A **04000CS402052002** **Pages: 3**

Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Eighth Semester B.Tech Degree Supplementary Examination August 2021

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 any four distinguishing features between OLAP and OLTP.	(4)
2		Perform data smoothing by bin means on 3 equi-width bins. Data: [23,27,29,15,17,32,33,29,36,39,36,45]	(4)
3	a)	What is the significance of concept hierarchy in data mining?	(2)
	b)	What is the difference between data warehouse and data mart?	(2)
4	a)	If we have a model that assigns a class label from ['sunny', 'rainy'] based on numerous factors like humidity, temperature, pressure and wind, is it classification or prediction? Justify your answer.	(2)
	b)	What is the role of activation function in neural network?	(2)
5		How is Gain Ratio calculated? What is the advantage of Gain Ratio over Information Gain?	(4)
6		Describe the working of kNN classification algorithm.	(4)
7		Discuss the role of CF (Clustering Feature) in BIRCH Algorithm. Write the CF for the cluster {(1, 1), (2, 1) (1, 2)}.	(4)
8		Dsicuss the objective function used in k-Means algorithm.	(4)
9		What is *soft focus approach* by focused crawler in web content mining?	(4)
10		Describe k-Medoids algorithm.	(4)

PART B

Answer any two full questions, each carries 9 marks.

11	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. Enumerate three classes of schemas that are popularly used for modeling data warehouses.	(3)
	b)	Draw a star schema diagram for the above data warehouse.	(3)

c) 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? (3)

12 a) How is correlation analysis done with Chi-square test? (4)

b) Illustrate three-tier data warehousing architecture. (5)

13 Describe various techniques for numerosity reduction in data mining. (9)

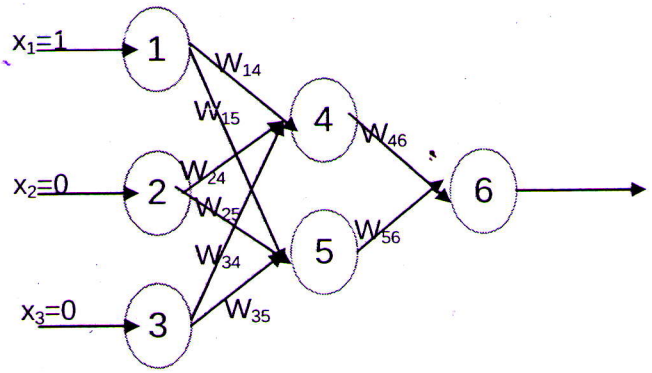
PART C

Answer any two full questions, each carries 9 marks.

14 The following shows 'Car Theft Database'. Using Naïve-Bayes algorithm, classify a Red Domestic SUV as 'Stolen' or 'Not Stolen'. (9)

Example No	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

15 The following figure shows a multilayer feed-forward neural network. Learning is done by backpropagation through time. Let the learning rate be 0.7. Bias for each neuron is 0.5. The initial weight of the network is as shown in the table below. The activation function used is the sigmoid function. Target output is 1. (9)



w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}
.3	.5	.7	.2	.4	.8	.5	.5

Calculate the updation of w_{46},w_{56},w_{14} and w_{25} for the first iteration.

16 a) What is overfitting in Decision tree? How pruning can be used to solve the problem? (4)

b) Describe the purpose of kernel function in SVM with a suitable example. (5)

PART D

Answer any two full questions, each carries 12 marks.

17 A database has five transactions. Let min sup = 60% and min_conf=50%. Find all frequent item sets using Apriori algorithm. (12)

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}

Identify strong association rules for the above table.

18 a) Suppose that our task is to cluster given sample data (height,weight) into two clusters. Let A1 and A2 be initial cluster centroids. Apply k-means algorithm to find a set of clusters. Use Manhattan distance function as dissimilarity measure. (6)

Id	Height	Weight
A1	185	72
A2	170	56
A3	168	60
A4	179	68
A5	182	72
A6	188	77

b) Describe any four challenges associated with link mining. (6)

19 a) Define the following terms related to DBSCAN algorithm with a suitable figure. (8)

i) Core object

ii) Directly density reachable

iii) Density reachable

iv) Density connected

b) How crawlers are used in web content mining? (4)

****