PVP 12

Code: **CSCS2T1**

I M.Tech - II Semester - Special Supplementary Examinations
March 2019

DATA WAREHOUSING AND DATA MINING
(COMPUTER SCIENCE & ENGINEERING)

Duration: 3 hours Max Marks: 70

Answer any FIVE questions. All questions carry equal marks

1. a) Briefly explain the following data mining functionalities. 6 M

   i) Classification and prediction
   ii) Association analysis
   iii) Cluster analysis

   b) Explain nominal, ordinal, interval and ratio attributes. Classify the following attributes as nominal, ordinal, interval or ratio. 8 M

   i) Time in terms of AM or PM.
   ii) Bronze, Silver, and Gold medals as awarded at the Olympics.
   iii) ISBN numbers for books.
   iv) Number of patients in a hospital.

2. a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 7 M

b) What is data integration? Explain the issues to consider during data integration. 7 M

3. a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. 8 M

   i) Draw a schema diagram for the above data warehouse using star schema.

   ii) Starting with the base cuboid [day; doctor; patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

b) With suitable example, illustrate how data can be generalized using attribute oriented induction? 6 M

4. a) What are the general strategies for cube computation? Explain. 7 M

b) How is discovery driven cube exploration mechanism a desirable way to mark interesting points among large number of cells in a data cube. Explain. 7 M

5. a) What is constraint-based frequent pattern mining? Explain. 6 M

b) A database has nine transactions as shown below. Let *min_sup* = 20% and *min_conf* = 80%.

TID	Items Bought
T1	{ I_1, I_2, I_5 }
T2	{ I_2, I_4 }
T3	{ I_2, I_3 }
T4	{ I_1, I_2, I_4 }
T5	{ I_1, I_3 }
T6	{ I_2, I_3 }
T7	{ I_1, I_3 }
T8	{ I_1, I_2, I_3, I_5 }
T9	{ I_1, I_2, I_3 }

i) Find all frequent itemsets using Apriori algorithm. 5 M

ii) Find all strong association rules for the frequent itemset { I_1, I_2, I_5 }. 3 M

6. a) Consider the following set of training examples. 6 M

Instance	Classification	a 1	a 2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

i) What is the entropy of this collection of training examples with respect to the target function classification?

ii) What is the information gain of **a2** relative to these training examples?

b) Explain classification by back propagation. 8 M

7. a) Illustrate k-means clustering algorithm with suitable example. Also state its strengths and weaknesses. 8 M

b) What is model based clustering? Explain briefly. 6 M

8. a) Explain any two statistical approaches for outlier detection. 8 M

b) What are the two types of proximity-based outlier detection methods? Explain. 6 M