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. Redundant attributes might be present when data from multiple sources are integrated. Give any two methods to identify such redundant attributes.   5.5

   b. A data cube 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.   5.0
      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. Discuss any five implementation issues associated with data mining?   5.0

   b. With neat diagram explain the architecture of a data warehouse.   5.5

3. a. Suppose that a data warehouse consists of three dimensions time, doctor and patients and two count measures charge and count, where charge is a fee that a doctor charges a patient for a visit.   6.5
      i. Draw a schema diagram for the above data warehouse using snowflake schema.
      ii. Starting with the base cuboid [day, doctor, patient], what basic OLAP operations should be performed in order to list the total fee collected by each doctor in 2010.

   b. With the help of an example describe why concept hierarchies are useful in data mining.   4.0

1

PART B

4. a. The training data for a classifier is given below: 6.5
   Using a Bayesian Classifier, classify the tuple (Red, SUV, Domestic) as stolen or
   not stolen.

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

   b. Differentiate between agglomerative clustering and divisive clustering.   4.0

5. a. Explain k-means and k-medoids algorithms that perform effective clustering.   5.5
   Illustrate the strength and weakness of k-means in comparison with the k-
   medoids algorithm.

   b. What are decision trees? Explain how decision trees are useful in data mining.   5.0

6. a. Suppose that the data mining task is to cluster points (with (x, y) representing   6.5
   location) into three clusters, where the points are $A1(2,10)$, $A2(2,5)$,$A3(8,4)$,
   $B1(5,8)$,$B2(7,5)$,$B3(6,4)$,$C1(1,2)$,$C2(4,9)$
   The distance function is Euclidean distance. Suppose initially we assign $A1$, $B1$,
   and $C1$ as the center of each cluster, respectively. Use the *k-means* algorithm to
   show
   (i) The three cluster centers after the first round of execution.
   (ii) The final three clusters

   b. What are the issues faced by decision tree based classification algorithms?   4.0

PART C

7. a. Explain how the spatial data structures R-Tree and KD Tree differs?   3.0

   b. How do context focused crawlers improve the performance of web search?   6.0

8. a. Illustrate Count Distribution Algorithm (CDA) with the help of an example.   6.0

   b. What are Hidden Markov Models or HMM's?   3.0

2

9. a. Consider the following transactional database, with set of items I={I₁, I₂, I₃, I₄,   6.0
       I₅}. Let minimum support is 40% and confidence is 60%. Find all frequent
       itemsets using Apriori algorithm (6 Marks)

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. How decision tree induction is made possible in spatial domain?   3.0