COMP1942 Exploring and Visualizing Data (Spring Semester 2016)
Midterm Examination (Question Paper)
Date: 17 March, 2016 (Thu)
Time: 9:00-10:15
Duration: 1 hour 15 minutes

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.

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 |

(i) What is the support of the rule "{A, E} → C"?

(ii) What is the confidence of the rule "{A, E} → C"?

(iii)What is the lift ratio of the rule "{A, E} → C"?

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

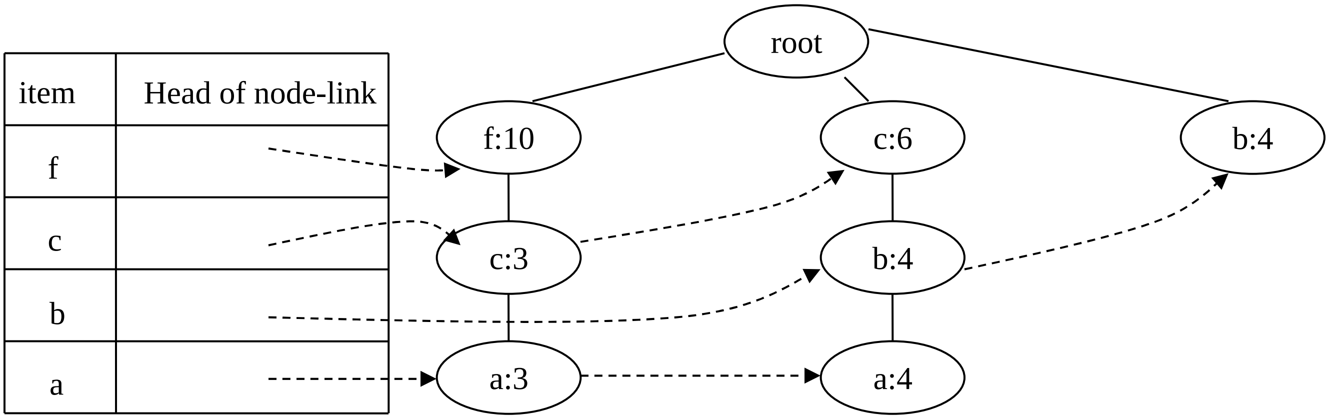
(b) This part is independent of part (a).

Suppose that we are also given another dataset with some transactions in binary format, and the support threshold = 2. Finally, we obtain the set S of all frequent itemsets equal to

{ {A}, {B}, {C}, {D},
{A, B}, {A, C}, {A, D}, {B, C}, {B, D},
{A, B, C}, {A, B, D} }

There are many possible datasets which have the same set S as the set of all frequent itemsets.
Please give one possible dataset which has the minimum number of transactions in binary format.
Assume that each transaction in this dataset contains A, B, C, D or E.

**Q2 (20 Marks)**



Consider the above FP-tree T and given the support threshold of 3.
Apply the algorithm of FP-growth(T, NULL) and generate all the conditional FP-trees.
What are the frequent patterns generated?

**Q3 (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) We are given five data points.

a: (1, 2), b: (2, 4), c: (7, 6), d: (6, 9), e: (8, 9)

Suppose that there are two clusters. The first cluster contains points a and b while the second cluster contains points c, d and e.

(i) (1) What is the center of the first cluster if we use the centroid linkage as a distance measurement?

(2) What is the center of the second cluster if we use the centroid linkage as a distance measurement?

(ii) Consider the agglomerative approach for hierarchical clustering.

Suppose that these two clusters are merged.

(1) What is the center of the merged cluster if we use the centroid linkage as a distance measurement?

(2) What is the center of the merged cluster if we use the median linkage as a distance measurement?

**Q4 (20 Marks)**

(a) Consider Algorithm original k-means clustering. Consider that at the beginning, k means are randomly chosen (not necessarily equal to the data points). At the beginning, we have these k means, and each data point belongs to one of the k clusters with these means. Consider a cluster. Is it always true that if there exists at least one data point which belongs to this cluster at the beginning, there at least one data point which belongs to this cluster after a lot of iterations (where each iteration corresponds to an update on k means in the algorithm)? If yes, please show the correctness of this statement. Otherwise, please give an example showing that this statement is incorrect.

(b) 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.)

**Q5 (20 Marks)**

The following shows a history of users with attributes “Has_Bicycle”, “Age” and “Income”. We also indicate whether they will buy a hoverhoard or not in the last column. You cannot use XLMiner in this question.

| No. | Has_Bicycle | Age | Income | Buy_Hoverboard |
|---|---|---|---|---|
| 1 | no | young | fair | yes |
| 2 | no | young | high | yes |
| 3 | yes | old | fair | yes |
| 4 | yes | middle | fair | yes |
| 5 | no | young | fair | no |
| 6 | no | middle | low | no |
| 7 | yes | old | low | no |
| 8 | yes | young | low | no |

(a) We want to train a C4.5 decision tree classifier to predict whether a new user will buy a hoverboard or not. We define the value of attribute Buy_Hoverboard to be the *label* of a record.

(i) Please find a C4.5 decision tree according to the above example. 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.

(ii) Consider a new young user with a bicycle whose income is fair. Please predict whether this new user will buy a hoverboard or not.

(b) What is the difference between the C4.5 decision tree and the ID3 decision tree? Why is there a difference?

# Part B (Bonus Question)

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

**Q6 (10 Additional Marks)**
We are given four items, namely A, B, C and D. Their corresponding unit profits are $p_A$, $p_B$, $p_C$ and $p_D$.

The following shows five transactions with these items. Each row corresponds to a transaction where a non-negative integer shown in the row corresponds to the total number of occurrences of the correspondence item present in the transaction.

| A | B | C | D |
|---|---|---|---|
| 0 | 0 | 3 | 2 |
| 3 | 4 | 0 | 0 |
| 0 | 0 | 1 | 3 |
| 1 | 0 | 3 | 5 |
| 6 | 0 | 0 | 0 |

The frequency of an itemset in a row is defined to be the minimum of the number of occurrences of all items in the itemset. For example, itemset {C, D} in the first row has frequency = 2. But, itemset {C, D} in the third row has frequency = 1.

The frequency of an itemset in the dataset is defined to be the sum of the frequencies of the itemset in all rows in the dataset. For example, itemset {C, D} has frequency = 2+0+1+3+0 = 6.

Define a function f on an itemset s. This function will be specified later. One example of this function is $f(s) = \sum_{i \in s} p_i$. In this example, if s = {C, D}, then $f(s) = p_C + p_D$.

The profit of an itemset s in the dataset is defined to be the product of the frequency of this itemset in the dataset and f(s).

For example, itemset {C, D} has profit = $6 \cdot f(\{C, D\})$

(a) Assume that we adopt function f such that $f(s) = (\sum_{i \in s} p_i)/|s|$ where |s| denotes the no. of items in s.
Suppose that we know that $p_A = 10$, $p_B = 10$, $p_C = 10$ and $p_D = 10$.
We want to find all itemsets with profit at least 50.
Can the Apriori Algorithm be adapted to find these itemsets?
If yes, please write down the pseudo-code (or steps) and illustrate it with the above example.
If no, please explain why the Apriori Algorithm cannot be adapted. In this case, please also design an algorithm, write down the pseudo-code and illustrate it with the above example.

(b) Assume that we adopt function f such that $f(s) = \sum_{i \in s} p_i$.
Suppose that we know that $p_A = 5$, $p_B = 10$, $p_C = 6$ and $p_D = 4$.
We want to find all itemsets with profit at least 50.
Can the Apriori Algorithm be adapted to find these itemsets?
If yes, please write down the pseudo-code (or steps) and illustrate it with the above example.
If no, please explain why the Apriori Algorithm cannot be adapted. In this case, please also design an algorithm, write down the pseudo-code and illustrate it with the above example.

**End of Paper**