COMP1942 Exploring and Visualizing Data (Spring Semester 2013)
Midterm Examination (Question Paper)
Date: 22 March, 2013 (Fri)
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 and Part B in the **answer sheet**.
(2) You can **optionally** answer the bonus question in Part C 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 = 3.

| P | Q | R | S | T |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 |

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

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

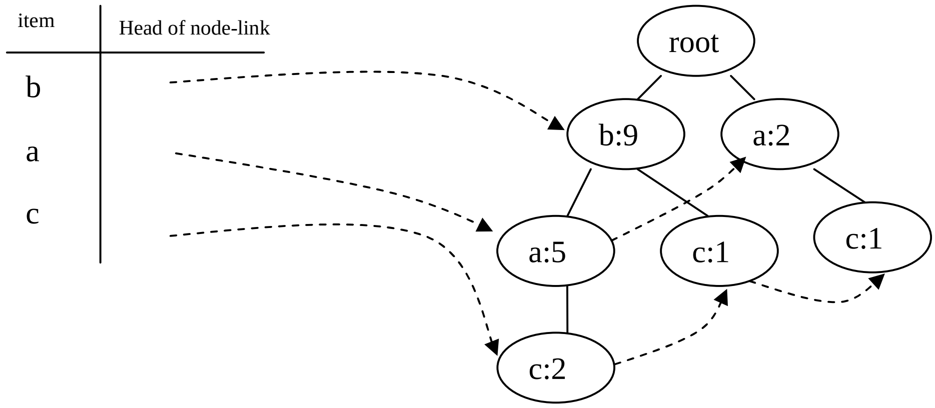
(iii) What is the lift ratio of the rule "{P, T} → R"?

(iv) Consider the Apriori algorithm which generates sets $C_2$ and $C_3$.

(1) After we execute the Join Step and the Prune Step, what is the content of $C_2$?

(2) After we execute the Join Step and the Prune Step, what is the content of $C_3$?

(b) The following shows an FP-tree which is constructed from a set of transactions. Let the support threshold be 1. Please write down the corresponding transactions which are used to generate the FP-tree.



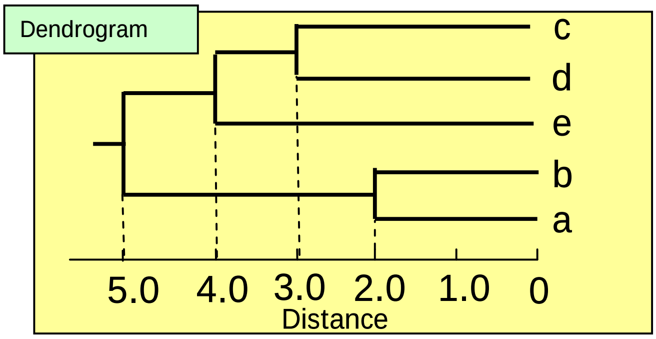
**Q2 (20 Marks)**

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

| Tuple No. | A | B | C |
|---|---|---|---|
| **1** | 1 | 1 | 0 |
| **2** | 1 | 1 | 0 |
| **3** | 1 | 0 | 1 |
| **4** | 0 | 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.

(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 possible to draw the other dendrogram which is constructed based on the complete linkage distance measurement? If yes, please draw the dendrogram. Otherwise, please explain it.

**Q3 (20 Marks)**

(a) Consider Algorithm sequential k-means clustering.
When it reads a data point x, it will update the mean m of a cluster with the following operation.

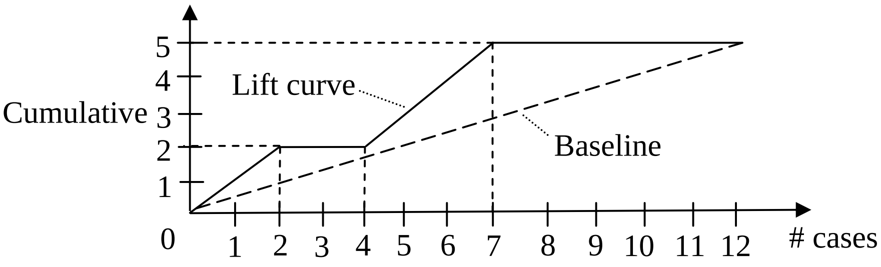
m ← m + 1/n (x - m)

where n is the size of the cluster including the new data point x.

(i) Please write down the steps for Algorithm sequential k-means clustering.

(ii) Please prove that, with the above operation, the mean m is calculated correctly. That is, the mean m calculated is equal to the expected vector among all data points in the cluster.
(Hints: Let $x_j$ be the j-th example in cluster i and $m_i(t)$ be the mean vector of cluster i containing t examples. Consider that x is the t-th example in cluster i. Note that $m_i(t) = \frac{x_1 + x_2 + ... + x_{t-1} + x_t}{t}$.)

(b) We are given the following lift chart for a decision tree.



Is it possible to know the error report for this decision tree according to this lift chart? If yes, please give the error report. If no, please explain it.

**Q4 (20 Marks)**

The following shows a history of customers with their incomes, ages and an attribute called “Have_iPhone” indicating whether they have an iPhone. We also indicate whether they will buy an iPad or not in the last column.

| No. | Income | Age | Have_iPhone | Buy_iPad |
|---|---|---|---|---|
| 1 | high | young | yes | yes |
| 2 | high | old | yes | yes |
| 3 | medium | young | no | yes |
| 4 | high | old | no | yes |
| 5 | medium | young | no | no |
| 6 | medium | young | no | no |
| 7 | medium | old | no | no |
| 8 | medium | old | no | no |

We want to train a CART decision tree classifier to predict whether a new customer will buy an iPad or not. We define the value of attribute Buy_iPad is the *label* of a record.

(a) 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.
(Hints: Even if attribute Income contains 3 possible values, namely “high”, “medium” and “low” (instead of the 2 values which appear in the above table (i.e., “high” and “medium”)), then the information gain we should calculate is just based on all values available in the table (i.e., 2 values in this case). This principle can also be applied to other attributes.)

(b) Consider a new young customer whose income is medium and he has an iPhone. Please predict whether this new customer will buy an iPad or not.

# Part B (Compulsory Multiple-Choice (MC) Questions)

In this part, there are 4 multiple-choice questions, namely Q5, Q6, Q7 and Q8. The total scores in this part are 20 scores. Each question weighs 5 scores.

Q5. [Removed]

A. [Removed]
B. [Removed]
C. [Removed]
D. [Removed]
E. [Removed]

Q6. [Removed]

A. [Removed]
B. [Removed]
C. [Removed]
D. [Removed]
E. [Removed]

Q7. [Removed]

A. [Removed]
B. [Removed]
C. [Removed]
D. [Removed]
E. [Removed]

Q8. [Removed]

A. [Removed]
B. [Removed]
C. [Removed]
D. [Removed]
E. [Removed]

# Part C (Bonus Question)

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

**Q9 (10 Additional Marks)**

(a) We are given two customers, namely X and Y. The following shows 5 transactions for these two customers. Each transaction contains three kinds of information: (1) customer ID (e.g., X and Y), (2) the time that this transaction occurred, and (3) all the items involved in this transaction.

Customer X, time 1, items A, B, C
Customer Y, time 2, items A, F
Customer X, time 3, items D, E
Customer X, time 4, item G
Customer Y, time 5, items D, E, G

For example, the first transaction corresponds to that customer X bought item A, item B and item C at time 1, while the last transaction corresponds to that customer Y bought item D, item E and item G at time 5.

A sequence is defined to be a series of itemsets in form of $<S_1, S_2, S_3, \ldots, S_m>$ where $S_i$ is an itemset for i = 1, 2, …, m. The above transactions can be transformed into two sequences as follows.

X: <{A, B, C}, {D, E}, {G}>
Y: <{A, F}, {D, E, G}>

After this transformation, each customer is associated with a sequence.

Given a sequence S in form of $<S_1, S_2, S_3, \ldots, S_m>$ and another sequence S' in form of $<S_1', S_2', S_3', \ldots, S_n'>$ , S is said to be a subsequence of S' if $m \leq n$ and there exist m integers, namely $i_1, i_2, \ldots, i_m$, such that (i) $1 \leq i_1 < i_2 < \ldots < i_m \leq n$, and (2) $S_j \subseteq S_{i_j}'$ for j = 1, 2, …, m. If S is a subsequence of S', then S' is defined to be a super-sequence of S.
The support of a sequence S is defined to be the total number of customers which sequences are super-sequences of S.

Given a positive integer k, a sequence in form of $<S_1, S_2, S_3, \ldots, S_m>$ is said to be a k-sequence if $\sum_{i=1}^{m} |S_i|=k$.

(Note that $|S_i|$ denotes the total number of elements in set $S_i$.)

Can the Apriori algorithm be adapted to mining all k-sequences with support at least 2 where k = 2, 3, 4, …. ? If yes, please write down the proposed method using the concept of the Apriori algorithm and illustrate your algorithm with the above example. If no, please explain the reason.

(b) We want to study the same problem described in (a). However, the support of a sequence is defined in another way. Now, the support of a sequence S is re-defined to be the total number of all possible occurrences of S in all customers divided by the total number of customers which sequences are super-sequences of S. An occurrence of a sequence S (in form of $<S_1, S_2, S_3, \ldots, S_m>$) in a customer who has his/her sequence S' ($<S_1', S_2', S_3', \ldots, S_n'>$) corresponds to one possible set of m integers, namely $i_1, i_2, \ldots, i_m$, such that (i) $1 \leq i_1 < i_2 < \ldots < i_m \leq n$, and (2) $S_j \subseteq S_{i_j}'$ for j = 1, 2, …, m. Note that if S is a subsequence of S', it is possible that there are multiple possible sets of m integers (or multiple possible occurrences). For example, suppose that there is a sequence S' for a customer as <{A}, {B}, {B}, {C}>. Consider a sequence S = <{A}, {B}, {C}>. There are two possible occurrences of S in this customer (and the corresponding two possible sets of integers are {1, 2, 4} and {1, 3, 4}).

Can the Apriori algorithm be adapted to mining all k-sequences with support at least 2 where k = 2, 3, 4, …. with this new definition of support? If yes, please write down the proposed method using the concept of the Apriori algorithm and illustrate your algorithm with the above example. If no, please explain the reason.

**End of Paper**