MCA 5th Sem-2019

Time: 3 Hours **Full Mark: 70**

(Answer all questions and the figures in the right hand margin indicates marks)

1. What is KDD? Discuss about the various pre-processing methods available in KDD process in brief. 14

OR

Define Data Mining. Describe the steps involved in data mining when viewed as a process of knowledge discovery. Describe why concept hierarchies are useful in data mining. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 14

2. (i) What is Data warehousing? Discuss about the differences between OLAP and ROLAP in detail. 7

(ii) Drill down and Roll up in a multi dimensional Data cube with an example. 7

OR

Discuss various data warehouse schema present along with their DMQL in neat diagrams. 14

3. A database has five transactions. Let *min sup* = 50% and *min conf* = 75% 14

T_ ID	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

(a) Find all frequent item-sets using Apriori algorithm. Do all the steps with necessary tables and Diagrams.

(b) List all of the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where *X* is a variable representing customer, and *item_i* denotes variables representing items (e.g., *A*, *B*, etc.):

for all x € transaction; buys(*X; item*1) ^ *buys*(*X; item*2) *)* → *buys*(*X; item*3) [*s; c*]

OR

Strong rules are not necessarily interesting- Discuss with an example with various measures such as: lift, Chi-square and cosine similarity 14

4. What are the basic steps for decision tree induction? For the dataset as below, construct the decision tree using algorithm. 14

department	*status*	*age*	*salary*	*count*
sales	senior	31. . . 35	46K. . . 50K	30
sales	junior	26. . . 30	26K. . . 30K	40
sales	junior	31. . . 35	31K. . . 35K	40
systems	junior	21. . . 25	46K. . . 50K	20
systems	senior	31. . . 35	66K. . . 70K	5
systems	junior	26. . . 30	46K. . . 50K	3
systems	senior	41. . . 45	66K. . . 70K	3
marketing	senior	36. . . 40	46K. . . 50K	10
marketing	junior	31. . . 35	41K. . . 45K	4
secretary	senior	46. . . 50	36K. . . 40K	4
secretary	junior	26. . . 30	26K. . . 30K	6

OR

Why naive Bayesian classification is called “naive"? Briefly, outline the major ideas of naive Bayesian classification. Discuss, how classification accuracy is measured? 14

5. Illustrate the strength and weakness of k-means algorithm. Discuss, the working of K-Means clustering with its Pseudo code. 14

OR

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): 6

(a) Compute the Euclidean distance between the two objects.

(b) Compute the Manhattan distance between the two objects.

(c) Compute the Minkowski distance between the two objects, using p = 3.

(d) Discuss about hierarchical clustering methods in brief. 8