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) Is strong association rule always feasible? Justify with example.

b) Differentiate between Enterprise Data Warehouse and Data Mart.

c) Given the following measurements of the variable profit

200, 300, 400, 600, 1000

Standardized the variable by following:

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

ii) z-score normalization

d) Why discriminate rule provides a *sufficient condition, but not a necessary one,* for tuple to be in the target class?

e) What is the use of Regression? What are the reasons for not using the linear regression model to estimate the output data?

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

g) Differentiate between support vector machine and neural networks.

(7x4)

2.

a) Briefly compare the following concepts. You may use an example to explain your point(s).
   i) Snowflake schema, fact constellation, starnet query model
   ii) Data cleaning, data transformation, refresh
   iii) Supervised and Unsupervised Learning

b) Discuss distributive, algebraic and holistic measures.

c) What is concept hierarchy? How are concept hierarchies useful in OLAP?

(9+6+3)

3.

a) Define sampling. Explain different type of sampling with example. What is the advantage of applying sampling to dataset?

b) How multilevel association rules can be mined efficiently using concept hierarchy?

(10+8)

4.

a) Describe the list of techniques for improving the efficiency of Apriori-based mining.

b) Explain how tree pruning performs in decision tree induction?

c) How can distance be computed for attributes that having missing valves in K-Nearest Neighbor classifier?

(9+6+3)

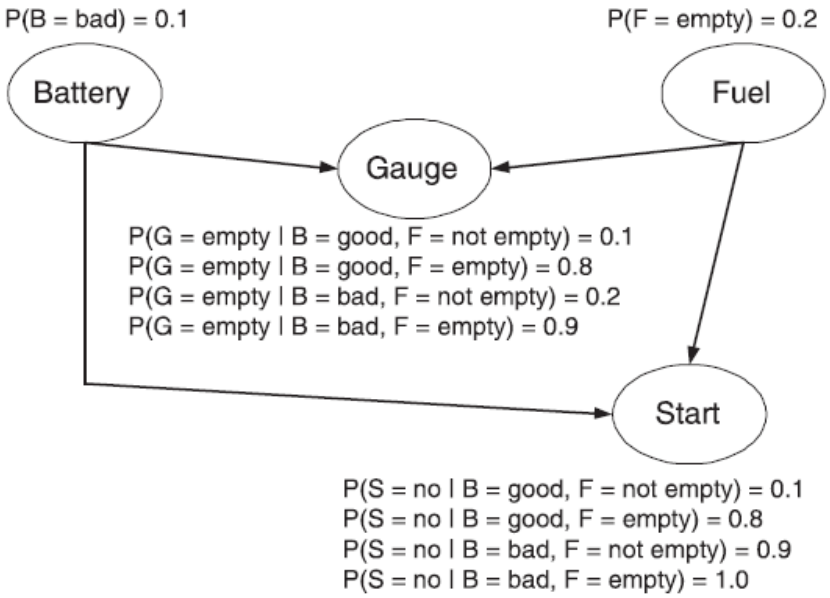
5.

a) Given then Bayesian network shown in below figure, compute the following probabilities:

i) P(B=good, F=empty, G=empty, S=yes)

iii) P(B=bad, F=empty, G=not empty, S=no)

iii) Given that the battery is bad, computer the probability that car will start



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

c) What are the advantages of Self Organizing Maps (SOM)? List also its application.

(9+6+3)

6.

a) Briefly describe the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. Give examples in each case.

b) Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm. Also, illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme.

c) Why BIRCH encounters difficulties in finding clusters of arbitrary shape but OPTICS does not?

(10+6+2)

7.

a) What is tilted time frame in stream data analysis? Explain different methods to design titled time frame with example.

b) What is Data generalization? Discuss basic principle of Attribute Oriented Indication.

(9+9)