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 different interestingness measures for pattern evaluation in data mining?

b) Differentiate between Enterprise data warehouse and data mart.

c) What is the difference between supervised and unsupervised learning? Give one example of each technique.

d) Why strong association rule is not always interesting? Explain with example.

e) What are the criteria to evaluate/compare classification and prediction methods?

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

g) Classify the following attributes as binary, discrete or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio).
   i) Times in terms of AM or PM
   ii) Angles as measures in degrees between 0 and 360
   iii) Number of patients in hospital
   iv) Olympics medal

(7x4)

2.

a) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.
   i) Draw a star schema diagram for the data warehouse.
   ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2004?
   iii) Bitmap indexing is useful in data warehousing. Taking this cube as an example, briefly discuss advantages and problems of using a bitmap index structure.

b) Differentiate OLAP and OLTP with respect to:
   i) Users and system orientation
   ii) Data contents
   iii) Database Design
   iv) View
   v) Access pattern

(10+8)

3.

a) What is decision tree induction? How is decision trees used for classification? Write Basic algorithm for inducing a decision tree from training tuples.

b) Explain conflict resolution strategies to handle situation when more than one rule is triggered for given tuple x in rule based algorithm.

c) List strengths and weakness of neural network as classifier.

(9+6+3)

4.

a) What is noise? Describe the possible reasons for noisy data. Explain the different techniques to remove the noise from data.

b) Suppose that the following table is derived by attribute-oriented induction.

class	birth place	count
programmer	USA	180
	others	120
DBA	USA	20
	others	80

i) Transform the table into a crosstab showing the associated t-weights and d-weights.

ii) Map the class Programmer into a (bidirectional) quantitative descriptive rule, for example,

$\forall X$, Programmer(X) $\Leftrightarrow$ (birth_place(X) = "USA"$\wedge$…)

[t : x% , d : y%] …Θ(…) [t : w%, d : z%]

c) Briefly describe case base reasoning classifier.

(9+5+4)

5.

a) Apriori algorithm is used to find the frequent item sets from candidate dataset. Explain the major two steps of the algorithm.

b) Explain three-tier data warehouse architecture.

c) What are the steps in KDD process? Explain in brief.

(6+6+6)

6.

a) Describe each of the following clustering algorithms in term of the criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations.

i) k-means

ii) BRICH

iii) DBSCAN

b) What are the requirements of clustering in data mining?

(9+9)

7.

a) What are the differences between mining association rules in multimedia databases versus that in transaction databases?

b) Write a short note on web usage mining.

c) TF-IDF has been used as an effective measure in document classification.

i) Give one example to show that TF-IDF may not be always a good measure in document classification.

ii) Define another measure that may overcome this difficulty.

(6+6+6)