E

APJ Abdul Kalam Technological University
First Semester M.Tech Degree Examination, 2015
Branch: Computer Science and Engineering (Cluster 01)
01CS6151 Data Warehousing & Mining

Time: 3 Hours   Max. Marks: 60

Instructions: Answer two questions from each module.

Part A

1.
   a. How is the effectiveness or usefulness associated with data mining tasks measured? (4)
   b. How are KDD and data mining related? How are they different?   (4)
   c. Illustrate the use of various OLAP primitives. (6.5)

2.
   a. A data warehouse can be modeled by either a *star schema* or a *snowflake schema*. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. (4)
   b. Data Mining is used in extracting information from data. Discuss four critical implementation issues associated with data mining.   (6.5)

3.
   a. When is data reduction used for preprocessing of data? How is data reduction done using principal component analysis? (6)
   b. What are the main characteristics that make a data warehouse distinctly different from a database? (1)
   c. Choose any normalization method to normalize the following data. 13,15,16,16,19,20,20,21,22,22,25,25,25,25,25,30,33,33,35. Justify the choice of normalization method. (3.5)

Part B

4.
   a. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8), compute the cosine similarity between the two tuples.   (3)
   b. With the training data given, derive a regression equation to model the data and classify data as short (represented using 0) or medium (represented using 1). The data is: { (1.6,0), (1.9,1), (1.88,1), (1.7,0), (1.85,1), (1.6,0), (1.7,0), (1.8,1), (1.95,1), (1.9,1), 1.8,1), (1.75,1).   (7.5)

5.
   a. When is 1R algorithm used for classification? Show how the algorithm can be used for classification. (5.5)
   b. How do K-means and K-medoids algorithm differ? Justify the contexts in which each of these algorithms can be used? (5)
6. The training data for a classifier is given below: http://www.ktuonline.com
   a. Using a Bayesian Classifier, classify the tuple (Adam, M, 1.95) as short, medium or tall. The training data for the classifier is given below:

Name	Gender	Height	Output
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium

Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

Use a Naïve Bayes Classifier to classify Adam who is Male and 1.95m tall as short, medium or tall.
(7.5)

b. When objects are described using binary valued attributes, what are the similarity measures used. (3)

Part C

7.
   a. How are support and confidence measures used in generating association rules? (4)
   b. How do context focused crawlers improve the performance of web search? (5)

8.
   a. Illustrate the use of two spatial data structures in storing spatial information (4)
   b. Apriori algorithm is used in finding large item sets. How does sampling improve its performance? (5)

9. a) When the database used in frequent itemset mining is very large, how can large itemsets be identified? Give an algorithm for the same. (6)
   b) How are spatial query processing handled by kd-trees ? (3)