Code: 15A05602 **R15**

B.Tech III Year I Semester (R15) Regular Examinations November/December 2017

DATA WAREHOUSING & DATA MINING

(Information Technology)

Time: 3 hours Max. Marks: 70

PART – A

(Compulsory Question)

*****

1 Answer the following: (10 X 02 = 20 Marks)

(a) What is an operational data store?

(b) How is a data warehouse different from a database? How are they similar?

(c) List and describe the five primitives for specifying a data mining task.

(d) What is clustering?

(e) What is information retrieval?

(f) What is data pre-processing?

(g) What is classification model?

(h) What are the security measures that have been taken to protect the privacy of individual while collecting and mining data?

(i) List four statistical methods for data analysis.

(j) How the multimedia data can be generalized?

PART – B

(Answer all five units, 5 X 10 = 50 Marks)

UNIT – I

2 (a) Define data mining. List and explain the different challenges of data mining.

(b) List the different purposes of dimensionality reduction and the different techniques used to reduce dimensionality.

OR

3 (a) What are “redundant “and “irrelevant features? With flow chart, explain the architecture for feature subset selection.

(b) Explain the operations of data cube with suitable examples.

UNIT – II

4 (a) Distinguish between OLTP and OLAP.

(b) Explain three tier data warehouse architecture with neat diagram.

OR

5 (a) Briefly compare the following concepts. You may use an example to explain your point (s). Snow flake schema, fact constellation, starnet query model.

(b) What is iceberg cube? Explain with example.

Contd. in page 2

UNIT – III

6 (a) Consider the following transaction data set:

TID	T100	T200	T300	T400	T500	T600	T700	T800	T900
ITEM IDS	I1, I2, I5	I2, T4	I2, I3	I1, I2, I4	I1, I3	I2, I3	I1, I3	I1, I2, I3, I5	I1, I2, I3

Generate the list of frequent item-set ordered by their corresponding suffixes.

(b) Explain FP-Growth algorithm for discovering frequent itemsets without candidate generation.

OR

7 (a) Consider the following training set for predicting the loan default problem

T-id	Home owner	Marital status	Defaulted borrower	Annual income
1	YES	SINGLE	NO	125K
2	NO	MARRIED	NO	100K
3	NO	SINGLE	NO	70K
4	YES	MARRIED	NO	120K
5	NO	DIVORCED	YES	95K
6	NO	MARRIED	NO	60K
7	YES	DIVORCED	NO	220K
8	NO	SINGLE	YES	85K
9	NO	MARRIED	NO	75K
10	NO	SINGLE	YES	90K

Find the conditional independence for given training set using BAYES theorem for classification.

(b) Consider the following dataset:

Instance	a_1	a_2	a_3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

(i) What is the entropy of this collection of training examples, with respect to the positive class?
(ii) What are the information gain of a_1 and a_2 relative to these training examples?
(iii) For a_3, compute information gain for every possible spilt.

UNIT – IV

8 (a) What is clustering analysis? Explain different types of clustering with an example

(b) What are the basic approaches for generating a agglomerative hierarchical clustering? Explain the algorithm.

OR

9 (a) Explain frequent pattern based clustering methods.

(b) Explain DBSCAN.

UNIT – V

10 (a) How does the Lossy Counting algorithm find frequent items?

(b) Explain Hoeffding Tree Algorithm.

OR

11 (a) Explain a pattern-growth algorithm for frequent substructure mining.

(b) What kinds of associations can be mined in multimedia data explain with example?

*****