

COMP1942 Exploring and Visualizing Data (Spring Semester 2018)

Midterm Examination (Question Paper)

Date: 16 March, 2018 (Fri)

Time: 3:15pm-4:15pm

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 confidence of the rule “{A, D}  $\rightarrow$  E”?
- (ii) 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).

Consider a data set containing 10 transactions and 6 items.

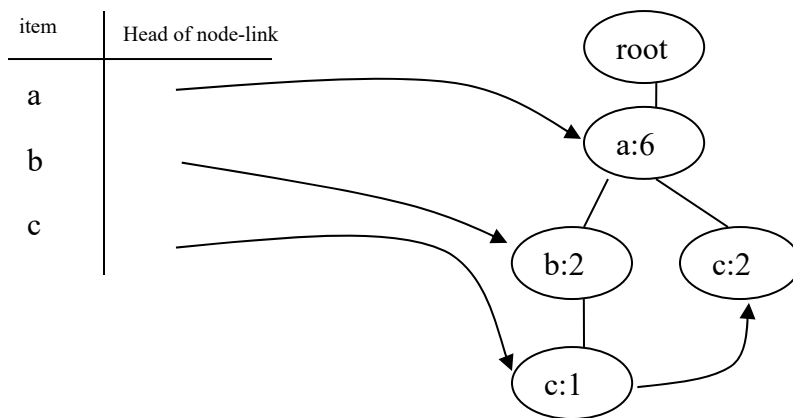
We know that the lift ratio of association rule “{A, B}  $\rightarrow$  C” is 1.25.

We also know that the support of {A} is 7, the support of {B} is 5, the support of {C} is 6, the support of {A, B} is 4, the support of {A, C} is 5 and the support of {B, C} is 3.

Is it always true that we can find the support of “{A, B}  $\rightarrow$  C”? If yes, please explain it and write down the support of “{A, B}  $\rightarrow$  C”. Otherwise, please elaborate it.

**Q2 (20 Marks)**

(a) 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.



(b) This part is independent of Part (a). We are given a support threshold greater than 1. We know that conditional FP-trees are constructed from an FP-tree. Is it always true that we can construct the FP-tree based on all conditional FP-trees constructed? Please elaborate it.

(c) This part is independent of Part (a) and Part (b).

In the Apriori algorithm, we know how to find some sets  $L_1, C_2, L_2, \dots$

- (i) Is it always true that the number of itemsets in  $L_2$  is smaller than or equal to the number of itemsets in  $C_2$ ? If yes, please explain it. Otherwise, please give a counter example.
- (ii) Is it always true that the number of itemsets in  $C_2$  is larger than or equal to the number of itemsets in  $L_1$ ? If yes, please explain it. Otherwise, please give a counter example.

**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.

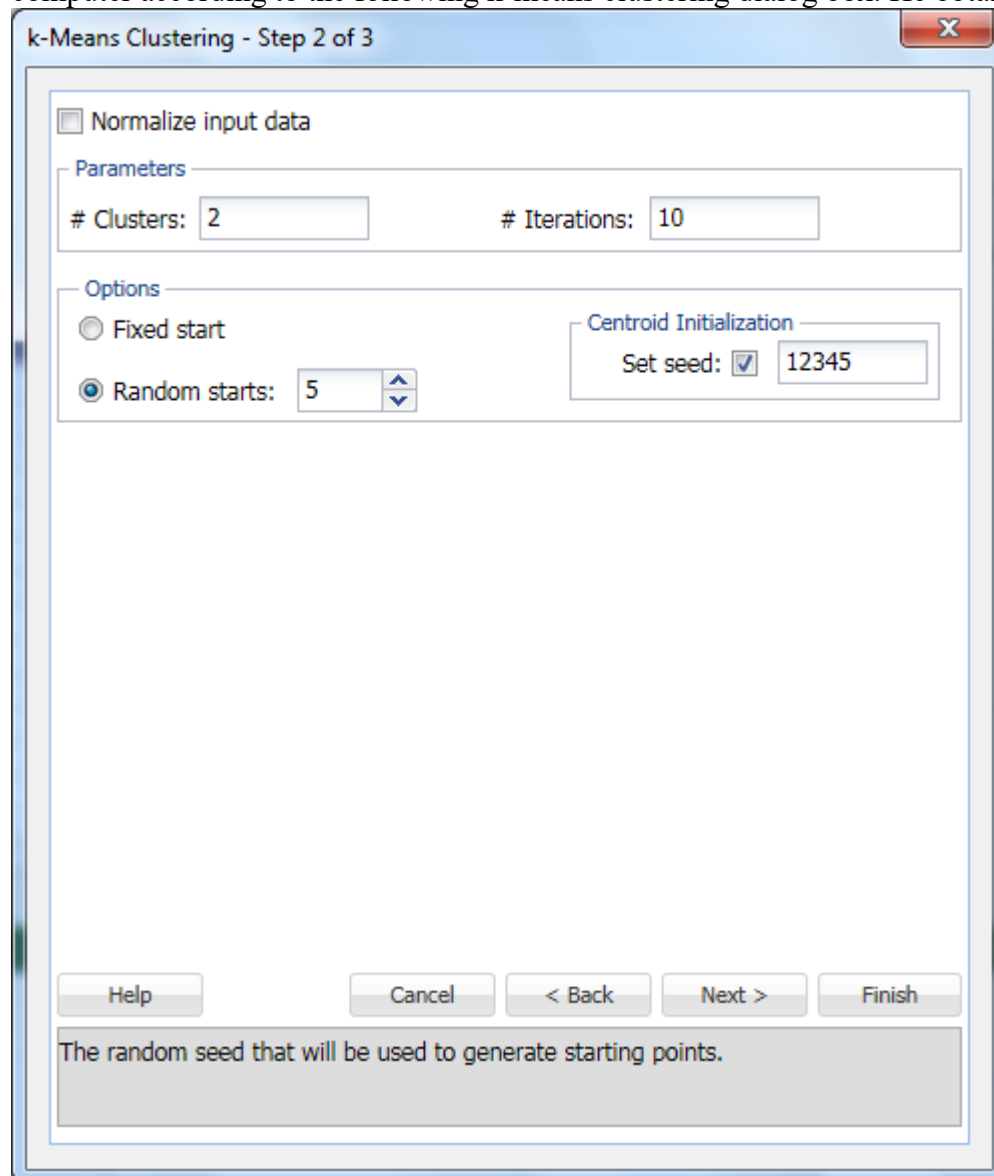
$$m \leftarrow m + 1/n (x - m)$$

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.  
(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  $m_i(t) = \frac{x_1 + x_2 + \dots + x_{t-1} + x_t}{t}$ .)

- (b) Consider the following two scenarios.

**Scenario 1:** Suppose that a user set some parameters for k-means clustering in XLMiner in his home computer according to the following k-means clustering dialog box. He obtained a clustering result R1.



**Scenario 2:** Suppose that the same user set some parameters for association rule mining in XLMiner in his home computer and obtained the following result R2.

**XLMiner : Association Rules**

**Output Navigator**

[Inputs](#) [List of Rules](#)

**Inputs**

Data	
# Transactions in Input Data	5
# Columns in Input Data	5
# Items in Input Data	5
# Association Rules	6
Minimum Support	3
Minimum Confidence	50.00%

**List of Rules**

Rule: If all Antecedent items are purchased, then with Confidence percentage Consequent items will also be purchased.

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	100	D	A	3	3	3	1.666666667
2	100	A	D	3	3	3	1.666666667
3	100	C	B	3	4	3	1.25
4	100	E	B	3	4	3	1.25
5	75	B	C	4	3	3	1.25
6	75	B	E	4	3	3	1.25

- In Scenario 1, if the user sets the same parameters for k-means clustering in XLMiner in his friend's computer according to the above k-means clustering dialog box, is it a must that he will obtain the same clustering result as R1? Please elaborate it.
- In Scenario 2, if the user changes the minimum support threshold of a rule to 4 only, is it a must that he will obtain the same association rule result as R2? Please elaborate it.
- In Scenario 2, if the user changes the minimum confidence threshold of a rule to 70% only, is it a must that he will obtain the same association rule result as R2? Please elaborate it.

**Q4 (20 Marks)**

(a) We are given the following 5 data points.

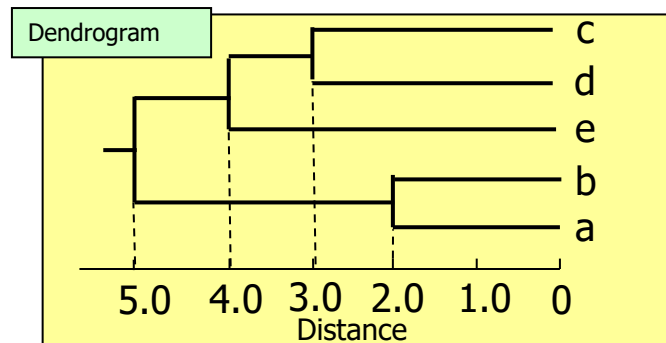
$$x_1: (1, 2), x_2: (4, 5), x_3: (3, 2), x_4: (10, 11), x_5: (12, 13)$$

Given a distance measurement between two clusters (e.g., group average linkage), we want to obtain a dendrogram based on the agglomerative approach. We want to find 2 clusters. Based on this dendrogram, we know that the first cluster contains  $x_1$ ,  $x_2$  and  $x_3$  and the second cluster contains  $x_4$  and  $x_5$ . In each of the following answers, please round the decimal number up to 2 decimal places (e.g., 4.12578 should be rounded to 4.13 in the final answer). In the following, it is sufficient to write down the final answer. If your final answer is correct, you will obtain full scores for that part. Otherwise, showing steps allows you to obtain partial “step” scores.

- (i) Please find the distance between these 2 clusters based on the distance measurement of “group average linkage”.
- (ii) Please find the distance between these 2 clusters based on the distance measurement of “centroid linkage”.
- (iii) Please find the distance between these 2 clusters based on the distance measurement of “median linkage”.

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

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.

**Q5 (20 Marks)**

The following shows a history of students with attributes “Gender”, “Year” and “School”. We also indicate whether they will study COMP1942 or not in the last column.

No.	Gender	Year	School	Study COMP1942
1	male	one	SHSS	yes
2	male	one	SSCI	no
3	male	two	SHSS	yes
4	male	three	SSCI	no
5	female	three	SHSS	yes
6	female	three	SHSS	no
7	female	three	SENG	yes
8	female	two	SSCI	no

- (a) We want to train a C4.5 decision tree classifier to predict whether a new student will study COMP1942 or not. We define the value of attribute Study\_COMP1942 to be the *label* of a record.
- 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.
  - Consider a new Year 3 male student whose school is SHSS. Please predict whether this new student will study COMP1942 or not.
- (b) What is the difference between the C4.5 decision tree and the ID3 decision tree? Why is there a difference?

**End of Paper**

[Scrap Paper]

[Scrap Paper]



[Scrap Paper]

[Scrap Paper]