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) Mention the different characteristics of data warehouse.

b) What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining? Discuss the motivation behind OLAP mining.

c) Why OLAP is performed? Mention 3 different OLAP operations along with examples.

d) Define the measures: Accuracy, Error rate, Precision, and Recall measure using True Positive, True Negative, False Positive, and False Negative.

e) Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Taking fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

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

g) Write short notes on different tools used for data mining.

(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?

b) Suppose the data for analysis includes the attribute Age. The age values for the data tuples(instances) are (in increasing order): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25,30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   Use binning (by bin means) to smooth the above data, using a bin depth of 3. Illustrate your steps, and comment on the effect of this technique for the given data**.**

(12+6)

3.

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) Describe the steps involved in Data Mining when viewed as a process of knowledge discovery.

(9+9)

4.

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

b) Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(4+14)

5. A database has four transactions. Let min sup = 50% and min conf = 80%.

TID	items bought
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

a) Find all frequent itemsets using Apriori

b) Find all frequent rules.

(9+9)

6.

a) There are several major differences between biological sequential patterns and transactional sequential patterns. First, in transactional sequential patterns, the gaps between two events are usually nonessential. For example, the pattern “purchasing a digital camera two months after purchasing a PC" does not imply that the two purchases are consecutive. However, for biological sequences, gaps play an important role in patterns. Second, patterns in a transactional sequence are usually precise. However, a biological pattern can be quite imprecise, allowing insertions, deletions, and mutations. Discuss how the mining methodologies in these two domains are influenced by such differences.

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

(9+9)

7.

a) Briefly describe the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, and constraint-based methods.

b) Consider the following 7 points:

Point	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Point 1 and Point 4 are initially assigned to cluster C1 and C2 respectively. Apply the k-means (k=2) algorithm using the Euclidean distance to cluster the above data.

(9+9)