Practice Paper 1

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) Explain features of Data Warehouse. [5 Marks]

(b) What is Market Basket Analysis? Explain with example. [5 Marks]

(c) Draw and explain the KDD (Knowledge Discovery in Databases) process with a neat diagram. [5 Marks]

(d) Calculate Accuracy, Recall, and Precision with the help of following data: [5 Marks]

- True Positive (TP) = 60
- True Negative (TN) = 30
- False Positive (FP) = 15
- False Negative (FN) = 20

Q.2 [20 Marks]

(a) Explain Star Schema and Snowflake Schema with suitable examples. Draw diagrams for both schemas. [10 Marks]

(b) Explain any four OLAP operations on multidimensional data with examples. [10 Marks]

Q.3 [20 Marks]

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

TID	Items
T1	A, B, C, E
T2	B, D, E
T3	A, B, D, E
T4	A, B, C, E

TID	Items
T5	A, C, D
T6	B, C, E

Use **Apriori Algorithm** with minimum support = 50% and minimum confidence = 70% to find all frequent itemsets and strong association rules.

(b) What is data preprocessing? Explain in detail any three data cleaning techniques. [10 Marks]

Q.4 [20 Marks]

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

Dataset: {5, 10, 15, 20, 25, 30, 35, 40, 45}

Show all iterations until convergence.

(b) The following table contains a training dataset of class-labeled tuples. Using Naive Bayesian Classification, predict the class label for the tuple X = (Age = Youth, Income = Medium, Student = Yes, Credit_rating = Fair). [10 Marks]

RID	Age	Income	Student	Credit_rating	Buys_computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle	Medium	No	Excellent	Yes
13	Middle	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

Q.5 [20 Marks]

(a) What is web mining? Explain the HITS (Hyperlink-Induced Topic Search) algorithm with a suitable example. [10 Marks]

(b) Explain the Decision Tree Induction algorithm for classification. Discuss the role of Information Gain in building the decision tree. [10 Marks]

Q.6 [20 Marks]

(a) Explain the ETL (Extract, Transform, Load) process in detail with a suitable diagram. [10 Marks]

(b) Explain the DBSCAN clustering algorithm with an appropriate diagram. Discuss its advantages over K-means. [10 Marks]

END OF PAPER

Prepared by: Nitin Gupta