BE6-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1.	**Answer question 1 and any FOUR from questions 2 to 7.**
2.	**Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1. (a) What is the importance of DM in KDD process?

(b) What is the difference between a data warehouse and a data mart?

(c) What is Apriori Property? How is the Apriori property used in the Apriori algorithm?

(d) How can the topology of the neural network be designed?

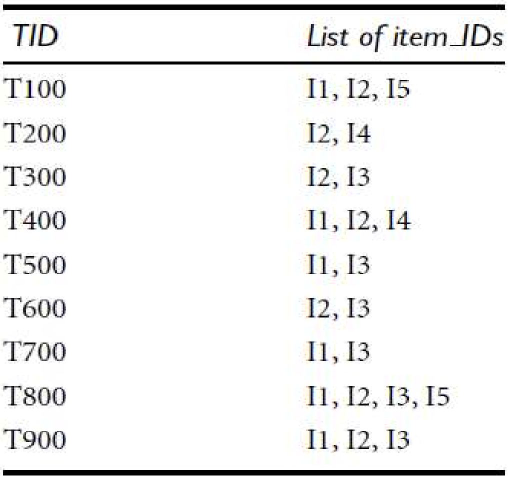
(e) The data set you have selected for analysis is Huge, which is sure to slow down the mining process. Which data preprocessing technique will you use to reduce the size of the data set, without jeopardizing the data mining results?

(f) What is Cluster Analysis? Explain with suitable example.

(g) What is Hypothesis Testing? How it is generated? Explain with an example. **(7×4)**

2. (a) What are major components of the architecture of a typical data mining system? Explain with neat sketch.

(b) Let the transaction database, D contains nine transactions in this database, Find all frequent item-sets using the Apriori algorithm. Also find strong association rules. Minimum Support count is 2 and minimum confidence threshold is 70%. **(9+9=18)**



TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

3. (a) Why not perform on-line analytical processing directly on databases instead of spending additional time and resources to construct a separate data warehouse?

(b) Consider the following data set shown below where each record represents the weather condition and class attributes shows whether people generally play sports in that weather condition or not.

Name	Gender	Height (In Metres)	Output
A	FEMALE	1.6	SHORT
B	MALE	2	TALL
C	FEMALE	1.9	MEDIUM
D	FEMALE	1.88	MEDIUM
E	FEMALE	1.7	SHORT
F	MALE	1.85	MEDIUM
G	FEMALE	1.6	SHORT
H	MALE	1.7	SHORT
I	MALE	2.2	TALL
J	MALE	2.1	TALL
K	FEMALE	1.8	MEDIUM
L	MALE	1.95	MEDIUM
M	FEMALE	1.9	MEDIUM
N	FEMALE	1.8	MEDIUM
O	FEMALE	1.75	MEDIUM

Draw the decision tree for the above table using ID3 algorithm.

(c) What is Data Stream? What are unique features of data stream? **(6+6+6)**

4. (a) What is noise? Given a numerical attribute, how can we "smooth" out the data to remove the noise. Explain with suitable example.

(b) What is meant by authoritative Web pages? How can a search engine automatically identify authoritative Web pages for given topic? **(8+10)**

5. (a) What is evolution analysis? Explain with suitable example.

(b) Why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distanced-based outlier detection, density-based local outlier detection, and deviation-based outlier detection.

(c) A data-warehouse for a university consists of four dimensions-student, course, semester, instructor. Two measures are maintained-count and average grade. Average grade is the average grade for a course, semester, instructor at the lowest level; count is the number of students. Draw a star schema for the data-warehouse. **(4+8+6)**

6. (a) A *spatiotemporal data stream* contains spatial information that changes over time, and is in the form of stream data. Discuss what kind of interesting knowledge can be mined from such data streams, with limited time and resources.

(b) How can we model data that does not show a linear dependence? What if a given response variable and predictor variable have a relationship that may be modeled by a polynomial function? **(10+8)**

7. (a) Your task as a Data Analytics at Software Company is to design a data mining system to examine university course database, which contains the following information: the name, address, and status of each student, the courses taken, and their cumulative grade point average (CGPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?

(b) What is Neural Network? Briefly describe multilayer-feed forward neural network.

(c) For the given six 2-dimensional data points (2,2),(2,3),(3,3),(1,2),(10,5),(10,8), find two clusters using k-means clustering algorithm assuming initial cluster centers are (2,3) and (10,5). Show cluster centers after each iteration and iterate the k-means algorithm for two times only. **(6+6+6)**