C5-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 are Multidimensional Association Rule? Explain in brief.
b) What do you mean by predictive and descriptive data mining?
c) Give a short example to show that items in a strong association may be negatively correlated.
d) Differentiate between lazy and eager learning.
e) Explain data smoothing techniques to remove the noise.
f) Why are decision tree classifiers so popular?
g) How multi-dimensional analysis in multimedia data is conducted?

(7x4)

2.

a) Differentiate between Star Schema and Snowflake Schema
b) From the architecture point of view, there are three data warehouse models: enterprise warehouse, data mart and virtual warehouse. Explain Data mart and Virtual warehouse.
c) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 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.
   i) Use min-max normalization to transform the value 35 for age onto the range [0:0; 1:0].
   ii) Use z-score normalization to transform the value 35 for age. Standard deviation for the age is 12.94 years.

(6+6+6)

3.

a) Suppose that a data warehouse for a big university consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg_grade. When 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 a snowflake schema diagram 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 one perform in order to list the average grade of CS courses for each Big University student.
b) Discuss following terms with suitable example:
   i) Schema Hierarchy
   ii) Set-grouping Hierarchy
c) What is Numerosity Reduction? Explain parametric methods for Numerosity Reduction.

(8+6+4)

4.

a) Describe the following correlation measures used in association rule mining:
   i) Lift
   ii) All_Confidence
   iii) Cosine

b) Give advantages and disadvantages of FP-Tree Algorithm.

c) Suppose there is a purchase data of games and videos. Out of 10,000 transactions analyzed, 6000 customer transactions included game, 7500 included videos and 4000 included both game and videos. Prepare a contingency table and check which type of correlation (positive or negative) exists between game and video.

(6+6+6)

5.

a) Consider the data set shown in the following table.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Using Naïve Bayesian Classifier, find the suitable class label (buys_computer = “yes” or buys_computer = “no”) for given sample:
X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

b) There are various classification methods. Differentiate between classification and prediction. How genetic algorithm can be used for classification?

c) Describe various methods which evaluate the accuracy of a classifier or a predictor.

(8+6+4)

6.

a) Outline the steps of ‘Clustering using Representative (CURE)’ algorithm.

b) How the accuracy of the text retrieval system can be accessed?

c) What is hierarchical clustering? Explain how agglomerative method is different from divisive clustering method.

(6+6+6)

7.

a) Simple classification of web mining is ‘Web Content Mining’ and ‘Web Usage Mining’. Write a short note on ‘Web Usage Mining’.

b) How does classification and prediction help in mining multimedia data? What kind of associations can be mined in multimedia data?

(9+9)