

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 | (4) |
| | i) min-max normalization by setting min=0 and max=1 | |
| | ii) z-score normalization | |
| 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)

| Docume nt/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 |
