B.E. INFORMATION TECHNOLOGY 4TH YEAR 1ST SEMESTER EXAMINATION-2018

SUPPLIMENTARY

Subject: Data Mining

Time: 3 Hours **Full Marks: 100**

Note: Attempt any *five* questions and all parts/subparts of a question shall be written together.

Q.1

a) The following table consists of five transactions with min_sup = 55% and min_conf = 75%.

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5
500	1 3 5

i) Find all frequent itemsets using *Apriori* and *FP-growth* respectively. Compare the efficiency of the two mining processes.

ii) List all the strong association rules (with support s =30% and confidence c=60%) matching the following *metarule*, where X is a variable representing customers, matching denotes variables representing items (e.g., A, B etc.):

$$\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c]$$

b) Explain with suitable example(s): normalized schema and galaxy schema as data warehouse model.

c) Explain with suitable examples of complete linkage algorithm.

(4x2+2+1)+(3+3) +3

Q.2

a)

Objects	X	Y
X1	2	3
X2	0	0
X3	10	-2
X4	-3	8
X5	4	5
X6	3	4
X7	-1	-2
X8	0	2
X9	-3	6
X10	0	8

The following table consists of ten objects (i.e., points) in 2-D systems.

Find best possible *two* clusters with the help of *successive iterations* formed by the objects using the algorithms:

i) k-means and

ii) k-medoids

b) Describe various methods/tests/techniques with suitable example(s) used in four levels of data preprocessing.

c) Define with suitable example(s): nominal attributes and ordinal attributes.

(5x2)+(2x4)+2

Q.3

a) Suppose that a data warehouse for Big University consists of the four dimensions *student*, *course*, *semester*, and *instructor*, and two measures *count* and *avg_grade*. At the lowest conceptual level (e.g., for a given *student*, *course*, *semester*, and *instructor* combination), the *avg_grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

i) Draw the snowflake and fact constellation schema diagrams for the data warehouse.

ii) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big University student.

iii) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids)?

b) Find out the outliers w.r.t. the data set, B={0.001, 1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20, 0.15, 0.49, 0.95, 0.24, 1.37, 0.17, 60.05, 6.98, 0.10, 0.94, 0.38, 49.15} using boxplots.

c) What are the advantages of using CLARA and CALARANS algorithms? Write the working principles of CF- tree.

(3x2+2+2)+ 4+(3+3)

Q.4

a)

RID	age	income	student	credit_rating	buys_computer
1	senior	high	yes	fair	yes
2	youth	medium	no	excellent	yes
3	middle_aged	high	no	excellent	no
4	senior	low	no	fair	yes
5	youth	low	yes	fair	no
6	senior	medium	yes	excellent	yes
7	middle_aged	low	no	excellent	no
8	youth	high	yes	fair	no

The following table consists of training data from an employee database.

Assuming *buys_computer* as class label attribute, select appropriate splitting attribute and draw the decision tree accordingly using:
i) Gain ratio and
ii) Gini index

c) What is k-nearest-neighbor graph approach? How can it be used in data mining?.
b) Describe the following with suitable examples: working principles of *FP-growth* from *FP-tree*

(5x2+2)+(3+2)+3

Q.5

a) The following contingency table summarizes supermarket transaction data. Suppose that the association rule *hot dogs* => *hamburgers* is mined with a minimum support threshold of 25% and a minimum confidence threshold of 50%.

	hot dogs	$\overline{\textit{hot dogs}}$
hamburgers	2000	500
$\overline{\textit{hamburgers}}$	1000	1500

i) Check whether the association rule is misleading strong or not?
ii) If it is misleading then what measure(s) will you take and how *(in different ways e.g., lift & chi-square)* to filter the misleading strong association rule?

b) Using the employee database given in Q.5(a) and assuming *buys_computer* as class label attribute, find out Information gain for measuring attribute selection.
c) What is/are the problem(s) and solution(s) of overfitting the data in making decision tree? Explain with suitable examples.

(2+(3+4))+5+(4+2)

Q.6 Explain the followings in brief:
a) CLIQUE algorithm (steps only).
b) OLAP vs. OLTP (w.r.t. DB design, priority, metric, unit of work, orientation, workers, function, and view)
c) Data Mart and Meta Data Repository (with examples).
d) Hierarchical clustering algorithm: BIRCH vs. CHAMELEON.
e) Relative interconnectivity and Relative closeness (with appropriate expressions).

4x5