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) Describe the types of situations that produce sparse or dense data cubes.

b) Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as PCA and SVD.

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

d) Explain how to categorize data mining system.

e) Why SVM is more effective on High Dimensional Data compared to Multi-Layer Feed-Forward Neural Network?

f) What is staging area? Why it need?

g) What are the differences in the analysis of stream data from that of relational and warehouse data?

(7x4)

2.

a) 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.

c) A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another.

(9+9)

3.

a) What is noisy data? Explain different smoothing techniques to smooth out noise in the dataset.

b) Use a flow chart to summarize the following procedures for attribute subset selection:

   i) stepwise forward selection

   ii) stepwise backward elimination

(10+8)

4.

a) What is frequent itemsets? How frequent itemsets can be mined without generating candidate sets?

b) 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.

c) What is a drawback of using a separate set of tuples to evaluate pruning in decision tree Induction?

(9+6+3)

5.

a) Briefly outline the major steps of decision tree classification.

b) It is difficult to assess classification *accuracy* when individual data objects may belong to more than one class at a time. In such cases, comment on what criteria you would use to compare different classifiers modeled after the same data.

c) Compare the advantages and disadvantages of eager classification versus lazy classification.

(6+6+6)

6.

a) Discuss different approaches to compute the dissimilarity between objects of mixed variable types.

b) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters:

$A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$:

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show

i) The three cluster centers after the first round execution.

ii) The final three clusters.

(10+8)

7.

a) What are Bayesian classifiers? Explain briefly Baye's theorem. Also explain how Naive Bayesian classifier works?

b) Write a short note on web usage mining.

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

(8+6+4)