Practice Paper 3

Data Warehousing and Mining

Mumbai University - BE Computer Engineering/AIDS - Semester V

Time: 3 hours
Max. Marks: 80

Instructions:

1. Question 1 is compulsory.
2. Attempt any 3 questions out of the remaining.
3. Assume suitable data if required.

Q.1 (Compulsory - Attempt All) [20 Marks]

(a) Why does every data structure in the data warehouse contain the time element? Justify your answer. [5 Marks]

(b) Define Metadata. Explain different types of metadata used in data warehousing. [5 Marks]

(c) What is outlier? Explain different types of outliers with examples. [5 Marks]

(d) Explain with example: Support, Confidence, and Lift function in association mining. [5 Marks]

Q.2 [20 Marks]

(a) Describe the various phases in the Knowledge Discovery (KDD) process with a neat diagram. [10 Marks]

(b) Calculate Accuracy, Recall, and Precision with the help of following data: [10 Marks]

- True Positive (TP) = 55
- True Negative (TN) = 40
- False Positive (FP) = 12
- False Negative (FN) = 18

Also, explain the significance of each metric in evaluating a classification model.

Q.3 [20 Marks]

(a) Explain the Extraction and Transformation phases in the ETL process with suitable examples. Discuss the common transformation operations. [10 Marks]

(b) Illustrate multidimensional association rules with suitable examples. Explain how they differ from single-dimensional association rules. [10 Marks]

Q.4 [20 Marks]

(a) Define classification. Discuss the major issues in classification. Explain Naive Bayesian classification with an example and show probability calculations. [10 Marks]

(b) Calculate the mean, median, mode, midrange, and variance for the following dataset: [10 Marks]

Dataset: 12, 15, 18, 18, 20, 22, 22, 22, 25, 28, 30, 30, 35, 38, 40, 42, 45, 50, 52, 55

Show all calculations step by step.

Q.5 [20 Marks]

(a) Explain the HITS (Hyperlink Induced Topic Search) algorithm. Illustrate its working with a sample web graph showing the calculation of hub and authority scores for at least 4 web pages. [10 Marks]

(b) Explain web structure mining in detail. Discuss its applications and how it differs from web content mining. [10 Marks]

Q.6 [20 Marks]

(a) What is clustering? Explain the K-means clustering algorithm with a flowchart. Apply K-means on the following dataset with K=2: [10 Marks]

Dataset: {4, 8, 15, 21, 25, 34, 38, 43, 49}

Initial centroids: C1 = 4, C2 = 49

Show all iterations with distance calculations until convergence.

(b) Explain the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm with an appropriate diagram. Discuss its advantages. [10 Marks]

Q.7 [20 Marks]

(a) Consider the following transaction database: [10 Marks]

T_ID	Items Purchased
T01	Milk, Bread, Butter, Eggs
T02	Milk, Bread, Cheese
T03	Bread, Butter, Eggs
T04	Milk, Bread, Butter, Cheese
T05	Milk, Butter, Eggs
T06	Bread, Cheese, Eggs

T_ID	Items Purchased
T07	Milk, Bread, Eggs

Use **Apriori Algorithm** with minimum support = 40% and minimum confidence = 75% to find all frequent itemsets and generate strong association rules.

(b) Explain different data sampling techniques with examples: [10 Marks]

1. **Simple Random Sampling**
2. **Stratified Sampling**
3. **Cluster Sampling**
4. **Systematic Sampling**

Discuss when each technique is most appropriate.

END OF PAPER

Prepared by: Nitin Gupta