A **04000CS402052001** **Pages: 5**

Reg No.:_______________ Name:__________________________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Eighth semester B.Tech degree examinations, September 2020

Course Code: CS402

Course Name: DATA MINING AND WAREHOUSING

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks. Marks

1 List out the four major features of data warehouse as defined by William H. Inmon, the father of data warehousing. (4)

2 What is the purpose of data discretization in data mining? List out any four data discretization strategies. (4)

3 a) Draw a suitable figure that shows data mining as a process of knowledge discovery. (2)

b) List out any four methods to handle missing attribute values in a dataset. (2)

4 a) How is entropy of a dataset calculated? (2)

b) What are the advantages of DBSCAN over k-Means clustering algorithm? (2)

5 What is confusion matrix? (4)

6 Describe the purpose of kernel function in nonlinear SVM with a suitable example. (4)

7 What is the significance of CF (Clustering Feature) in BIRCH Algorithm? (4)

8 The transaction details are given in the following table, what is the confidence and support of the association rule {Diapers} ⇒ {Coffee, Nuts}? (4)

T_id	**Items bought**
10	Beer, Nuts, Diapers
20	Beer, Coffee, Diapers, Nuts
30	Beer, Diapers, Eggs
40	Beer, Nuts, Eggs, Milk
50	Nuts, Coffee, Diapers, Eggs, Milk

9 How can we compute the dissimilarity between two binary objects? (4)

10 Describe the following activities involved in the web usage mining. (4)

i)Pre-processing activity ii) Pattern analysis activity

PART B

Answer any two full questions, each carries 9 marks.

11 a) Suppose a group of 15 sales price records has been given as follows: (3)

5, 10, 11, 13, 15, 5,8,12,11,13,18,20,18,19,19

Draw a three bucket equi-width histogram.

b) Draw a three-bucket equi-depth histogram. (3)

c) How numerosity reduction is done by MaxDiff histogram. (3)

12 a) Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. (5)

Draw a snowflake schema diagram for the data warehouse.

Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student.

b) A set of data is given: A= {115,233,484,543}. Normalize the data by Min-max normalization (range: [0.0,1.0]). (4)

13 a) Explain different OLAP operations on multi-dimensional data with suitable examples. (6)

b) A data warehouse can be modeled by either a star schema or a snowflake schema. Describe the similarities and the differences of the two models. (3)

PART C

Answer any two full questions, each carries 9 marks.

14 a) Why linear SVM is known as maximal margin classifier? Explain with suitable figure. (4.5)

b) Consider the collection of training samples (S) in the table given below. Loan_risk is the target attribute which describes the risk associated with loan for each customer. Find the value of the following. (4.5)

i) Gain(S, Sex) ii) Gain (S,Credit_rating)

Cust_ID	Age	Sex	Income	Credit_rating	Loan_risk
1000	Young	F	High	Normal	Safe
1001	Young	F	High	High	Safe
1002	Middle Age	F	High	Normal	Risky
1003	Senior	F	Normal	Normal	Risky
1004	Senior	M	Low	Normal	Risky
1005	Senior	M	Low	High	Safe
1006	Middle Age	M	Low	High	Risky
1007	Young	F	Normal	Normal	Safe
1008	Young	M	Low	Normal	Risky
1009	Senior	M	Normal	Normal	Risky
1010	Young	M	Normal	High	Risky
1011	Middle Age	F	Normal	HIgh	Risky
1012	Middle Age	M	High	Normal	Risky
1013	Senior	F	Normal	High	Safe

15 Suppose we have data on few individuals randomly surveyed. The data gives the responses towards interests to promotional offers made in the areas of Finanace, Travel, Reading, and Health. Sex is the output attribute to be predicted. Apply Naïve Bayesian classification algorithm to classify the new instance (9)

(Finance = No,Travel = Yes, Reading = Yes, Health = No).

Finance	Travel	Reading	Health	Sex
Yes	No	Yes	No	Male
Yes	Yes	No	No	Male
No	Yes	Yes	Yes	Female
No	Yes	No	Yes	Male
Yes	Yes	Yes	Yes	Female
No	No	Yes	No	Female
Yes	No	No	No	Male
Yes	Yes	No	No	Male
No	No	No	Yes	Female
Yes	No	No	No	Male

16 a) The following table shows the midterm and final exam grades obtained for students in a database course. (6)

x(Mid-term Exam)	Y(Final Exam)
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

b) Predict the final exam grade of a student who received 86 marks on the midterm exam with the above model. (3)

PART D

Answer any two full questions, each carries 12 marks.

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

(i) Compute the Euclidean distance between the two objects. (6)

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

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

b) Explain frequent subgraph mining using Apriori method. (6)

18 A database has five transactions.Let min sup=60% and min confidence=50%. Find all frequent patterns using FP-growth algorithm. (12)

Tid	Items_bought
T1000	{M,O,N,K,E,Y}
T2000	{D,O,N,K,E,Y}
T3000	{M,A,K,E}

T4000	{M,U,C,K,Y}
T5000	{C,O,O,K,I,E}

Find all strong association rules for the above table.

19 a) Explain BIRCH algorithm (9)

b) Explain the application of Naive Bayes Classifier in web content mining. (3)

****