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) 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 or simple statistical analysis?

b) Describe why concept hierarchies are useful in data mining.

c) The Apriori algorithm makes use of prior knowledge of subset support properties. Prove that all nonempty subsets of a frequent itemset must also be frequent.

d) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

e) Show that accuracy is a function of sensitivity and specificity.

f) What is boosting? State why it may improve the accuracy of decision tree induction.

g) What is the difference between supervised and unsupervised learning? Give an example each techniques.

(7x4)

2.

a) Briefly outline how to compute the dissimilarity between objects described by the following:
   i) Nominal attributes
   ii) Asymmetric binary attributes
   iii) Numeric attributes

b) 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 2010?
   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.

(9+9)

3.

a) What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)? Explain with the help of an example.

b) The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, hotdogs refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and hamburgers refers to the transactions that do not contain hamburgers.

	$hot\ dogs$	$\overline{hotdogs}$	Σ_{row}
$hamburgers$	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
Σ_{col}	3000	2000	5000

i) Suppose that the association rule "hot dogs $\Rightarrow$ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

ii) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?

c) Describe each of the following clustering algorithms in terms of the following criteria: (1) shapes of clusters that can be determined; (2) input parameters that must be specified; and (3) limitations.

i) k-means

ii) k-medoids

iii) CLARA

(3+6+9)

4.

a) Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semitight coupling, and tight coupling. State which approach you think is the most popular, and why?

b) In data warehouse technology, a multiple dimensional view can be implemented by a relational database technique (ROLAP), or by a multidimensional database technique (MOLAP), or by a hybrid database technique (HOLAP). Briefly describe each implementation technique.

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

(5+6+7)

5.

a) Briefly describe the classification processes using i) genetic algorithms, ii) rough sets, and iii) fuzzy sets.

b) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

c) Suppose that a base cuboid has three dimensions, A, B, C, with the following number of cells: |A| = 1, 000, 000, |B| = 100, and |C| = 1000. Suppose that each dimension is evenly partitioned into 10 portions for chunking.

i) Assuming each dimension has only one level, draw the complete lattice of the cube.

ii) If each cube cell stores one measure with 4 bytes, what is the total size of the computed cube if the cube is dense?

(6+6+6)

6.

a) What is time-series database? How to characteristics the time-series data using trend analysis.

b) State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.

c) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

(6+6+6)

7.

a) The following table consists of training data from an employee database. The data have been generalized. For example, "31 . . . 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let status be the class label attribute.
Given a data tuple having the values “systems”, “26. . . 30”, and “46–50K” for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?

b) Explain the following Methodologies for Stream Data Processing and Stream Data Systems.
   i) Random Sampling
   ii) Sliding Windows
   iii) Sketches

(9+9)