Practice Paper 2

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) Demonstrate with a diagram the architecture of a typical Data Mining system. [5 Marks]

(b) What is prediction? Explain the Linear Regression method with an example. [5 Marks]

(c) Explain confusion matrix, accuracy, and precision with a suitable example. [5 Marks]

(d) What is hierarchical clustering? Explain the divisive clustering approach. [5 Marks]

Q.2 [20 Marks]

(a) Suppose that a data warehouse for a Sports Academy consists of four dimensions: Date, Player, Coach, and Sport, and two measures: count and performance_score. [10 Marks]

Tasks:

1. Draw a **Star Schema** diagram for this data warehouse.
2. Create a base cuboid for dimensions [Date, Player, Sport].
3. Apply any three OLAP operations on this cuboid (Roll-up, Drill-down, Slice).

(b) Briefly outline with examples how to compute dissimilarity between objects for: [10 Marks]

1. **Nominal attributes**
2. **Asymmetric binary attributes**

Q.3 [20 Marks]

(a) Explain the three-tier data warehouse architecture in detail. [10 Marks]

(b) Define multidimensional and multilevel association mining with suitable examples. [10 Marks]

Q.4 [20 Marks]

(a) The following table shows protein and fat content of 6 food items. Apply Agglomerative Hierarchical Clustering with single linkage method and draw a dendrogram. [10 Marks]

Food Item	Protein	Fat
1	2.0	50
2	7.5	18
3	5.0	30
4	2.5	22
5	8.0	12
6	3.0	45

Note: Use Euclidean distance to calculate dissimilarity.

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

T_ID	Items
T100	P, Q, R, S
T200	P, R, T
T300	Q, R, S, T
T400	P, Q, T
T500	P, Q, R, T

Use **Apriori Algorithm** with minimum support count = 3 and minimum confidence = 65% to find frequent itemsets and strong association rules.

Q.5 [20 Marks]

(a) What is clustering? Explain the K-means clustering algorithm. Apply the algorithm on the following dataset with K=3: [10 Marks]

Dataset: {3, 8, 12, 18, 22, 28, 33, 38, 42}

Show all iterations and final cluster assignments.

(b) Explain web content mining in detail. Discuss its applications and challenges. [10 Marks]

Q.6 [20 Marks]

(a) Illustrate any one classification technique for the following dataset. Show how to classify a new tuple with (Experience = High, Qualification = Masters, Salary_Expected = Medium). [10 Marks]

Sr. No	Experience	Qualification	Salary_Expected	Hired
1	High	Masters	High	Yes
2	Low	Bachelors	Low	Yes
3	Medium	Masters	Medium	Yes
4	High	Bachelors	High	No
5	Low	PhD	Low	Yes
6	Medium	Bachelors	Medium	Yes
7	High	PhD	High	Yes
8	Low	Masters	Low	Yes
9	Medium	PhD	Medium	Yes
10	High	Bachelors	Medium	No

(b) Explain the following with examples: [10 Marks]

1. **Different data sampling techniques** (Random, Stratified, Systematic)
2. **Data normalization techniques** (Min-Max, Z-score)

END OF PAPER

Prepared by: Nitin Gupta