CE3-R3: DATA WAREHOUSING AND MINING

NOTE:

> **1. Answer question 1 and any FOUR questions from 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) List the advantages of update driven approach (data warehousing) over query driven approach (wrapper and integrators).

b) What is spatial data mining? Give two applications of spatial data mining.

c) Data mining is a multi-disciplinary area which draws heavily from statistics, artificial intelligence, high performance computing and databases among others. Give one idea from each of these fields that is used in data mining.

d) What are the measurements for pattern interestingness? Explain each measurement in brief.

e) What is classification? Explain with the help of an example. List four classification techniques.

f) What are the steps in the process of knowledge discovery?

g) What is a concept hierarchy? How is it used during data mining?

(7x4)

2.

a) List and explain four differences between OLAP and OLTP.

b) What is cluster analysis? What are the criteria to judge goodness of clusters? Give two applications of cluster analysis.

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

(6+8+4)

3.

a) 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. Give your opinion of which might be more empirically useful and state the reasons for your opinion.

b) What do you understand by Market Basket Analysis? Explain in brief with example.

c) What is meant by authoritative web page? How can a search engine automatically identify authoritative web pages for the topic? How can we use hub pages to find authoritative pages?

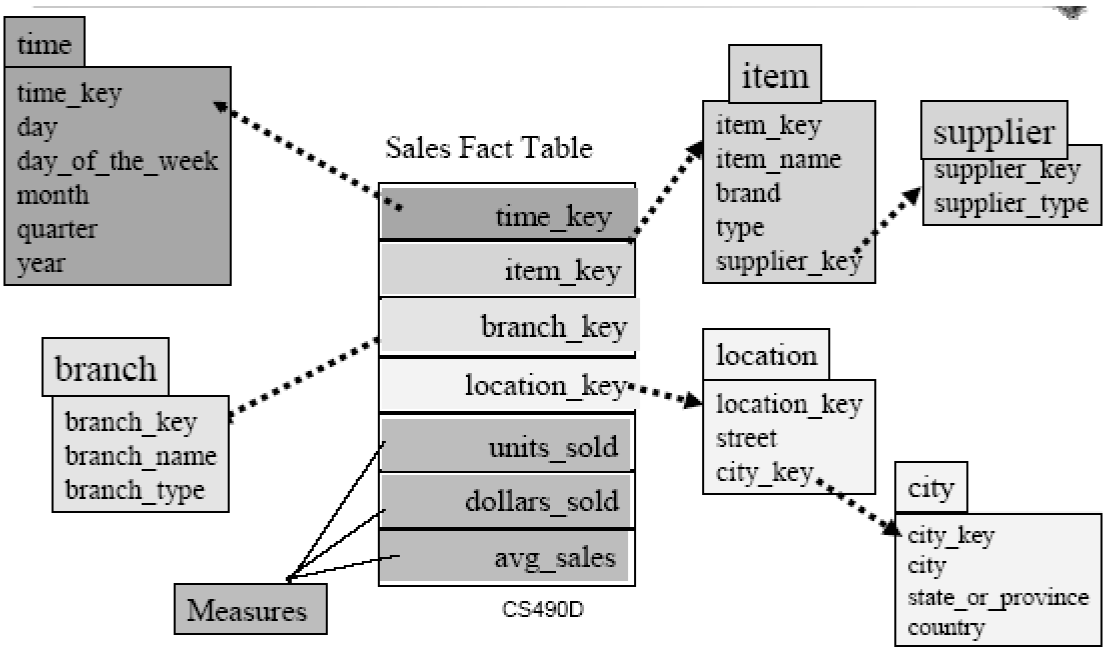
d) List four characteristics of a clustering algorithm.

(6+4+6+2)

4.

a) What is an artificial neural network? What is the difference between a feed-forward and back propagation network?

b) Following is a snowflake schema. Define the snowflake schema in DMQL.



A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—Explain each of these characteristics in detail.

(5+5+8)

5.

a) On-Line Analytical Mining (OLAM) integrates on-line analytical processing with data mining and mining knowledge in multidimensional databases. Describe the importance of OLAM. Also explain an integrated OLAM and OLAP architecture with diagram.

b) Explain Apriori Algorithm for finding frequent itemsets using candidate generation with an example.

c) What do you understand by data cleaning? Why is it an important step in the discovery process?

(8+6+4)

6.

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

b) List and explain the following OLAP operations, with the help of an example.

i) Roll-up
ii) Drill-down
iii) Slice-and-dice
iv) Pivot

c) Why there is a need for data preprocessing? What are the data preprocessing tasks?

(8+6+4)

7.

a) Write a generic algorithm for decision tree induction. What are the two approaches to avoid overfitting in classification?

b) Give mathematical formulation of itemset and frequent itemset, support and confidence? Explain with the help of an example.

c) List and explain one application of each of the following:
   i) Mining Image database
   ii) Text mining
   iii) Temporal data mining

iv) Web usage mining

(8+6+4)