Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

EIGHTH SEMESTER B.TECH DEGREE EXAMINATION, MAY 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 mining related to business intelligence?	(4)
2		Differentiate between OLTP and OLAP.	(4)
3		Why do we need data transformation? What are the different ways of data transformation?	(4)
4		An airport security screening station wants to determine if passengers are criminals or not. To do this, the faces of passengers are scanned and kept in a database. Is this a classification or prediction task? Justify	(4)
5		Where do we use Linear regression? Explain linear regression.	(4)
6		What is the significance of tree pruning in decision tree algorithms?	(4)
7		What are the two measures used for rule interestingness?	(4)
8		Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8) Compute the Manhattan distance between the two objects.	(4)
9		How density based clustering varies from other methods?	(4)
10		Differentiate web content mining and web structure mining.	(4)

PART B

Answer any two full questions, each carries 9 marks.

11	a)	Explain various stages in knowledge discovery process with neat diagram	(5)
	b)	Use the two methods below to normalize the following group of data: 1000,2000,3000,5000,9000 i)min-max normalization by setting min=0 and max=1 ii) z-score normalization	(4)
12		Suppose that a data warehouse for University consists of four dimensions date, spectator, location and game and two measures count and charge, where charge is the fare that a spectator pays when watching a game on the given date. Spectator may be students , adults or seniors ,with each category having its own charge rate	

a) Draw a star scheme for the data warehouse. (6)

b) Starting with the basic cuboid [date,spectator,location,game] ,what specific OLAP operation should be performed in order to list the total charge paid by student spectators at GM_PLACE in 2010. (3)

13 Summarize the various pre-processing activities involved in data mining (9)

PART C

Answer any two full questions, each carries 9 marks.

14 Based on the following data determine the gender of a person having height 6 ft., weight 130 lbs. and foot size 8 in. (use Naive Bayes algorithm). (9)

person	height (feet)	weight (lbs)	foot size (inches)
male	6.00	180	10
male	6.00	180	10
male	5.50	170	8
male	6.00	170	10
female	5.00	130	8
female	5.50	150	6
female	5.00	130	6
female	6.00	150	8

15 (9)

The "Restaurant A" sells burger with optional flavours: Pepper, Ginger and Chilly. Every day this week you have tried a burger (A to E) and kept a record of which you liked. Using Hamming distance, show how the 3NN classifier with majority voting would classify
{pepper = false, ginger =true, chilly = true}

	Pepper	Ginger	Chilly	liked
A	true	true	true	false
B	true	false	flase	true
C	false	true	true	false
D	false	true	false	true
E	true	false	false	true

16 a) How C4.5 differs from ID3 algorithm? (3)

b) How does backpropagation algorithm works? (6)

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 FP Growth Algorithm. (8)

b) Generate strong association rules. (4)

18 a) Explain BIRCH Clustering Method. (8)

b) What are the advantages of BIRCH compared to other clustering method. (4)

19 a) Explain k-means partition algorithm. What is the drawback of K-means? (6)

b) Term frequency matrix given in the table shows the frequency of terms per document. Calculate the TF-IDF value for the term T4 in document 3. (6)

Document/term	T1	T2	T3	T4	T5	T6
D1	5	9	4	0	5	6
D2	0	8	5	3	10	8
D3	3	5	6	6	5	0
D4	4	6	7	8	4	4

****

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)

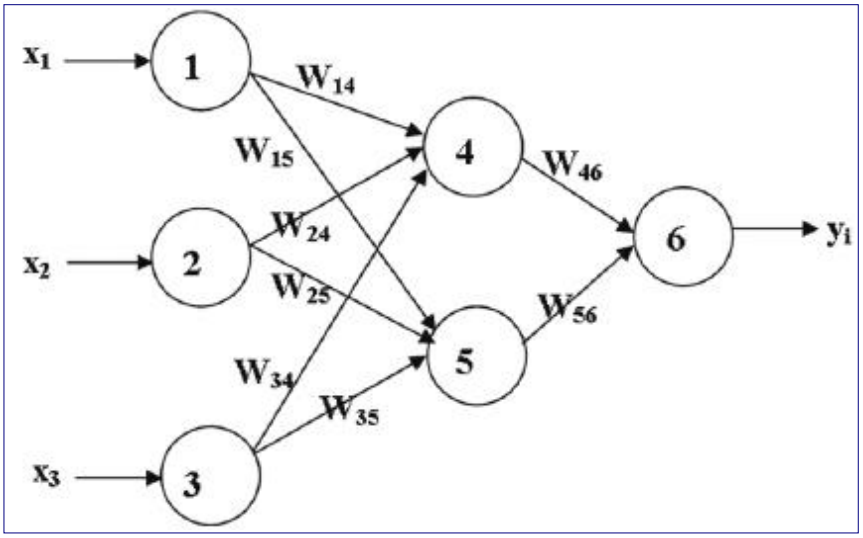
PART C

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)

****

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. i)Pre-processing activity ii) Pattern analysis activity	(4)

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: 5, 10, 11, 13, 15, 5,8,12,11,13,18,20,18,19,19 Draw a three bucket equi-width histogram.	(3)
	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. 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.	(5)
	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 (Finance = No,Travel = Yes, Reading = Yes, Health = No). (9)

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.

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

(6)

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%. (12)
Find all frequent patterns using FP-growth algorithm.

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)

****