COMP1942 Exploring and Visualizing Data (Spring Semester 2017)
Midterm Examination (Question Paper)
Date: 17 March, 2017 (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 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 set of transactions with a certain number of items.
Consider association rule mining where the confidence threshold is 50% and the support threshold is 3.
Let S_0 be the set of all possible association rules with their confidence at least 50% and their support at least 3. In other words, S_0 is the set of all association rules we want to find.

In the class, we learnt the following two-step method of generating a set of association rules.
The description of the two-step method is given as follows.

- **Step 1:** To generate a set S_1 of all itemsets with their support at least 3
- **Step 2:** For any two itemsets in S_1, namely X and Y, where $X \subseteq Y$,
  if $\text{supp}(Y)/\text{supp}(X) \geq 50\%$, ……..(*)
  generate an association rule in the form of "$X \rightarrow Y - X$"
  Let S_2 be the set of all association rules generated in this step.

In the class, we study the following two claims.

- **Claim 1**: Each association rule in S_2 has its support at least 3.
- **Claim 2**: For each association rule in S_0, it is in S_2.

(a) Consider that the original Step 1 of the two-step method is changed to
"To generate a set S_1 of all itemsets with their support at least 2"
(i.e., number "3" is changed to number "2" in Step 1).

(i) Is it always true that Claim 1 is correct? If the answer is "yes", please show the correctness of the following "simplified" form of Claim 1 where B and C are two items (similar to the form shown in the class):
If "$B \rightarrow C$" is in S_2, the support of "$B \rightarrow C$" is at least 3.
If the answer is "no", please give a concrete example containing the *smallest* possible number of transactions and illustrate with this example that "$B \rightarrow C$" is in S_2 but the support of "$B \rightarrow C$" is smaller than 3.

(ii) Is it always true that Claim 2 is correct? If the answer is "yes", please show the correctness of the following "simplified" form of Claim 2 where B and C are two items (similar to the form shown in the class):
If "$B \rightarrow C$" is in S_0, it is in S_2.
If the answer is "no", please give a concrete example containing the *smallest* possible number of transactions and illustrate with this example that "$B \rightarrow C$" is in S_0 but "$B \rightarrow C$" is not in S_2.

(b) Consider that the original Step 1 of the two-step method is changed to
"To generate a set S_1 of all itemsets with their support at least 4"
(i.e., number "3" is changed to number "4" in Step 1).
(i) Is it always true that Claim 1 is correct? Please elaborate it following the instruction in (a)(i).
(ii) Is it always true that Claim 2 is correct? Please elaborate it following the instruction in (a)(ii).

Q2 (20 Marks)

Given the following transactions and the support threshold = 2.

TID	Items bought
1	b, c, d, p
2	f, j, q
3	c, i
4	a, d
5	c, m
6	b, d, f
7	a, d, f
8	e, f
9	f, h
10	b, d
11	a, l
12	c, g
13	c, k
14	f, n, o

Follow the steps of the FP-growth algorithm to find all frequent itemsets.
Please show the steps (including showing all conditional FP-trees generated and showing all frequent itemsets generated based on each conditional FP-tree) and list all frequent itemsets at the end. In the list of frequent itemsets, you are not required to give the frequency of each frequent itemset.

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.

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

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

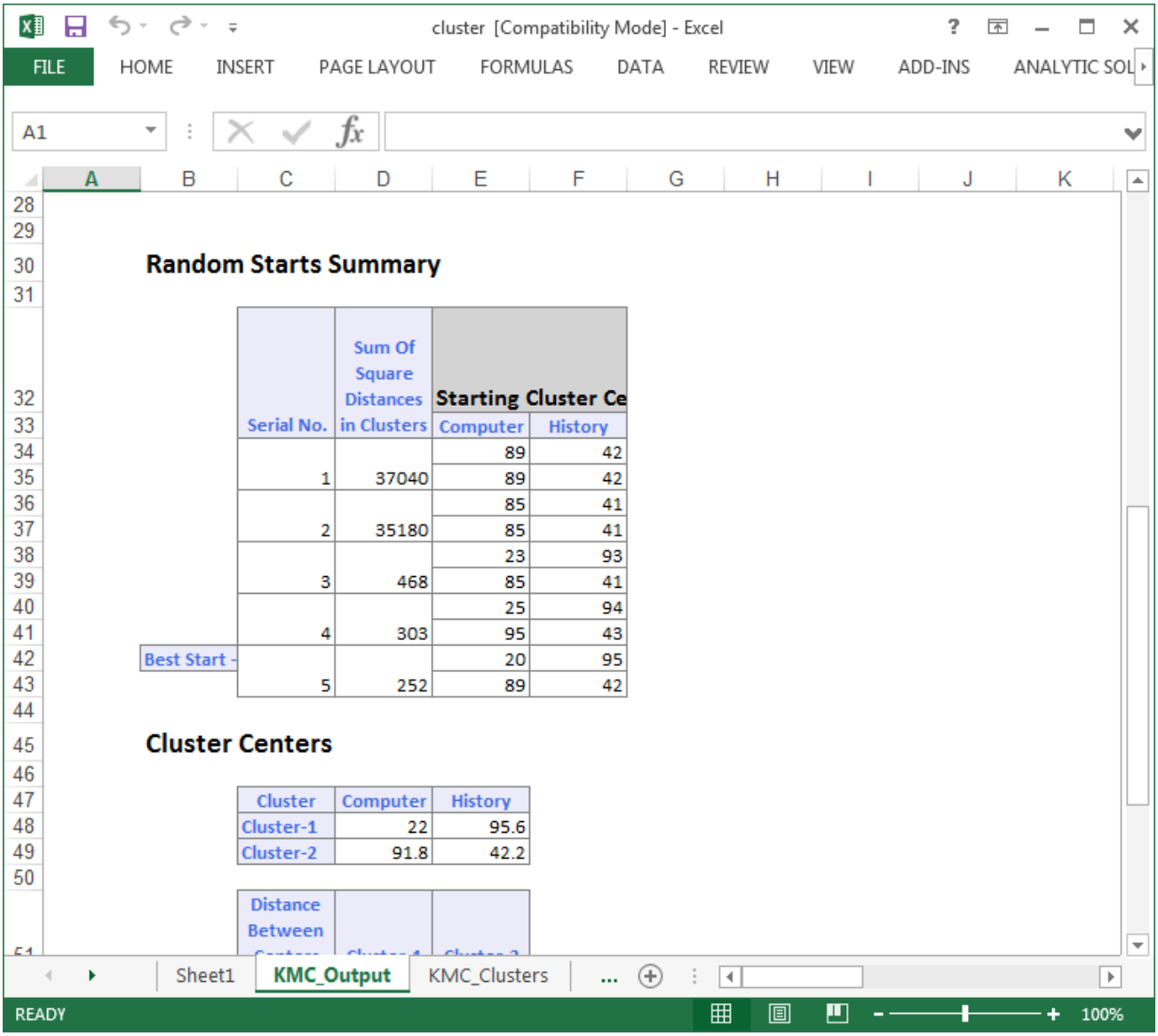
(i) Please write down the steps for Algorithm sequential k-means clustering.

(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.

(b) The following shows the output generated by XLMiner for k-means clustering. As we know, the k-means algorithm performs some operations and obtains multiple clustering results. But, it *finally* uses one clustering result. Please state the following based on the following output.

(i) the total number of clusters used *finally*, and

(ii) the final mean calculated and the corresponding initial randomly generated mean for each cluster used *finally*.



Q4 (20 Marks)

There are four two-dimensional points:

x_1 (2, 3),
x_2 (3, 3),
x_3 (6, 3),
x_4 (7, 2).

(a) 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.
(b) 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.

Q5 (20 Marks)

The following shows a history of customers with their ages, incomes and credit ratings. We also indicate whether they will buy an apple watch or not in the last column.

No.	Age	Income	Credit_Rating	Buy_AppleWatch
1	young	high	high	yes
2	young	high	fair	yes
3	old	medium	fair	yes
4	old	low	fair	yes
5	young	medium	low	no
6	young	high	fair	no
7	old	low	low	no
8	old	high	low	no

(a) We want to train a C4.5 decision tree classifier to predict whether a new customer will buy an apple watch or not. We define the value of attribute Buy_AppleWatch 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 70% 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 old customer whose income is high but his credit rating is fair. Please predict whether this new customer will buy an apple watch 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)

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.

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 $H(Y) < \omega$.

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

[Scrap Paper]

[Scrap Paper]

[Scrap Paper]

[Scrap Paper]