

JK LAKSHMIPAT UNIVERSITY, JAIPUR

INSTITUTE OF ENGINEERING AND TECHNOLOGY

End Term Examination (Regular), November-December 2015

B. Tech in Computer Science Engineering, Semester VII

Subject Code: CSE 727

Roll No.....

Subject Name: Data Warehousing and Mining

Time: 3 hours

Max. Marks: 100

Instructions to students:

1. Do not write anything other than your roll number on question paper.
2. Assume suitable data wherever required and mention it clearly.
3. Writing appropriate units, nomenclature, and drawing neat sketches/schematics/ flow chart wherever, required are an integral part of the answer.

Q.1A Objective type multiple choice questions (10 questions @ 1 mark)

10

- i. Concept/class description is one of the ----- of data mining.
(a) Techniques (b) Needs (c) Functionalities (d) Issues
- ii. Which of the following is not used for measuring the dispersion of data?
(a) Range (b) Interquartile range (c) Mode (d) Variance
- iii. ----- provides the analytical foundation for data preprocessing.
(a) Data warehousing (b) Descriptive data summarization
(c) Knowledge discovery process (d) None of these
- iv. The itemsets that have support above the minimum prespecified support are known as ----- itemsets.
(a) Small (b) Cube-based (c) Medium (d) Large
- v. If an attribute A has v possible values, then there are ----- possible ways to form two partitions of the data, D , based on a binary split on attribute A .
(a) v (b) $2v - 1$ (c) $2v - 2$ (d) $2v$
- vi. Which of the following method is used for improving the accuracy of classifier and predictor?
(a) Bootstrap (b) Bagging (c) Cross-validation (d) Holdout
- vii. The PAM algorithm is an example of ----- clustering?
(a) Partitioning (b) Hierarchical (c) Density-based (d) Model-based
- viii. The dissimilarity between asymmetric objects i and j is computed by using
(a) Manhattan distance (b) Simple matching coefficient
(c) Jaccard coefficient (d) Minkowski distance
- ix. Which of the following keywords is not a major feature of a data warehouse?
(a) object-oriented (b) integrated (c) time-variant (d) nonvolatile
- x. ----- is not an OLAP operation.
(a) Pivoting (b) Slice and dice (c) Bitmap indexing (d) Roll up and drill down

Q.2 Short answer type questions (4 questions @ 2.5 marks)

10

- i. What is data mining? What does motivate one to use data mining?
- ii. Briefly elaborate the concept of evolution analysis.
- iii. Briefly describe the heuristic (greedy) methods for obtaining a good subset of

attributes from original attribute set.

iv. Briefly discuss the key features of a data warehouse.

Q.3A Suppose a group of 12 sales price records has been sorted as follows: 05
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Answer the following:

- Use *smoothing by bin means* to smooth the data, using a bin depth of 3.
- Use *smoothing by bin boundaries* to smooth the data, using a bin depth of 3.

Q.3B Table below shows a set of paired data where x is the number of years of work experience of a college graduate and y is the corresponding salary of the graduate. 05

x years experience	y salary (in 1000\$)
20	10
30	20
30	30
35	35
40	40
50	45

Calculate the *correlation coefficient* (Pearson's product moment coefficient) between these variables, where the mean of x and y (i.e. $\bar{x}$ and $\bar{y}$) are 34.2 and 30, and the standard deviation of x and y (i.e. σ_x and σ_y) are 9.3 and 11.9, respectively. Are these two variables positively or negatively correlated?

Q.4A How the analysis of attribute (dimension) relevance is performed? Give the mathematical expression for any measure used for analysis of attribute relevance. 05

Q.4B Consider the sales price records given in Q. 3A. Determine the following: 05

- Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data?
- Give the five-number summary of the data.
- Show a box plot of the data.

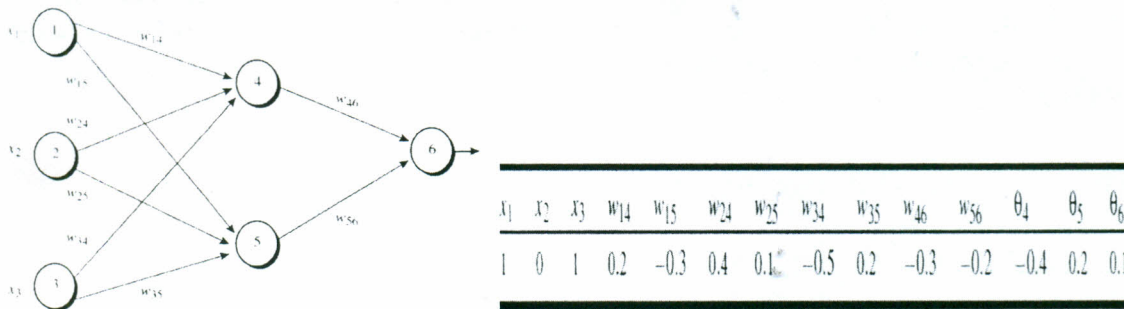
Q.5A Using Apriori algorithm (assume minimum support level is 30%), generate all the frequent itemsets on the following transactional database: 05

TID	Items_bought
T100	B, C, H
T200	B, F, S
T300	A, F, G
T400	B, C, H
T500	B, F, G
T100	B, E, O

Q.5B Suppose half of all the transactions in a clothes shop consist of jeans and one-third of all transactions in the shop consist of T-shirts. Also, suppose that half of the transactions that contain jeans also contain T-shirts. Write down all the (non-trivial) association rules you can deduce from the above information, giving support and confidence of each rule. 05

Q.6A Why is *tree pruning* useful in decision tree induction? Discuss the approaches of tree pruning. 05

Q.6B Figure below shows a multilayer feed-forward neural network. Let the learning rate be 0.9. The initial weight and bias values of the network are given in Table below, along with the first training tuple, $X = (1, 0, 1)$, whose class label is 1. 05



Calculate

- Net input and output
- Error values at each node
- Weight and bias updates values

OR

Q.6B Consider the data given in Q.3B. Determine the following: 05

- Plot the data. Do x and y seem to have a linear relationship?
- Use the *method of least squares* to find an equation for the prediction of a graduate's salary based on the graduate's experience.
- Predict the salary of a graduate who has 10 years of experience.

Q.7A What is cluster analysis? What are the various application areas of clustering? 05

Q.7B Suppose that a patient record table below contains the attributes *name*, *gender*, *fever*, *cough*, *test-1*, *test-2*, *test-3*, and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary. 05

A relational table where patients are described by binary attributes.

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
...	...	...	...	...	...	...	...

For asymmetric attribute values, let the values $Y(\text{yes})$ and $P(\text{positive})$ be set to 1, and the value $N(\text{no or negative})$ be set to 0. Suppose that the distance between objects (patients) is computed based only on the asymmetric variables.

- Show the *contingency matrix* for each pair given Jack, Marry and Jim.
- Compute the *Jaccard coefficient* for each pair.
- Who do you suggest would make the best pair of patients? Which pair of individuals would be the least compatible?

OR

Q.7B Explain in detail the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm. 05

Q.8A Write and explain the *k-means* algorithm of clustering. Suppose that the data mining task is to cluster the records (objects) given in Q.3B into two clusters C_1 10

and C_2 (i.e. $k=2$).

The distance function is Euclidian distance. Let initially we assign record 2 (i.e. (30, 20)) and record 5 (i.e. (40, 40)) as the center (centroid) of C_1 and C_2 . Use the *k-means* algorithm to show the two cluster centers after each round execution and the final two clusters.

- Q.9A What is a schema? Discuss Star, Snowflake and Fact Constellation Schemas used in data warehouse. 10
- Q.9B Discuss the following OLAP operations with an example. 10
- i. Pivoting ii. Slice and dice iii. Rollup and drill-down
- OR
- Q.9B: Differentiate between the following (any two): 10
- i. Operational database systems Vs Data warehouse
- ii. Multidimensional Data Modelling Vs Relational Data Modelling
- iii. OLTP System Vs OLAP System
- iv. ROLAP Vs MOLAP Vs HOLAP