SPRING END SEMESTER EXAMINATION-2018

2nd Semester M.Tech / Ph.D Coursework

DM&DW

CS-6301

[For 2017 Admitted Batch]

Time: 3 Hours Full Marks: 50

Answer any SIX questions including question No.1 which is compulsory.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

1. Answer the following questions [1 × 10]

(a) What is meant by pruning in a decision tree induction?

(b) Define snowflake schema.

(c) Define association and correlation.

(d) Compare the advantage of FP Growth algorithm over Apriori algorithm.

(e) "Data mining is a multidisciplinary field". Justify your answer.

(f) How to find interestingness of an association rule?

(g) What is ETL Process? How it is helpful in preprocessing the data?

(h) Distinguish STARjoin and STARindex.

(i) What is outlier? How to handle outlier?

(j) What is data mart?

2. (a) What is KDD Process? Explain briefly how the knowledge is extracted from it. [4]

(b) What is data warehouse? Give the steps for design and construction of Data Warehousesand explain with three-tier architecture diagram. [4]

3. (a) What is Classification? Explain decision tree classification algorithm with a suitable example. [4]

(b) Suppose that a data warehouse consists of four dimensions customer, product, salesperson and sales time, and the three measure sales Amt(in rupees), GST(in rupees) and payment_type(in rupees). Draw the different classes of schemas that are popularly used for modeling data warehouses and explain it. [4]

4. (a) Describe the k-means and k-medoids algorithm in terms of shape of cluster that can be determined and parameter that must be specified? [4]

(b) The sales of a company (in crores) for each year are shown in the table below. [4]

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

Find the least square regression line $y = ax + b$. Use the least squares regression line as a model to estimate the sales of the company in the year 2012.

5. (a) From the given data, using k-nearest neighbor classifier, find the class of the data tuple (38, 45) where k value is 3. [4]

A	B	Class		A	B	Class
26	30	L		36	52	P
30	32	L		40	62	P
36	42	L		43	70	P

(b) The given data set describes two categorical input variables and a class variable that has two outputs. [4]

Weather	Car	Class		Weather	Car	Class
sunny	working	go-out		rainy	broken	stay-in
rainy	broken	go-out		rainy	broken	stay-in
sunny	working	go-out		sunny	working	stay-in
sunny	working	go-out		sunny	broken	stay-in
sunny	working	go-out		rainy	broken	stay-in

Predict the class label of second instance using Naïve Bayes classifier.

6. (a) What is a confusion matrix and what is its role in evaluation of a classifier's performance? [4]

(b) Briefly outline the major steps of decision tree classification. State the stopping criteria of Decision Tree algorithm. [4]

7. (a) A database has five transactions. Let min_support =50% and min_confidence=75%. [4]

TID	Items_bought
T100	Bread,Cheese,Eggs,Juice
T200	Bread,Cheese,Juice
T300	Bread,Milk,Yoghurt
T400	Bread, Juice, Milk
T500	Cheese,Juice,Milk

Find all frequent item sets using Apriori Algorithm.

(b) Suppose that the data for analysis include the attributed age. The age values for the data tuples are 13, 15, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. [4]

(i) Use smoothing by bin means to smooth the above data using a bin depth of 3. Illustrate your steps.

(ii) Classify the various methods for data smoothing.

8. Short notes (Any TWO) [4×2]

(a) Classification by Backpropagation

(b) Neural Network classifier

(c) OLAP vrs. OLTP

*****