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 the goals of a data warehouse?

b) List out the significant issues in automatic cluster detection.

c) Explain various methods for generation of concept hierarchy for categorical data

d) How can we compute dissimilarity between two interval-scaled variables?

e) Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.

f) Describe following methods which evaluate the accuracy of a classifier

i) Holdout Method

ii) Random subsampling

iii) K-fold cross validation

g) Is strong association rule always feasible? Justify with example.

(7x4)

2.

a) Differentiate between Multidimensional data modeling and Relational Data Modeling.

b) Explain 3-tier Data Warehouse Architecture.

c) Suppose that a data warehouse for 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.

(3+6+9)

3.

a) Given the following transactional database. Generate all frequent itemset using Apriori algorithm with minimum support 30%

TID	ITEM
1	C,B,H
2	B,F,S
3	A,F,G
4	C,B,H
5	B,F,G
6	B,E,O

b) Describe Roll-up, Slice, Dice, Pivot and Drill-down OLAP operations

c) What business analyst gain from having a data warehouse?

(8+6+4)

4.

a) Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.

b) Describe Possible improvements for Apriori Algorithm.

c) Consider the sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34. Bins=3

Perform i) Partition into equal-frequency bins
ii) Smoothing by bin mean
iii) Smoothing by bin boundaries

(6+6+6)

5.

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

No.	**Outlook**	**Temperature**	**Humidity**	**Windy**	**Class**
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

Show how the induction of a decision tree is done using the information gain measure?

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

c) Discuss application of data warehousing and data mining in government sectors.

(8+6+4)

6.

a) Briefly discuss the various types of data that are considered in cluster analysis.

b) What are the advantages of Self Organizing Maps (SOM)? List also its application.

c) Discuss following clustering methods:
i) Hierarchical Methods
ii) Density-based Methods
iii) Grid-based methods

(6+3+9)

7.

a) What is tilted time frame in stream data analysis? Explain different methods to design titled time frame with example.

b) Write a short note on web usage mining.

c) Why trend analysis is performed on time series database?

(8+6+4)