COMP192 Exploring and Visualizing Data (Spring Semester 2011)
Midterm Examination (Question Paper)
Date: 25 March, 2011 (Fri)
Time: 13:35-14:35
Duration: 1 hour

Student ID:_____ Student Name:_____

Seat No. :_____

Instructions:
(1) Please answer **all** questions in the **answer sheet**.
(2) You can use a calculator.

Question Paper

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.

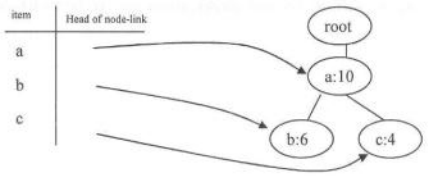
(b) This part is independent of part (a).
Suppose that we are also given another dataset with some transactions in binary format, and the support threshold = 2. Finally, we obtain the set S of all frequent itemsets equal to

{ {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.

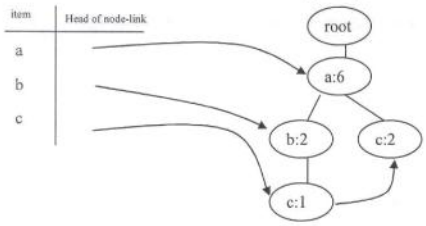
Q2 (20 Marks)

(a) The following shows an FP-tree which is constructed from a set of transactions. Let the support threshold be 4.



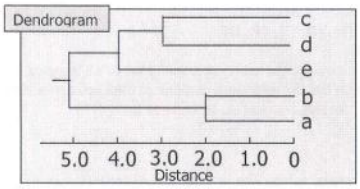
(i) Please draw the conditional FP-tree on c.
(ii) Please draw the conditional FP-tree on b.
(iii) Please draw the conditional FP-tree on a.
(iv) Please list all frequent itemsets. You do not need to give the frequency of each frequent itemset.

(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.



Q3 (20 Marks)

(a) The following shows a dendrogram for clustering five data points, namely a, b, c, d and e.



Suppose that we are interested in finding two clusters. For each cluster, please state all data points in this cluster.

(b) There are four two-dimensional points:

x_1 (1, 2),
x_2 (2, 2),
x_3 (5, 2),
x_4 (6, 1).

(i) Assume that we adopt the Euclidean distance metric as the distance between any two points. Please write a matrix where its entries correspond to the pairwise distances between any two points.
(ii) Please use the agglomerative approach to group these points with distance median linkage. Draw the corresponding dendrogram for the clustering. You are required to specify the distance metric in the dendrogram.

Q4 (20 Marks)

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.

$$m \leftarrow m + 1/n\ (x - m)$$

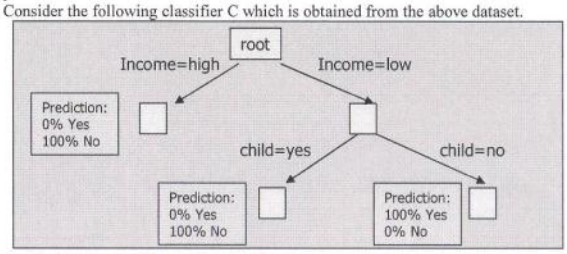
where n is the size of the cluster including the new data point x.
(a) Please write down the pseudo-code for Algorithm sequential k-means clustering.
(b) Please prove that, with the above operation, the mean m is calculated correctly. That is, the mean m calculated is equal to the expected vector among 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 $m_i(t) = \frac{x_1 + x_2 + \ldots + x_{t-1} + x_t}{t}$.)

Q5 (20 Marks)

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	low	no	no

(a) 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) Consider the following classifier C which is obtained from the above dataset.



(i) What is the lift chart for classifier C?
(ii) What is the decile-wise lift chart for classifier C?
(iii)Is C a good classifier? Why? Please explain it by using the decile-wise lift chart in part (ii).

End of Paper