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 major challenges of mining a huge amount of data in comparison with mining a small amount of data?

b) Briefly outline the major steps of decision tree classification?

c) Prove that all nonempty subsets of a frequent itemset must also be frequent in Apriori algorithm.

d) Briefly outline how to compute the dissimilarity between objects described by the following:
   1) Nominal attributes
   2) Numeric attributes

e) Are there cases where accuracy may not be appropriate to evaluate performance of classifier? Justify with appropriate example.

f) What can business analysts gain from having a data warehouse?

g) Explain with the help of suitable example that items in a strong association rule may actually be negatively correlated.

(7x4)

2.

a) What is under fitting and over fitting in decision tree algorithm? How is tree pruning useful to handle these issues in decision tree induction? Explain any one method of tree pruning.

b) What are the limitations of support and confidence framework to assess interestingness of association rules? How this limitation can be overcome using lift measure? Explain with suitable example.

(9+9)

3.

a) Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing?

b) Present conditions under which density-based clustering is more suitable than partitioning-based clustering and hierarchical clustering. Give application examples to support your argument

c) 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.
   1) Draw a star schema diagram for the data warehouse.
   2) 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 2010?
   3) Bitmap indexing is useful in data warehousing. Taking this cube as an example, briefly discuss advantages and problems of using a bitmap index structure.

(3+6+9)

4.

a) What kind of data mining can be performed on spatial databases?

b) How are data actually stored in ROLAP and MOLAP architectures? Explain with a suitable example.

c) Why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distance-based outlier detection, and deviation-based outlier detection.

(5+6+7)

5.

a) Why is naïve Bayesian classification called "naïve"? Briefly outline the major ideas of naïve Bayesian classification.

b) A data cube, *C*, has *n* dimensions, and each dimension has exactly *p* distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions.

1) What is the *maximum number of cells* possible in the base cuboid?

2) What is the *minimum number of cells* possible in the base cuboid?

3) What is the *maximum number of cells* possible (including both base cells and aggregate cells) in the data cube, *C*?

4) What is the *minimum number of cells* possible in the data cube, *C*?

(9+9)

6.

a) Use the methods below to normalize the following group of data:

20, 30, 40, 60, 100

1) min-max normalization by setting min = 0 and max = 1.

2) z-score normalization.

3) z-score normalization using the mean absolute deviation instead of standard deviation.

4) normalization by decimal scaling.

b) How can the data be preprocessed in order to help improve the quality of the data and consequently of the mining results?

c) Briefly describe the classification processes using

i) genetic algorithms,

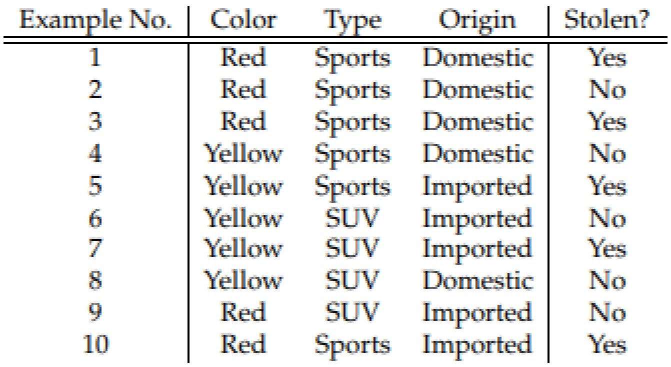
ii) rough sets, and

iii) fuzzy sets.

(6+6+6)

7.

a) The following table consists of training data from an car theft database with attributes Color , Type , Origin, and the class label stolen can be either yes or no..



Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Given a data tuple having the values “Red”, “Domestic”, “SUV” for the attributes color, Type, and Origin, respectively, what would a naive Bayesian classification of the stolen status for the tuple be?

b) Explain the following Methodologies for Stream Data Processing and Stream Data Systems.

1) Random Sampling,

2) Sliding Windows

3) Sketches

(9+9)