COMP1942 Exploring and Visualizing Data (Spring Semester 2018)
Final Examination (Question Paper)
Date: 25 May, 2018 (Fri)
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)

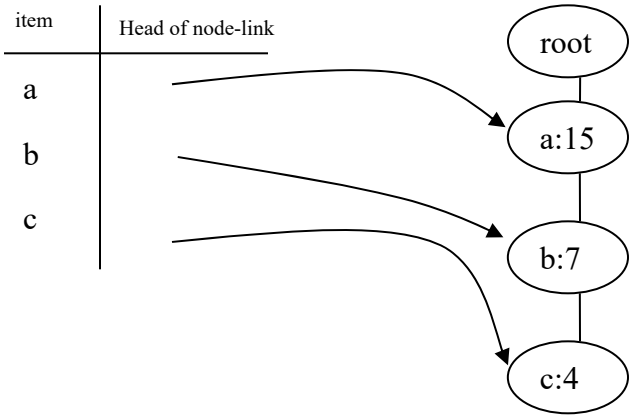
Q1 (20 Marks)

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

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

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? 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 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) There are the following four records with three binary attributes, namely A, B and C.

Tuple No.	A	B	C
1	0	0	1
2	1	1	0
3	1	1	0
4	1	0	1

Please use the monothetic approach to perform hierarchical clustering over these records. If there is any tie for choosing the attributes, we choose the attributes in the order: A, B and C.
Draw the dendrogram (without specifying the distance metric). You are required to show your steps.

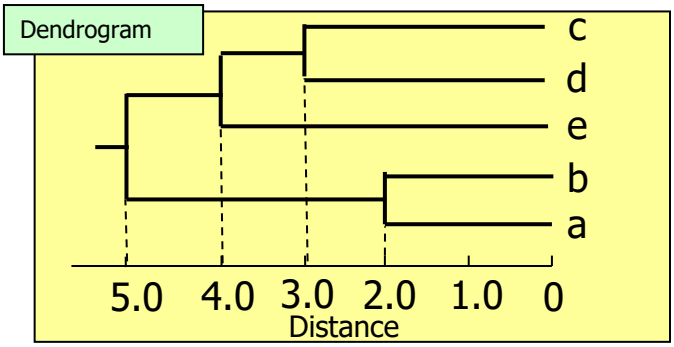
Q3 (20 Marks)

(a) We are given the following table with 3 input attributes, namely "LocalStudent", "CGA" and "Gender", and 1 target attribute, namely "LikeUST". "Actual LikeUST" corresponds to the actual values for attribute "LikeUST" and "Predicted LikeUST" corresponds to the values for attribute "LikeUST" given by a classification model (e.g., decision tree).

LocalStudent	CGA	Gender	Actual LikeUST	Predicted LikeUST
Yes	High	Male	Yes	Yes
Yes	Low	Male	No	Yes
No	High	Male	No	No
Yes	High	Female	Yes	No
Yes	Medium	Female	No	Yes
No	Medium	Female	Yes	Yes
No	Low	Female	No	No

(i) Please give the confusion matrix.
(ii)Please give the lift chart.

(b) The following shows a dendrogram for clustering five data points, namely a, b, c, d and e, based on the single linkage distance measurement.

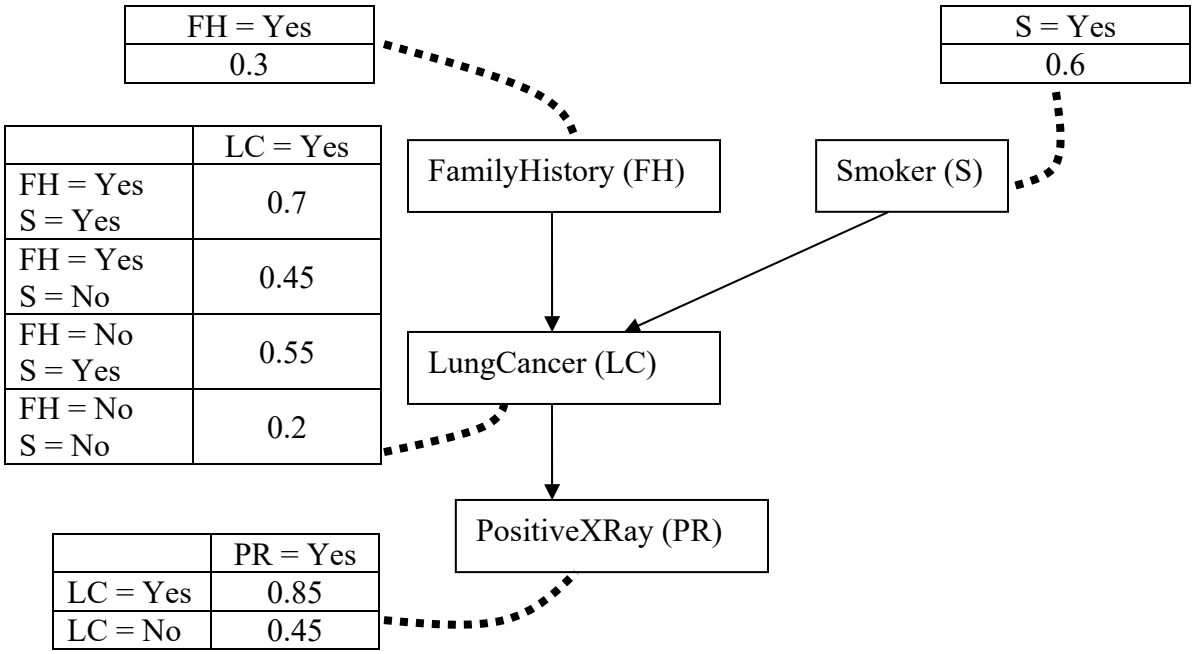


Suppose that we know that the greatest distance between a point and another point is 12 and the distance between point c and point e is 6.

Is it always true that we could draw the other dendrogram which is constructed based on the complete linkage distance measurement? If yes, please draw the dendrogram. Otherwise, please explain it.

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	?

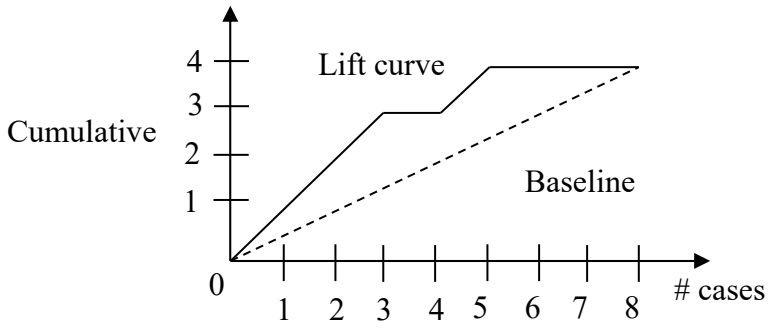
(a) Please use Bayesian Belief Network classifier with the use of Bayesian Belief Network to predict whether he is likely to have Lung Cancer.

(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)

Consider a classification problem where the target attribute contains two possible values, “Yes” and “No”.

We are given a training dataset. We generate a classifier based on this dataset. Besides, we derive the lift chart of this classifier based on this dataset where “Yes” is considered as a “success”. The chart is shown as follows.



(a) Is it possible to know the specificity of this classifier? If yes, please write down the specificity of this classifier. Otherwise, please explain it.
(b) Is it possible to know the precision of this classifier? If yes, please write down the precision of this classifier. Otherwise, please explain it.
(c) Is it possible to know the recall of this classifier? If yes, please write down the recall of this classifier. Otherwise, please explain it.
(d) Is it possible to know the f-measure of this classifier? If yes, please write down the f-measure of this classifier. Otherwise, please explain it.

Q6 (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.

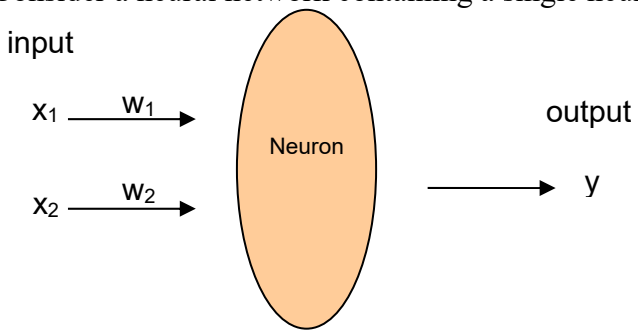
"If X is conditionally independent of Y given Z, then P(X, Y | Z) = P(X | Z) x P(Y | Z)."

(b) One reason why we need to study subspace clustering is "curse of dimensionality". What is the meaning of "curse of dimensionality"?

(c) 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 five instances by the following 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?

Q7 (20 Marks)

(a) We are given the following 4 data points: (6, 6), (8, 8), (5, 9), (9, 5). Use PCA to reduce from two dimensions to one dimension for each of these 4 data points. In this part, please show your steps.
(b) We are given the following 4 data points: (5, 5), (7, 7), (4, 8), (8, 4). Use PCA to reduce from two dimensions to one dimension for each of these 4 data points. In this part, please just write down the answer. You do not need to show your steps. Hints: You may use the answer of Part (a) for this part.
(c) We are given the following 4 data points: (18, 18), (24, 24), (15, 27), (27, 15). Use PCA to reduce from two dimensions to one dimension for each of these 4 data points. In this part, please just write down the answer. You do not need to show your steps. Hints: You may use the answer of Part (a) for this part.

Q8 (20 Marks)

(a) Data warehouse requires materializations of some views. Why do we need to materialize some views?
(b) In the view materialization method discussed in class, the constraint is given by the number of views that can be materialized. However, different views are of different sizes and the same number of views may not occupy the same amount of storage. In a more realistic problem setting, we are given available memory of size X. We still assume the equal probability of querying of each possible view. Please propose a solution for this problem setting.

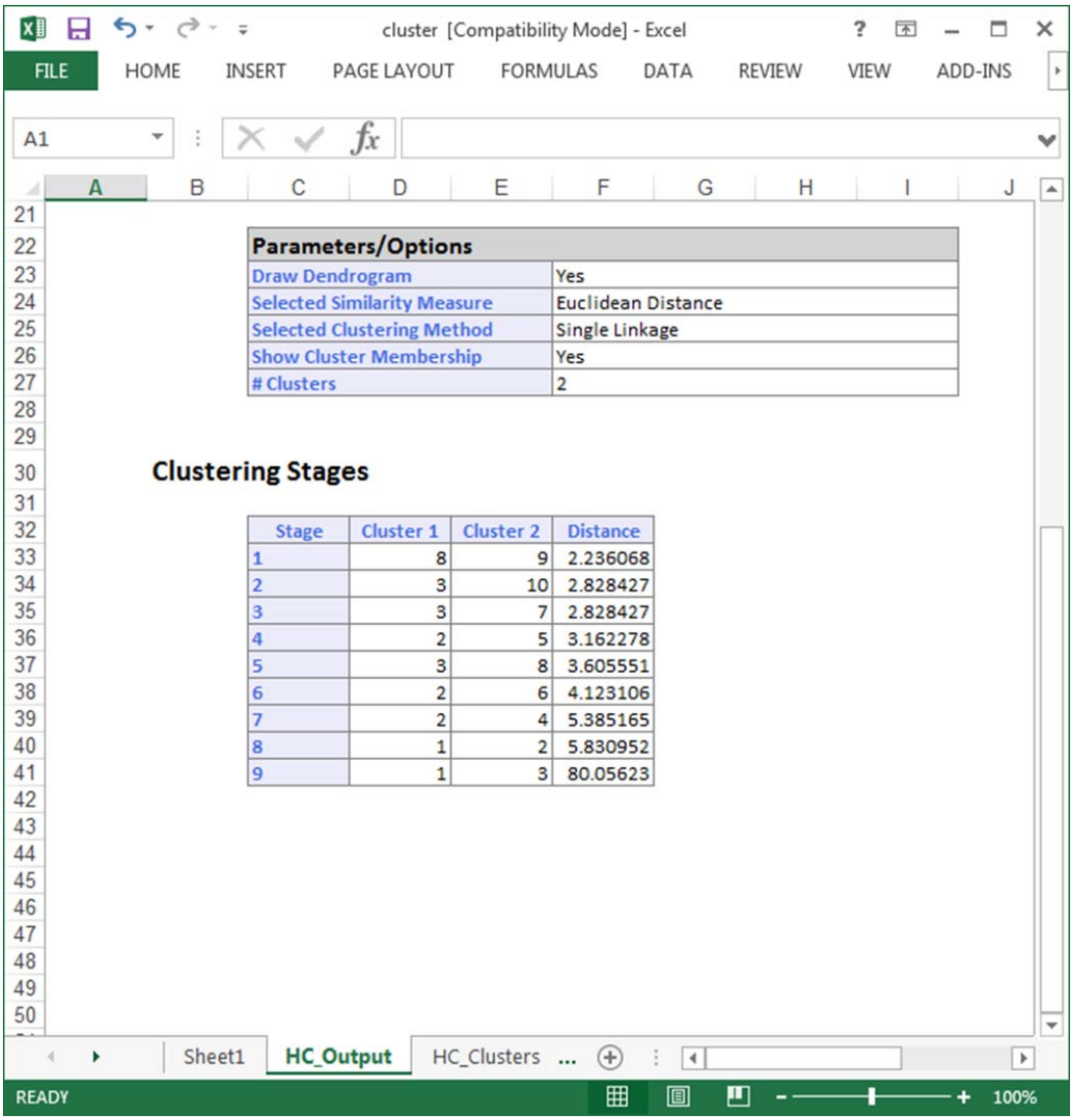
Q9 (20 Marks)

(a) We are doing clustering with a dendrogram according to a distance measurement. Given a set A of points and another set B of points, we denote the distance between A and B by D(A, B) which is calculated based on the given distance measurement.
Suppose that we have the following information.

- There are 5 data points, namely a, b, c, d, and e.
- According to this dendrogram, if we want to find two clusters, we find that the two clusters are {a, b} and {c, d, e} where the distance between these two clusters according to the distance measurement is 5.0
- D({c, d}, {e}) = 4.0, D({c}, {d}) = 3.0, and D({a}, {b}) = 2.0

Is it always true that we can draw the corresponding dendrogram? If yes, please draw the dendrogram. In this case, you are required to specify the distance metric in the dendrogram. If no, please explain what additional information we need to draw the dendrogram. In this case, please give the minimum possible sources of additional information.

(b) There are 10 data points in the dataset, namely data points 1, 2, …, 10. When we use the XLMiner software to perform "Hierarchical Clustering", we obtain the following result. In class, we learnt how to analyze the table in the result. Suppose that we want to find two clusters. Please give all data points in each of these two clusters.



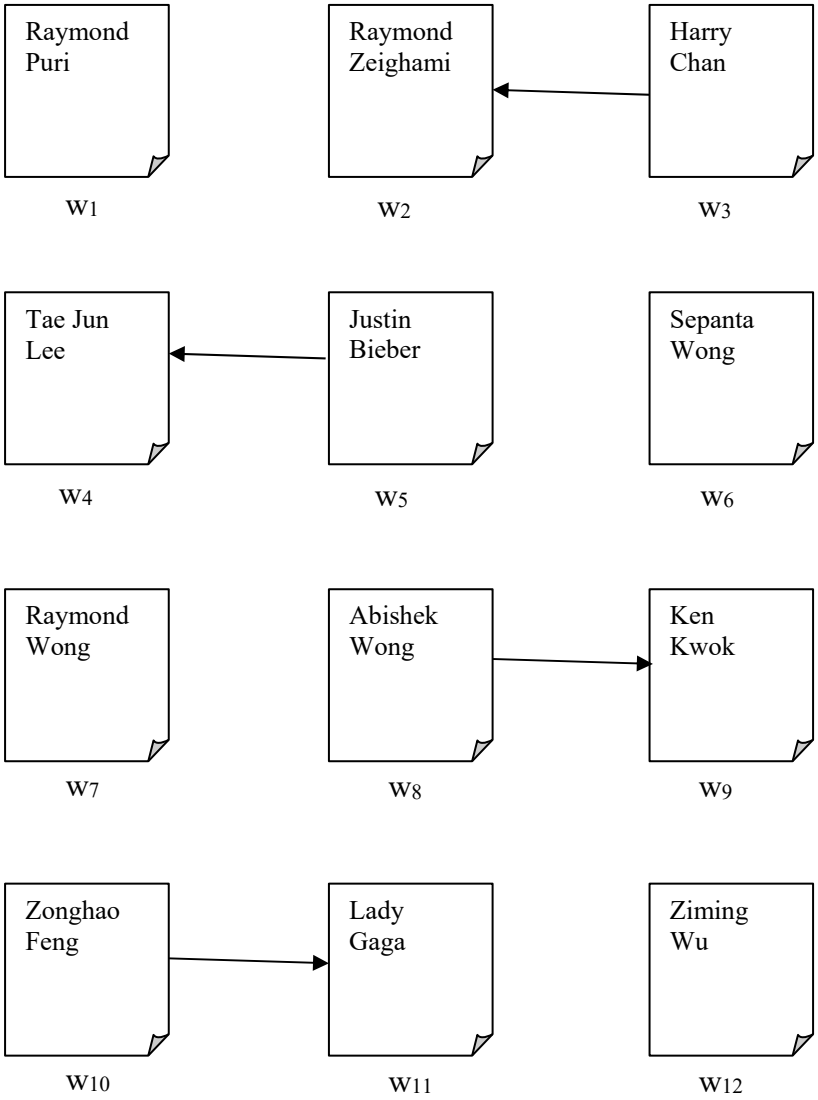
Q10 (20 Marks)

We are given the following adjacency matrix according to three sites, namely x, y and z.

$$\begin{array}{c} \\ x \\ y \\ z \end{array} \begin{array}{c} \begin{array}{ccc} x & y & z \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

(a) Is it possible to find the corresponding stochastic matrix? If yes, write down the stochastic matrix. Otherwise, please explain it.

(b) We are given the following 12 webpages, namely w_1, w_2, …, w_{12}.



The query terms typed by the user are "Raymond" and "Wong".

(i) What is the root set in this query? Please list the webpages in this set.

(ii) What is the base set in this query? Please list the webpage in this set.

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)

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}).

(a) Let A be a view. Is it always true that
Gain(V ∪ {top view}, {top view}) ≥ Gain(V ∪ {top view, view A}, {top view, view A}) ?
Please write down "yes" or "no". Then, elaborate it briefly.

(b) 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. Is it always true that
Gain({view P} ∪ {top view}, {top view}) ≥ Gain({view C} ∪ {top view}, {top view}) ?
Please write down "yes" or "no". Then, elaborate it briefly.

(c) 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.)
Is it 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}) ?
Please write down "yes" or "no". Then, elaborate it briefly.

End of Paper