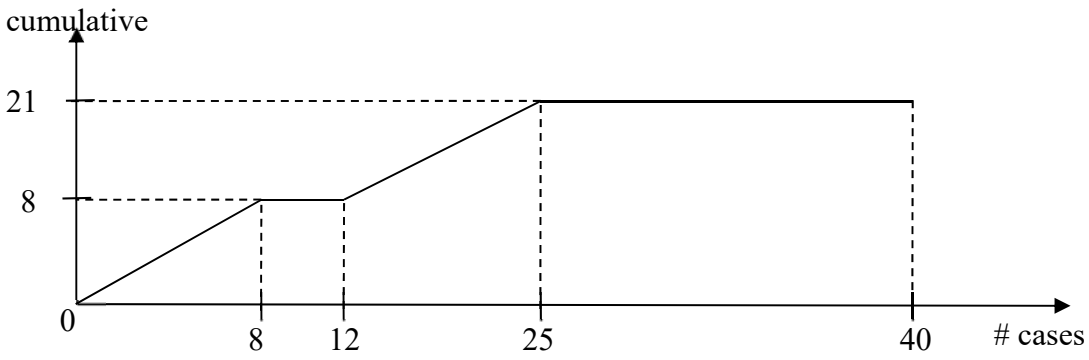
We would like to know whether he is likely to smoke or not.

Please use Bayesian classifier with the use of Bayesian Belief Network to predict whether he is likely to smoke.

Please show all steps and all fractional numbers up to 6 decimal places.

Q5 (20 Marks)

(a) We are given the following lift chart based on a classifier.



In the following, please show all fractional numbers up to 4 decimal places.

(i) Is it possible to find the accuracy of the classifier? If yes, please write down the number. Otherwise, please explain it.
(ii)Is it possible to find the precision of the classifier? If yes, please write down the number. Otherwise, please explain it.
(iii) Is it possible to find the specificity of the classifier? If yes, please write down the number. Otherwise, please explain it.
(iv) Is it possible to find the number of false negatives? If yes, please write down the number. Otherwise, please explain it.
(v) Is it possible to find the decile-wise lift chart? If yes, please give the chart. In the chart, please write down the height of the bar explicitly (next to the top of the bar). Otherwise, please explain it.

(b) In the training phase, usually, the original dataset containing the target attribute is split into three data sets. Please give the names of these three datasets. Besides, please give the purpose of using each of these datasets.

Q6 (20 Marks)

(a) Consider a table T: (part, supplier, customer, price) where "part" is an attribute for parts, "supplier" is an attribute for suppliers, "customer" is an attribute for customers and "price" is an attribute for prices. A record (p, s, c, x) means that the part p, supplied by supplier s and bought by customer c, has its price x. Suppose that the total size of this table is 20GB. We materialize this table.

Consider the following seven queries, namely Q1, Q2, Q3, Q4, Q5, Q6 and Q7.

Q1: We want to find the total price (or the sum of the prices) for each combination of part and customer.

Q2: We want to find the total price (or the sum of the prices) for each part.

Q3: We want to find the total number of records in T for each combination of part and customer.

Q4: We want to find the total number of records in T for each part.

Q5: We want to find the total number of records in T for each supplier.

Q6: We want to find the average price for each combination of part and customer.

Q7: We want to find the average price for each part.

Suppose that we materialize the answers of Q1, Q3 and Q6. Each of these answers occupies 4GB storage.

We know that we can find the answer of Q2 from the answer of Q1 only in class.

(i) Is it a must that we can find the answer of Q4 from the answer of **only** one of the materialized answers of the queries (except the original table) (i.e., Q1, Q3 and Q6)? If yes, please give the only one query with the materialized answer (except the original table) and elaborate how to find the answer of Q4 from the materialized answer of this query. Otherwise, please give what materialized answers of queries together with the original table (if needed) that we can use with the minimum overall access cost (in GB) and explain it. For each of the above cases, please write down the minimum access cost.

(ii) Is it a must that we can find the answer of Q5 from the answer of **only** one of the materialized answers of the queries (except the original table) (i.e., Q1, Q3 and Q6)? If yes, please give the only one query with the materialized answer (except the original table) and elaborate how to find the answer of Q5 from the materialized answer of this query. Otherwise, please give what materialized answers of queries together with the original table (if needed) that we can use with the minimum overall access cost (in GB) and explain it. For each of the above cases, please write down the minimum access cost.

(iii) Is it a must that we can find the answer of Q7 from the answer of **only** one of the materialized answers of the queries (except the original table) (i.e., Q1, Q3 and Q6)? If yes, please give the only one query with the materialized answer (except the original table) and elaborate how to find the answer of Q7 from the materialized answer of this query. Otherwise, please give what materialized answers of queries together with the original table (if needed) that we can use with the minimum overall access cost (in GB) and explain it. For each of the above cases, please write down the minimum access cost.

(b) In class, we learnt "Sequential K-means Clustering" and "Forgetful Sequential K-means Clustering". What is the scenario or application that "Forgetful Sequential K-means Clustering" is better used compared with "Sequential K-means Clustering"?

Q7 (20 Marks)

Suppose that c is a positive real number where we do not know the exact value.
Similarly, d is also another positive real number where d is equal to c+7.

In the following, please show all fractional numbers up to 4 decimal places.

(a) Consider the four 2-dimensional data points:

a:(6 + c, 6 + c), b:(9 + c, 10 + c), c:(4 + c, 11 + c), d:(10 + c, 5 + c)

We can make use of PCA for dimensionality reduction. In dimensionality reduction, given an L-dimensional data point, we want to transform this point to a K-dimensional data point where K < L such that the information loss during the transformation is minimized. Suppose that L = 2 and K = 1.
Please illustrate with the above example. Please show all the steps.

(b) Consider the four 2-dimensional data points:

a:(6 – d, 6 – d), b:(9 – d, 10 – d), c:(4 – d, 11 – d), d:(10 – d, 5 – d)

We can make use of PCA for dimensionality reduction. In dimensionality reduction, given an L-dimensional data point, we want to transform this point to a K-dimensional data point where K < L such that the information loss during the transformation is minimized. Suppose that L = 2 and K = 1.
Can we make use of the answers in part (a) to perform the dimensionality reduction? If yes, please write down each transformed data point. If no, please write down the reasons why we cannot make use of the answers of part (a).

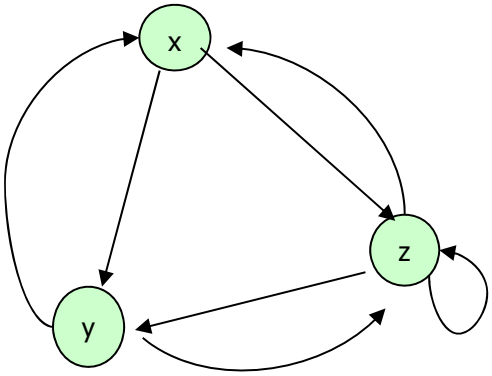
(c) Consider the four 2-dimensional data points:

a:(6c, 6c), b:(9c, 10c), c:(4c, 11c), d:(10c, 5c)

We can make use of PCA for dimensionality reduction. In dimensionality reduction, given an L-dimensional data point, we want to transform this point to a K-dimensional data point where K < L such that the information loss during the transformation is minimized. Suppose that L = 2 and K = 1.
Can we make use of the answers in part (a) to perform the dimensionality reduction? If yes, please write down each transformed data point. If no, please write down the reasons why we cannot make use of the answers of part (a).

Q8 (20 Marks)

In the following, please show all fractional numbers up to 3 decimal places.
The following shows three sites, namely x, y and z, with their linkage.



(a) What is the adjacency matrix?
(b) What is the stochastic matrix?
(c) Suppose that site y and site z wants to form a group which would like to become a "spider trap". Please list out the set of all operations that site y and site z have to do such that the total number of operations involved is minimum.
(d) In the PageRank algorithm, we need to update a ranking vector r by "$0.7 \cdot M \cdot r + c$" iteratively where M is the stochastic matrix and c is a vector $(0.3, 0.3, \ldots 0.3)^T$. Suppose that r_n is the ranking vector after the update and r_0 is the ranking vector before the update. For simplicity, you can assume that there are only four sites in this PageRank algorithm. Prove that if the sum of the values in r_0 is equal to 4, then the sum of the values in r_n is equal to 4. In the proof, please use the following notations.

$$r_0 = \begin{pmatrix} r_{0,1} \\ r_{0,2} \\ r_{0,3} \\ r_{0,4} \end{pmatrix}, \qquad r_n = \begin{pmatrix} r_{n,1} \\ r_{n,2} \\ r_{n,3} \\ r_{n,4} \end{pmatrix}, \qquad M = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix}$$

Part B (Multiple-Choice Question)

In this part, there are 8 multiple-choice questions, namely Q9, Q10, Q11, Q12, Q13, Q14, Q15 and Q16. The total scores in this part are 40 scores (out of 200). Each question weighs 5 scores (out of 200). 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
Q9	
Q10	
Q11	
Q12	
Q13	
Q14	
Q15	
Q16	

Q9. The following shows a history of people with their income, gender and education. We also indicate whether they will study the PhD program or not in the last column.

No.	Income	Gender	Education	Study_PhD
1	low	female	bachelor	yes
2	medium	female	bachelor	yes
3	high	male	bachelor	yes
4	high	male	master	yes
5	medium	male	diploma	no
6	high	male	bachelor	no
7	high	female	diploma	no
8	low	female	diploma	no

We want to train a C4.5 decision tree classifier to predict whether a new person will study the PhD program or not. We define the value of attribute Study_PhD 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 statements are correct?

(1) There are four terminal nodes in this decision tree.
(2) Consider a new female person whose income is high but her education is bachelor. According to this decision tree, it is very likely that this person will not study the PhD program.
(3) There are six nodes in this decision tree (where nodes includes both the decision nodes and the terminal nodes).

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

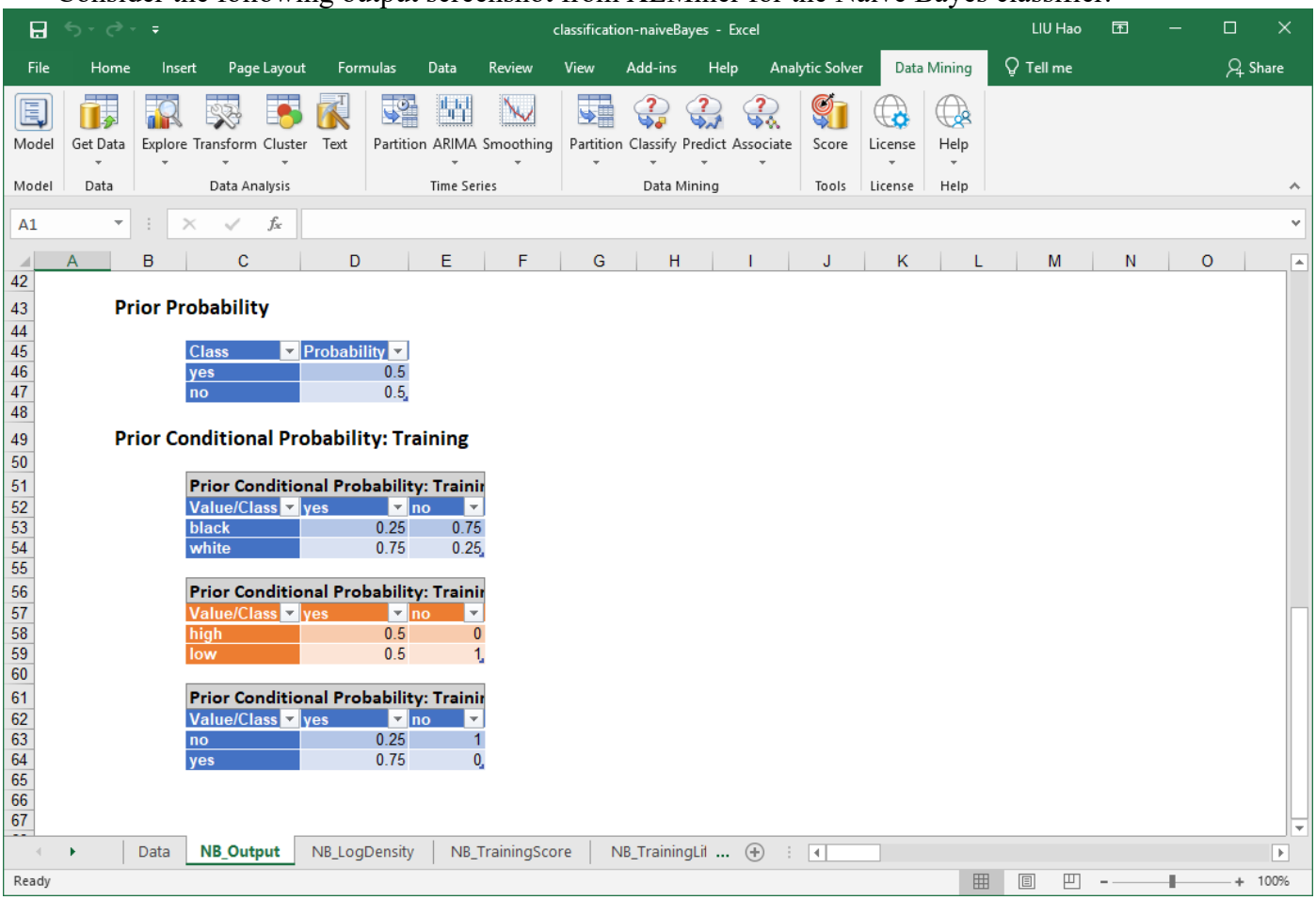
Q10. Which of the following statements are correct?

(1) In the k-nearest neighbor classifier, it is more preferable to setting parameter k to an even number compared with an odd number.
(2) LSTM has a more complex structure compared with GRU.
(3) The rollup operation could be performed when we want to change from one query specifying some detailed descriptions to another query specifying more general descriptions.

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. We are given a table with four attributes, namely Race, Income, Child and Insurance, where the first 3 are input attributes and the last one is the target attribute for classification. Note that attribute Race has values "black" and "white", attribute Income has values "high" and "low", and attribute Child has values "yes" and "no". The target attribute Insurance has values "yes" and "no".

Consider the following output screenshot from XLMiner for the Naïve Bayes classifier.



Which of the following statements are true?

(1) According to the above information only, we could derive that P(Insurance = no | Race = black) = 0.75.

(2) According to the above information only, we could derive that P(Child = no | Insurance = no) = 1.

(3) According to the above information only and P(Income=low)=3/4, we could derive that P(Insurance = no | Income = low) = 2/3.

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

Q12. Which of the following statements are true?

(1) Consider association rule mining on a given table. It is a must that the number of association rules is larger if we set the confidence threshold value to a smaller value (and keep the support threshold value).
(2) Consider association rule mining on a given table. It is a must that the lift ratio of an association rule in the form of "A → B" is equal to the lift ratio of an association rule in the form of "B→A" where A and B are two items in the table.
(3) It is a must that if the support of an association rule in the form of "A→B" is at least 4, then both the support of A and the support of B is at least 4.

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

Q13. We are given a support threshold equal to 10. Which of the following statements are true?

(1) It is always true that the total number of non-root nodes in the FP-tree is smaller than the total number of occurrences of frequent items in the given table.
(2) It is always true that the original table could be constructed according to the FP-tree.
(3) It is always true that all frequent itemsets could be generated based on the FP-tree.

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

Q14. Which of the following statements are true?

(1) Consider the agglomerative approach. It is possible that the mean of each cluster found based on the centroid linkage is exactly equal to the center of one of the clusters found based on the median linkage.
(2) Consider the agglomerative approach. It is possible that the mean of each cluster found based on the centroid linkage is not exactly equal to the centers of all clusters found based on the median linkage.
(3) It is a must that we should know the number of clusters to be found before we perform the agglomerative approach.

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

Q15. Consider the data warehouse technique we learnt in class. Given a set V of views to be materialized, we know how to compute Gain(V ∪ {top view}, {top view}). Which of the following statements are true?

(1) Suppose that S and T are two sets of views to be materialized such that "S ⊆ T".
(Note that "S ⊆ T" means that each view in set S can be found in set T.)
It is always true that for any view x,
Gain({x} ∪ S ∪ {top view}, {top view}) - Gain(S ∪ {top view}, {top view})
≥ Gain({x} ∪ T ∪ {top view}, {top view}) - Gain(T ∪ {top view}, {top view}).
(2) Suppose that view P and view C are two nodes where P is a parent node of C (i.e., P is just above C) in the relationship graph. It is always true that
Gain({view P} ∪ {top view}, {top view}) ≥ Gain({view C} ∪ {top view}, {top view}).
(3) Let A be a view. It is always true that
Gain(V ∪ {top view}, {top view}) ≥ Gain(V ∪ {top view, view A}, {top view, view A}).

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

Q16. 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 we can construct the decile-wise lift chart according to the lift chart without the original training dataset.
(2) It is always true that we can construct the lift chart according to the decile-wise lift chart without the original training dataset.
(3) 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.

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

End of Paper