

No. of Pages: 3

E

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY
FIRST SEMESTER M.TECH DEGREE EXAMINATION, DECEMBER 2017

Branch: Computer Science and Engineering

Stream(s): Computer Science and Engineering

Course Code & Name: 01CS6151 Data Warehousing & Mining
(Elective I)

Answer any two full questions from each part

Limit answers to the required points.

Max. Marks: 60

Duration: 3 hours

PART A

1. a. What are the basic steps in Knowledge discovery in databases (KDD)? 6.5
- b. A datacube C has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions. 4
 - i. What is the maximum number of cells possible in the base cuboid?
 - ii. What is the minimum number of cells possible in the base cuboid?
 - iii. What is the maximum number of cells possible (including both base cells and aggregate cells) in the data cube, C?
 - iv. What is the minimum number of cells possible in the data cube, C?
2. a. Briefly describe the different OLAP operations. 5
- b. A popular data warehouse implementation is to construct a multidimensional database, known as data cube. Unfortunately, this may often generate a huge, yet very sparse multidimensional matrix. Present an example illustrating such a huge and sparse data cube. 5.5
3. a. I. Differentiate between star schema and snowflake schema. 6.5
- II. Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. Draw a star schema diagram for the data warehouse.
- b. With the help of an example describe why concept hierarchies are useful in data mining. 4

PART B

4. a. Why is naive Bayesian classification called “naive”? Briefly outline the major ideas of naïve Bayesian classification. 6.5
- b. Use single and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendrograms. 4

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

5. a. Differentiate between k-means and k-medoids algorithms that perform effective clustering. 4
- b. Consider the following data set for a binary classification. Calculate information gain for each attribute and draw decision tree by selecting the best split. 6.5

Tid	Refund	Marital Status	Taxable income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

6. a. Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: 5.5
 $A_1=(2,10)$, $A_2=(2,5)$, $A_3=(8,4)$, $A_4=(5,8)$, $A_5=(7,5)$, $A_6=(6,4)$, $A_7=(1,2)$, $A_8=(4,9)$.
- b. What are the issues faced by decision tree based classification algorithms? 5

PART C

7. a. Explain how the spatial data structures R-Tree and KD Tree differs? 3
- b. With an example differentiate between Trie and suffix trees. 6
8. a. Illustrate Data Distribution Algorithm (DDA) with the help of an example 6
- b. What are Hidden Markov Models or HMM's? 3

9. a. Consider the following transactional database, with set of items $I=\{I_1, I_2, I_3, I_4, I_5\}$. Let minimum support is 40% and confidence is 60%. Find all frequent item sets using Apriori algorithm. 6

TID	List of Items
T1	I_1, I_2, I_5
T2	I_2, I_4
T3	I_2, I_3
T4	I_1, I_2, I_4
T5	I_1, I_3
T6	I_2, I_3
T7	I_1, I_3
T8	I_1, I_2, I_3, I_5
T9	I_1, I_2, I_3
T10	I_2, I_4, I_5

- b. Describe the spatial data mining primitives. 3

http://www.ktuonline.com

Whatsapp @ 9300930012

Your old paper & get 10/-

पुराने पेपर्स भेजे और 10 रुपये पायें,

Paytm or Google Pay से