53 (IT 701) DMDW

2018

DATA MINING AND DATA WAREHOUSING

Paper : IT 701

Full Marks : 100

Time : Three hours

The figures in the margin indicate full marks for the questions.

*Answer **any five** questions.*

1. *(a)* How is a data warehouse different from a database? How are they similar?

   (b) Describe the steps involved in data mining when viewed as a process of knowledge discovery. 20

2. Briefly compare the following concepts. You may use an example to explain your points. 20

   (a) Snowflake schema, fact constellation, starnet query model.

   (b) Data cleaning, data transformation, refresh.

3. *(a)* In realworld data, tuples with missing values for some attributes are a common occurrance. Describe various methods for handling this problem.

(b) Consider the following data (in increasing order) for the attribute age : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. 20

(i) Use smoothing by bin means to smooth this data, using a bin depth of three. Illustrate your steps, comment on the effect of this technique for the given data.

(ii) How might you determine outliers in the data ?

(iii) What other methods are there for data smoothing ?

4. A database has five trasactions. Let min sup D 60% and min conf D 80%

TID	**items bought**
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

Find all the frequent item sets using Apriori and FP-growth respictively. Compare the efficiency of the two mining process. 20

5. *(a)* Briefly outline the major steps of decision tree classification.

(b) Why naive Bayesian classification is called 'naive' ? Briefly outline the major ideas of naive Bayesian classification. 20

6. *(a)* Briefly describe and give examples of each of the following approaches to clustering : partitioning methods, hierarchical methods, density based methods and grid-based methods.

(b) Suppose that the data mining task is to a cluster points (with (x, y)representing location) into three cluster, where the points are

A_1 (2,10), A_2 (2,5), A_3 (8,4), B_1(5,8), B_2(7,5), B_3 (6,4), C_1 (1,2), C_2 (4,9).

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, C_1 as the center of each cluster, respectively. Use the *K*-means algorithim to show only— 20

(a) The three cluster centers after the first round of execution

(b) The final three clusters.

7. Draw the decision tree for the following data sets. Use entropy as a node selection mechanism. 20

RID	age	income	student	credit rating	class : buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no