Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

EIGHTH SEMESTER B.TECH DEGREE EXAMINATION(S), OCTOBER 2019

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	How is data warehouse different from a database? How are they similar?	(4)
2	Compare star and snowflake schema dimension table.	(4)
3	Use the two methods below to normalize the following group of data: 100,200,300,500,900 i) min-max normalization by setting min=0 and max=1 ii) z-score normalization	(4)
4	Explain the attribute selection method in decision trees .	(4)
5	Distinguish between hold out method and cross validation method.	(4)
6	Explain prepruning and postpruning approaches in decision tree algorithm.	(4)
7	Differentiate between support and confidence.	(4)
8	How to compute the dissimilarity between objects described by binary variables?	(4)
9	Differentiate between Agglomerative and Divisive hierarchical clustering method.	(4)
10	Explain web content mining?	(4)

PART B

Answer any two full questions, each carries 9 marks.

11		The following data is given in increasing order for the attribute age: 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,36,40,45,46,52,70.	
	a)	Use smoothing by bin boundaries to smooth these data, using bin depth of 3.	(3)
	b)	How might you determine outliers in the data?	(3)
	c)	What other methods are there for data smoothing?	(3)
12)		Explain the following procedures for attribute subset selection	
	a)	Stepwise forward selection	(3)
	b)	Stepwise backward elimination	(3)
	c)	A combination of forward selection and backward elimination	(3)

13 a) Suppose a datawarehouse consists of three measures customer, account and branch and two measures count (number of customers in the branch) and balance. Draw the schema diagram using snowflake schema. (4)

b) Real-world data tend to be incomplete, noisy, and inconsistent. What are the various approaches adopted to clean the data? (5)

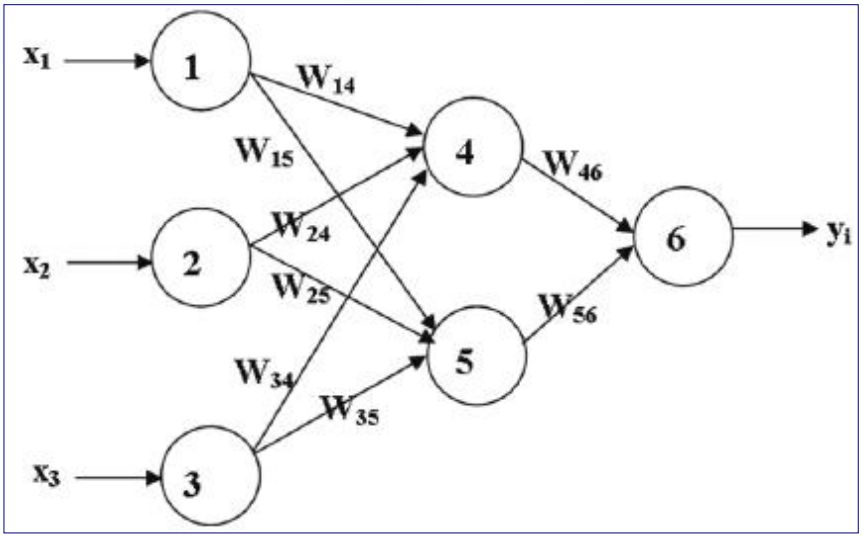
PART C

Answer any two full questions, each carries 9 marks.

14 Given the following data on a certain set of patients seen by a doctor, can the doctor conclude that a person having chills, fever, mild headache and without running nose has the flu?(Use Naive Bayes algorithm for prediction) (9)

chills	running nose	headache	fever	has flu
Y	N	mild	Y	N
Y	Y	no	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	no	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

15 The following figure shows a multilayer feed-forward neural network. Let the learning rate be 0.9. The initial weight and bias values of the network is given in the table below.The activation function used is the sigmoid function. (9)



x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1

Show weight and bias updation with the first training sample (1,0,1) with class label 1, using backpropagation algorithm

16 a) Explain classification by C4.5 algorithm. (6)

b) What is meant by Maximum Marginal Hyperplane (MMH)? (3)

PART D

Answer any two full questions, each carries 12 marks.

17 Consider the transaction database given below. Set minimum support count as 2 and minimum confidence threshold as 70%

Transaction ID	List of Item_Ids
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

a) Find the frequent itemset using Apriori Algorithm. (8)

b) Generate strong association rules . (4)

18 a) Explain DBSCAN algorithm . (8)

b) State the pros and cons of DBSCAN method. (4)

19 a) Explain clustering by k-medoid algorithm. (6)

b) Explain Apriori based frequent subgraph mining. (6)

****