

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.	In which of the following processes data objects are analyzed without consulting a known class label? (a) Cluster analysis (b) Evolution analysis (c) Outlier analysis (d) None of these	
ii.	----- is not an attribute subset selection method. (a) Attribute construction (b) Stepwise forward selection (c) Decision tree induction (d) Stepwise backward elimination	
iii.	If attribute A has c distinct values $a_1, \dots, a_c$ and attribute B has r distinct values $b_1, \dots, b_r$ in a dataset, then the degree of freedom (DF) is calculated as (a) $r \times c$ (b) $(r-1) \times c$ (c) $(r-1) \times (c-1)$ (d) $r \times (c-1)$	
iv.	If the horizontal data format represents the data in the form as TID : itemset, then vertical data format represents the data in the form as ----- (a) TID : itemset (b) Item : TID_set (c) Itemset : TID_set (d) TID_set : item	
v.	Which is not a true statement w.r.t. Naïve Bayes classifier? (a) It is less prone to errors (b) It provides a theoretical justification for other classifiers (c) It works well with real world situations (d) None of these	
vi.	The type of data used in clustering which has only two states is known as ----- variable (a) Interval-scaled (b) Binary (c) Ordinal (d) Categorical	
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.	The ----- hierarchy deals with the ordering of attributes of a particular dimension. (a) Schema (b) Set-grouping (c) Concept (d) Partial	
ix.	----- provide multidimensional data analysis and summarized data. (a) OLAP tools (b) Data mining tools (c) Data warehousing (d) None of these	
x.	The 'slice' OLAP operation deals with ----- (a) Selecting all but one dimension of the data cube (b) Merging the cells along one dimension (c) Selecting the cells of any one dimension of the data cube (d) None of these	

Q.2	Short answer type questions (4 questions @ 2.5 marks)	10																
i.	"Data mining can be viewed as searching for knowledge (interesting patterns) in the given data" What makes a pattern interesting?																	
ii.	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.																	
iii.	What are the various forms to represent a classification model?																	
iv.	Briefly explain the major features of a data warehouse which distinguish it from other data repository systems.																	
Q.3A	Suppose a group of 12 <i>sales price</i> (in dollars) records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 Using this data for <i>sales price</i> , answer the following:	05																
i.	Use min-max normalization to transform the value 35 for <i>sales price</i> on to the range [0.0, 1.0].																	
ii.	Use z-score normalization to transform the value 35 for <i>sales price</i> , where the standard deviation of <i>sales price</i> is 69.86 dollars.																	
iii.	Use normalization by decimal scaling to transform the value 35 for <i>sales price</i> .																	
Q.3B	Suppose that a group of 1,500 people was surveyed. Respondents were classified by gender (male or female) and by their preferred type of reading material (fiction, non-fiction). Thus, we have two attributes, <i>gender</i> and <i>preferred_reading</i> . The observed frequency (or count) of each possible joint event is summarized in the following contingency table:	05																
<table><tr><td></td><td><i>fiction</i></td><td><i>non-fiction</i></td><td>Row Total</td></tr><tr><td><i>male</i></td><td>250</td><td>50</td><td>300</td></tr><tr><td><i>female</i></td><td>200</td><td>1000</td><td>1200</td></tr><tr><td>Column Total</td><td>450</td><td>1050</td><td>1500</td></tr></table>				<i>fiction</i>	<i>non-fiction</i>	Row Total	<i>male</i>	250	50	300	<i>female</i>	200	1000	1200	Column Total	450	1050	1500
	<i>fiction</i>	<i>non-fiction</i>	Row Total															
<i>male</i>	250	50	300															
<i>female</i>	200	1000	1200															
Column Total	450	1050	1500															
Suppose, for 1 degree of freedom, the χ^2 (Chi square) value needed to reject the hypothesis at the 0.001 significance level is 10.828. Determine whether the attributes <i>gender</i> and <i>preferred_reading</i> are independent or strongly correlated?																		
Q.4A	"How a target class is compared with its contrasting class(es) performed?" Explain the procedure with example.	05																
Q.4B	Consider the temperature (in degree Celsius) values for the data tuples are (in increasing order): 25, 26, 26, 27, 27, 28, 29, 29, 30, 32, 32, 33, 33, 33, 35, 35, 36, 37, 38, 38, 38, 39, 40 Determine the following:	05																
i.	Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data?																	
ii.	Give the five-number summary of the data.																	
iii.	Show a box plot of the data.																	
Q.5A	Consider the association rule below, which was mined from the student database at <i>Big University</i> : $major(X, 'science') \Rightarrow status(X, 'undergrad')$ [s, c] Suppose that the number of students at the university is 5000, that 70% of the	05																

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.5B Explain, with example, how multilevel association rules can be mined?

05

Q.6A The following table consists of training data from an employee database. The data have been generalized. For example, "31...35" for *age* represents the age range of 31 to 35. For a given row entry, *count* represents the number of data tuples having the values for *department*, *status*, *age*, and *salary* given in that row.

05

<i>department</i>	<i>status</i>	<i>age</i>	<i>salary</i>	<i>count</i>
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let *status* be the class label attribute. Given a data tuple having the values "systems," "26...30," and "46-50K" for the attributes *department*, *age*, and *salary*, respectively, what would a naïve Bayesian classification of the *status* for the tuple be?

Q.6B The following table shows the midterm and final exam grades obtained for students in a database course.

05

x Midterm Exam	y Final Exam
20	10
30	20
30	30
35	35
40	40
50	45

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.

- 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 student's final exam grade based on the student's midterm grade in the course.
- Predict the final exam grade of a student who received an 86 on the midterm exam.

OR

Q.6B Consider a confusion matrix below for the classes *buys computer = yes* and *buys computer = no*, where an entry is row i and column j shows the number of tuples

05

of class i that were labeled by the classifier as class j .

Classes	<i>buys_computer</i> = yes	<i>buys_computer</i> = no	Total
<i>buys_computer</i> = yes	6,954	46	7,000
<i>buys_computer</i> = no	412	2,588	3,000
Total	7,366	2,634	10,000

Determine the following measures w.r.t. the given classifier:

i. Sensitivity ii. Specificity iii. Precision iv. Accuracy

Q.7A What are the typical requirements of clustering in data mining? 05

Q.7B How can data for a variable be standardized? How can we compute dissimilarity between two interval-scaled variables? 05

OR

Q.7B Write a short note on CURE algorithm. 05

Q.8A Write and explain the *k-means* algorithm of clustering. Suppose that the data mining task is to cluster the following six points (with (x, y) representing location) into two clusters C_1 and C_2 (i.e. $k=2$). 10

$A_1(20, 10)$, $A_2(30, 20)$, $A_3(30, 30)$, $B_1(35, 35)$, $B_2(40, 40)$, $B_3(50, 45)$.

The distance function is Euclidian distance. Let initially we assign A_2 and B_2 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 Describe the 3-tier data warehouse architecture in detail. 10

Q.9B Explain various kinds of OLAP servers in detail. 10

OR

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

i. Data Cube Aggregation

ii. Mining Multi-Dimensional Association rules from Relational Databases

iii. K-nearest neighbor classifiers

iv. Concept hierarchy and its types