KUMARAGURU college of technology
character is life

Register Number: ……..

B.TECH DEGREE EXAMINATIONS: NOV 2015

(Regulation 2009)

Seventh Semester

INFORMATION TECHNOLOGY

ITY117 : Data Warehousing and Data Mining

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 1 = 10 Marks)

1 Which of the following regression models the data to fit a straight line with equation $Y=\alpha+\beta X$?

a) Linear regression b) Multiple regression

c) Data set regression d) Data regression

2 Identify the most popular data model for a data ware house.

a) UML b) ER

c) Multi dimensional model d) Database

3 Capturing user access patterns in such distributed information environments is called mining path

a) Cluster pattern b) Data Pattern

c) Traversal pattern d) Classify pattern

4 The mean and standard deviation of the values for the attribute income are 54,000 and 16,000 respectively. With Z-score normalization ,a value of 73,600 for income is transformed to

a) 0.716 b) 1.225

c) 0.817 d) 2.225

5 Which of the following cube is for the highest level of abstraction?

a) Apex Cuboids b) Top Cuboids

c) Bottom Cuboids d) Base Cuboids

6 If the rule concerns associations between the presence or absence of items

a) Single-level association rule b) Multidimensional association rule

c) Multilevel association rule d) Boolean association rule

7 Which of the following is the collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters?

a) Predictive Data mining
b) Clustering
c) Supervised Learning
d) Descriptive Data mining

8 Which of the following is related to STING?

a) Classical Partitioning Methods
b) Model based methods
c) Hierarchical Methods
d) Grid based methods

9 Identify the system that supports retrieval based on the image content, such as color histogram, texture, shape, objects and wavelet transforms.

a) Description –based retrieval
b) Information retrieval
c) Content-based retrieval
d) Concept-based retrieval

10 Which of the following data mining technique uses audio signals to indicate data patterns or features of data mining results?

a) Web Structure Mining
b) Web content Mining
c) Audio content Mining
d) Audio Mining

PART B (10 X 2 = 20 Marks)

11. Define Snowflake schema.
12. What is an iceberg query?
13. In real-world data, tuples with missing values for some attributes are a common occurrence. List out various methods for handling this problem.
14. What is mean by k-fold cross validation? Given a dataset with 1200 instances, how k-fold cross validation is done when k=10?
15. Calculate the information gain with 150 samples of class postgraduate and 120 samples of class undergraduate.
16. Give an example for Multilevel and Multi dimensional Association rule Mining.
17. Given two objects represented by the tuples (23, 1, 41, 10) and (20, 0,32, 8).Calculate the Minkowski Distance between the two objects using q=3.
18. Define outlier mining.
19. What are the basic measures for text retrieval?
20. List any four applications of data mining.

PART C (5 X 14 = 70 Marks)

21 a) i) Explain the architecture of a data warehouse with a neat sketch. (8)

ii) Differentiate between OLAP and OLTP. (6)

(OR)

b) i) What is a data cube? How it is created? Explain the operations performed on data cubes. (8)

ii) Suppose a data warehouse consists of three dimensions time, doctor and patient and the two measures count and charge, where charge is the fee that a doctor charges for a visit. Enumerate and draw all the schema diagrams. (6)

22 a) i) Discuss the issues to consider during data integration. (8)

ii) Explain entropy discretization with necessary formulas. (6)

(OR)

b) i) Summarize the major tasks involved in data preprocessing. (8)

ii) The data for analysis include the attribute age. The age values for the data tuples are (in increasing order) (6)

13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,35,35,35,35,36,40,45,46,52,70

1. Use min-max normalization to transform the value 35 for age onto the range[0.0,1.0]
2. Use Z-score normalization to transform the values 35 for age , where the standard deviation of age is 12.94 years.
3. Use Normalization by decimal scaling to transform the value 35 for age.

23 a) A data base has four transactions with min_sup=60%

TID	date	Items_bought
T100	10/15/2010	{K,A,D,B}
T200	10/15/2010	{D,A,C,E,B}
T300	10/15/1998	{C,A,B,E}
T400	10/15/1995	{B,A,D}

Find all frequent item set using apriori and FP-growth, respectively

(OR)

b) Explain the working of Naïve Bayesian Classifier. Find the class(X) for the following dataset by executing it in the given training set.

X=(age<30, Income=medium, student=yes ,credit_ rating=Fair)

Training Set:

Age	Income	Student	Credit _rating	Buys _laptop
<= 30	High	No	Fair	No
<=30	High	No	Excellent	No
<=30	Low	Yes	Fair	Yes
31..40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes

>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31..40	Low	Yes	Excellent	Yes
<= 30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<= 30	Medium	Yes	Excellent	Yes
31..40	Medium	No	Excellent	Yes
31..40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

24 a) What is an outlier? Explain in detail about statistical outlier detection and distance based outlier detection.

(OR)

b) i) What is clustering? Briefly describe the Partitioning and Hierarchical clustering methods. Give examples in each case. (10)

ii) Summarize the typical requirements for clustering in data mining. (4)

25. a) Discuss in detail the vital role played by data mining in the field of text databases.

(OR)

b) i) Explain about spatial data mining with suitable examples (8)

ii) Summarize the role of data mining in web. (6)