

JK LAKSHMIPAT UNIVERSITY, JAIPUR

INSTITUTE OF ENGINEERING AND TECHNOLOGY

End Term Examination (Regular), November 2016
B. Tech in Computer Science Engineering, Semester VII

Subject Code: CSE701

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.	Knowledge is referred to (a) Non-trivial extraction of implicit previously unknown and potentially useful information from data (b) Set of columns in a database table that can be used to identify each record within this table uniquely (c) Collection of interesting and useful patterns in a database (d) None of these	
ii.	 is an essential process where intelligent methods are applied to extract data patterns. (a) Data warehousing (b) Data mining (c) Text mining (d) Data selection	
iii.	Any subset of a frequent set is a frequent set. This is _____. (a) Upward closure property (b) Downward closure property (c) Maximal frequent set (d) Border set	
iv.	If T consist of 500000 transactions, 20000 transaction contain bread, 30000 transaction contain jam, 10000 transaction contain both bread and jam. Then the support of bread and jam is _____. (a) 2% (b) 20% (c) 3% (d) 30%	
v.	Bayesian classifiers is (a) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation. (b) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory. (c) Any mechanism employed by a learning system to constrain the search space of a hypothesis. (d) None of the above	
vi.	In a feed- forward networks, the connections between layers are _____. from input to output. (a) Bidirectional (b) Unidirectional (c) Multidirectional (d) Directional	
vii.	The dissimilarity between symmetric objects i and j is computed by using	

	(a) Manhattan distance	(b) Simple matching coefficient
	(c) Jaccard coefficient	(d) Minkowski distance
viii.	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8). The Minkowski distance (using $p = 3$) between the two objects is -----	
	(a) 5.15	(b) 6.15
	(c) 7.55	(d) 8.15
ix.	Data warehouse contains _____ data that is never found in the operational environment.	
	(a) Normalized	(b) Informational
	(c) Summary	(d) Denormalized
x.	A data warehouse is which of the following?	
	(a) Can be updated by end users (b) Contains numerous naming conventions and formats	
	(c) Organized around important subject areas (d) Contains only current data	

Q.2	Short answer type questions (4 questions @ 2.5 marks)	10
i.	In real-world data, tuples with <i>missing values</i> for some attributes are a common occurrence. Briefly describe various methods for handling this problem.	
ii.	Differentiate between Data Characterization and Data Discrimination.	
iii.	What is cluster analysis? How it is different from classification?	
iv.	Briefly discuss the basic characteristics of a data warehouse.	

Q.3A	Demonstrate data mining as a confluence of multiple disciplines. Also give the classification of data mining systems.	05
Q.3B	Suppose a group of 27 values for attribute <i>age</i> records has been sorted as follows: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70 Using this data for <i>age</i> values, answer the following: (i) Use min-max normalization to transform the value 35 for <i>age</i> on to the range [0.0, 1.0]. (ii) Use z-score normalization to transform the value 35 for <i>age</i> , where the standard deviation of <i>age</i> is 12.94 years. (iii) Use normalization by decimal scaling to transform the value 35 for <i>age</i> .	05

Q.4A	A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preferences (Republican, Democrat, or Independent). Results are shown in the table below:	05
------	--	----

Voting Preferences				
	Republican	Democrat	Independent	Row Total
Male	200	150	50	400
Female	250	300	50	600
Column Total	450	450	100	1000

Suppose, for 2 degree of freedom, the χ^2 (Chi square) value needed to reject the hypothesis at the 0.05 significance level is 14.280. Determine whether the attributes gender and voting preferences are independent or strongly correlated?

- Q.4B Consider the *age* records given in Q. 3B. Determine the following: 05
- What is the median?
 - What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
 - 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
T600	B, E, O

- Q.5B Consider the association rule below, which was mined from the student database at Big University: 05

$major(X, 'science') \Rightarrow status(X, 'undergrad') [s, c]$

Suppose that the number of students at the university is 5000, that 70% of the students are majoring in science, that 64% of the students are registered in programs leading to undergraduate degrees, and that 56% of the undergraduates at the university major in science. Compute the support, s , and confidence, c , for the given rule.

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

- Q.6B What are neural networks? Describe the various factors which make them useful for classification and prediction in data mining. Also explain the multilayer feed-forward network. 05

OR

- Q.6B Consider the following expected (y_i) and observed (y_i') outcomes (sanctioned loan amount in 1000\$) obtained for a prediction mining task. 05

y_i	10	15	19	25	30	35
y_i'	11.5	13.5	20.4	23.8	32	35.5

Compute the following errors:

- Mean absolute error
- Mean squared error
- Relative absolute error
- Relative squared error

- Q.7A 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 (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?

Q.7B Suppose that we have a sample data of Table (below), where test data is ordinal. 05

Object Identifier	test (ordinal)
1	excellent
2	fair
3	good
4	excellent

Compute the dissimilarity matrix for the given data.

OR

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

Q.8A Write down the *k-means* algorithm of clustering. Suppose that the data mining task is to cluster the records (objects) given in Table (below) into two clusters C_1 and C_2 (i.e. $k=2$). 10

Age	20	30	30	35	40	50
Income	10	20	30	35	40	45

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. Rollup ii. Drill-down iii. Slice and dice iv. Pivoting

OR

Q.9B Write short note on the following (any two): 10

- Heuristic methods for attribute subset selection
- Multi-Dimensional and Multi-level Association rule mining
- K-nearest neighbor classifiers
- Process of Data Warehouse Design