COMP1942 Exploring and Visualizing Data (Spring Semester 2017)
Final Examination (Question Paper)
Date: 27 May, 2017 (Sat)
Time: 16:30-19:30
Duration: 3 hours

Student ID:_____ Student Name:_____

Seat No. :_____

Instructions:
(1) Please answer **all** questions in Part A in the **answer sheet**.
(2) You can **optionally** answer the bonus question in Part B in the answer sheet. You can obtain additional marks for the bonus question if you answer it correctly.
(3) You can use a calculator.

Question Paper

Part A (Compulsory Short Questions)

Q1 (20 Marks)

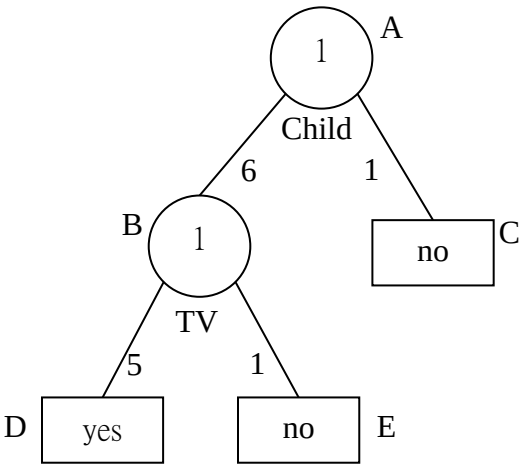
We are given a dataset with the following transactions.

TID	Items
1	c, d
2	b, d
3	a, b, d
4	a, d, f
5	a, e

(a) (i) What is the confidence of "d→a"?
(ii) What is the lift ratio of "d→a"?
(b) Let the support threshold be 2.
(i) Draw the FP-tree.
(ii) Draw the conditional FP-tree on a.
(c) What are the disadvantages of the Apriori Algorithm?

Q2 (20 Marks)

(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, "Child" corresponds to the number of children and "TV" corresponds to the number of TVs.



(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 "Child" = 1 and "TV"=1. What is the predicted value of the target attribute of this new record according to this decision tree?

(b) Consider the C4.5 decision tree. In the original formula, the "log" function has its base 2 (i.e., $\log_2 x$ where x is a number). Based on this base, we obtain a decision tree T_2. Let b is a positive number. Suppose that we change the base of the "log" function from 2 to b. Based on this new base, we obtain a decision tree T_b. Is it always true that T_2 is equal to T_b? Please elaborate. If yes, please prove it. If no, please give an example showing that T_2 is not equal to T_b and explain it.

Q3 (20 Marks)

(a) (i) What is the major difference between classification and clustering?
(ii) What is the major similarity between classification and clustering?

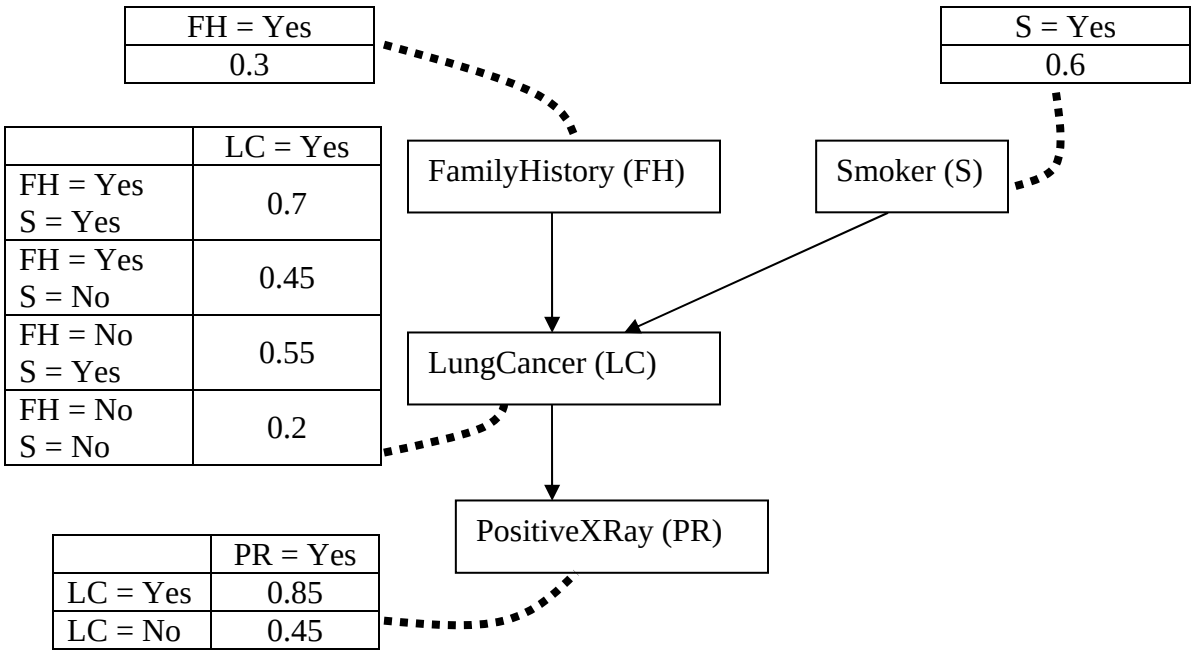
(b) Consider seven data points.
The following matrix shows the pairwise distances between any two points.

	1	2	3	4	5	6	7
1	0						
2	10	0					
3	7	7	0				
4	30	23	21	0			
5	29	25	22	7	0		
6	38	34	31	10	11	0	
7	42	36	36	13	17	9	0

Please use the divisive (polythetic) approach to divide these seven points into two groups/clusters by using group average linkage.

Q4 (20 Marks)

We have the following Bayesian Belief Network.



	LC = Yes
FH = Yes S = Yes	0.7
FH = Yes S = No	0.45
FH = No S = Yes	0.55
FH = No S = No	0.2

	PR = Yes
LC = Yes	0.85
LC = No	0.45

Suppose that there is a new person. We know that
(1) he has his family history
(2) he is a non-smoker
(3) his result of X-Ray is positive

We would like to know whether he is likely to have Lung Cancer.

Family History	**Smoker**	**PositiveXRay**	**Lung Cancer**
Yes	No	Yes	?

(a) Please use Bayesian Belief Network classifier with the use of Bayesian Belief Network to predict whether he is likely to have Lung Cancer.

(b) Although Bayesian Belief Network classifier does not have an independent assumption among all attributes (compared with the naïve Bayesian classifier), what are the disadvantages of this classifier?

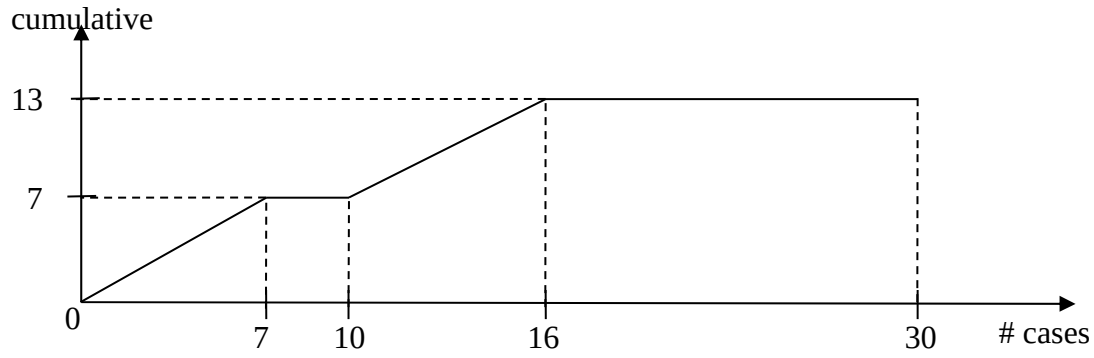
Q5 (20 Marks)

(a) Consider the following table where the first three columns correspond to the input attributes and the fourth column corresponds to the target attribute.

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	high	no	no

In XLMiner, we want to find a decision tree (or classification tree) from the above table. But, we cannot directly use this table. Instead, we need to do a transformation process on this table so that we can use the transformed table to find a decision tree by XLMiner. What is this transformation process?

(b) This part is independent of Part (a). We are given the following lift chart based on a classifier.



(i) Is it possible to find the number of false positives? If yes, please write down the number. Otherwise, please explain it.

(ii)Is it possible to find the number of true positives? If yes, please write down the number. Otherwise, please explain it.

(iii) Is it possible to find the number of false negatives? If yes, please write down the number. Otherwise, please explain it.

(iv) Is it possible to find the number of true 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. Otherwise, please explain it.

Q6 (20 Marks)

(a) 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.

(b) Consider a classification problem and a classifier. Suppose that there is at least one record with the target attribute equal to "Yes" in the training dataset and there is at least one record with the predicted target attribute equal to "Yes" in the same dataset.

(i) Is it possible that the recall is equal to 50% but the specificity is equal to 100%? Please explain.

(ii) Is it possible that the precision is equal to 50% but the specificity is equal to 100%? Please explain.

(iii) Is it possible that the precision is equal to 100% but the recall is equal to 0%? Please explain.

(iv) Is it possible that the precision is equal to 0% but the recall is equal to 100%? Please explain.

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+5.

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

a:(7 + c, 7 + c), b:(9 + c, 9 + c), c:(6 + c, 10 + c), d:(10 + c, 6 + 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.

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

a:(7 – d, 7 – d), b:(9 – d, 9 – d), c:(6 – d, 10 – d), d:(10 – d, 6 – 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).

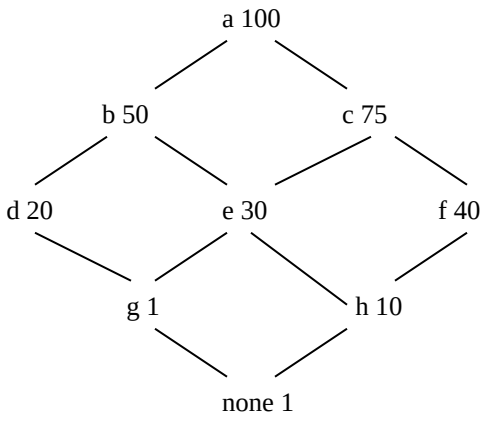
(c) Consider the four 2-dimensional data points:

a:(7c, 7c), b:(9c, 9c), c:(6c, 10c), d:(10c, 6c)

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)

We are given the following lattice containing nodes where each node corresponds to a possible query. The number associated with each query corresponds to the cost of answering this query.



Assume that we do not consider “none 1”. Suppose 3 views are to be materialized (other than the top view). Apply the greedy algorithm and find the resulting views.

Q9 (20 Marks)

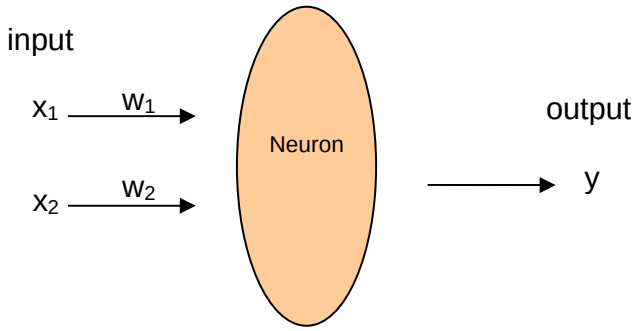
(a) In class, we learnt that given three random variables, namely X, Y and Z, X is said to be conditionally independent of Y given Z if P(X | Y, Z) = P(X | Z).
Please state whether the following statement is true or not according to the above concept.
If yes, please give a proof. If no, please explain it.

"If X is conditionally independent of Y given Z, then P(X, Y | Z) = P(X | Z) x P(Y | Z)."

(b) The following shows the AND function where x_1 and x_2 are two inputs and y is the output.

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Consider a neural network containing a single neuron.



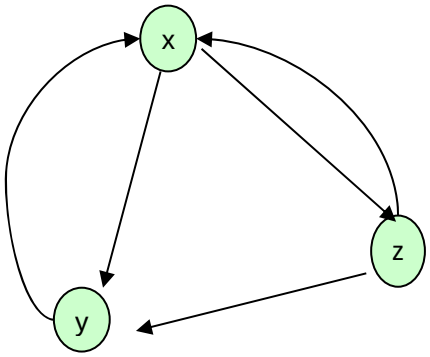
Initially, we set the values of w_1, w_2 and b to be 0.1 where b is a bias value in the neuron.
Suppose the learning rate is denoted by α. Let $\alpha = 0.5$.
Suppose we adopt the threshold function as an activation function.
Please try to train the neural network with the following 5 inputs in the given sequence.

1. $(x_1, x_2) = (0, 0)$
2. $(x_1, x_2) = (0, 1)$
3. $(x_1, x_2) = (1, 0)$
4. $(x_1, x_2) = (1, 1)$
5. $(x_1, x_2) = (0, 0)$

What are the final values of w_1, w_2 and b after these five instances are read?

Q10 (20 Marks)

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 x would like to become a "spider trap". Please list out all possible operations that site x has to do.
(d) In the PageRank algorithm, we need to update a ranking vector r by "$0.8 \cdot M \cdot r + c$" iteratively where M is the stochastic matrix and c is a vector $(0.2, 0.2, \ldots 0.2)^T$. 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 three sites in this PageRank algorithm. Prove that if the sum of the values in r_0 is equal to 3, then the sum of the values in r_n is equal to 3. In the proof, please use the following notations.

$$r_0 = \begin{pmatrix} r_{0,1} \\ r_{0,2} \\ r_{0,3} \end{pmatrix}, \qquad r_n = \begin{pmatrix} r_{n,1} \\ r_{n,2} \\ r_{n,3} \end{pmatrix}, \qquad M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

Part B (Bonus Question)

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

Q11 (20 Additional Marks)

In our class, we learnt that in the problem of finding large itemsets, we know how the Apriori algorithm finds large itemsets. In fact, the algorithm uses the Apriori property for this purpose. This property is that if an itemset S is large, then any proper subset of S must be large.

In fact, the Apriori algorithm could be modified for the other problem when the Apriori property is modified for the other problem.

Let us consider the other problem. Consider 20 dimensions, namely X_1, X_2, ..., X_{20}. There are 10,000 data points. Suppose that each dimension contains integer values ranging from 1 to 40. For each dimension, we divide it into 4 equal parts called *grids* or *units*. For example, dimension X_1 has 4 *grids* or *units*, namely $X_{1,1}$, $X_{1,2}$, $X_{1,3}$ and $X_{1,4}$ where $X_{1,1}$, $X_{1,2}$, $X_{1,3}$ and $X_{1,4}$ correspond to "X_1:[1, 10]", "X_1:[11, 20]", "X_1:[21, 30]" and "X_1:[31, 40]", respectively. Note that "X_1:[1, 10]" means the grid/unit representing a range between 1 and 10 in dimension X_1 and this grid is said to come from dimension X_1.

Given a set Y of dimensions and a set S of grids, S is said to be a *cell* for set Y if for each dimension X_i in Y, there exists exactly one grid in S such that this grid comes from dimension X_i.

Given a set Y of dimensions, we define A(Y) to be a set of all possible cells for set Y.

Given a cell c for a set Y of dimensions, we define d(c) to be the total number of points falling in the ranges specified in all grids of c divided by 10,000.

(a) The other problem could be **Problem 1** (to be described next).

Given a cell c, c is said to be a *dense* cell if $d(c) \geq 0.2$.

Problem 1: We want to find all possible dense cells.

Please modify the Apriori algorithm for finding all possible dense cells. In your answer, please state the Apriori property for this problem and show the correctness of this property. Besides, please elaborate how you modify the Apriori algorithm for this problem.

(b) This part is independent of Part (a).

The other problem could be **Problem 2** (to be described next).

We define $H(Y) = -\sum_{c \in A(Y)} d(c) \log d(c)$

Given a non-negative real number ω and a set Y of dimensions, Y is said to have *good clustering* if $H(Y) < \omega$.

Problem 2: We are given a value of ω in this problem. We want to find all possible sets of dimensions where each possible set Y of dimensions has good clustering.

Please modify the Apriori algorithm for finding all possible sets of dimensions where each possible set Y of dimensions has good clustering. In your answer, please state the Apriori property for this problem and show the correctness of this property. Besides, please elaborate how you modify the Apriori algorithm for this problem.

End of Paper