Atal Bihari Vajpayee Indian Institute of Information Technology & Management, Gwalior

IT405: Data Mining

Major Examination (Session 2023–24)

Maximum Time: 3 Hours **Max Marks: 70**

Note: Answer all questions. Wherever required, show intermediate steps and state assumptions.

1. (a) Define data mining. List and briefly explain four common tasks in data mining. (8 Marks)
   (b) What is the difference between data mining and machine learning? Give examples. (4 Marks)

2. (a) Describe the Apriori algorithm for frequent itemset generation (steps and pruning). (8 Marks)
   (b) Given transactions: T1: A,B,C T2: A,C T3: A,B T4: B,C T5: A,B,C Using minimum support = 40% and minimum confidence = 60%, find frequent itemsets and strong association rules. Show calculations. (10 Marks)

3. (a) Explain entropy and information gain used in decision tree learning (ID3/C4.5). Provide formulae and intuition. (6 Marks)
   (b) Given a dataset with binary class labels and attribute splits, compute information gain for one split (details provided in exam room by instructor) — show steps. (6 Marks)

4. **Clustering (numerical + concept) — 12 Marks**: (a) Describe the working of DBSCAN and how it differs from k-means (advantages/disadvantages). (5 Marks)
   (b) Given 2D points: (1,1),(1.5,2),(3,4),(5,7),(3.5,5),(4.5,5),(3.5,4.5) explain how DBSCAN with eps=1 and MinPts=3 would cluster them (show core/border/noise classification). (7 Marks)

5. (a) What are common methods for feature selection? Explain filter and wrapper approaches with pros/cons. (6 Marks)
   (b) Discuss curse of dimensionality and effects on distance-based algorithms. (4 Marks)

6. **Model Evaluation and Validation — 8 Marks**: (a) Explain k-fold cross validation and when to use stratified k-fold. (3 Marks)
   (b) A classifier yields: TP=80, FP=20, FN=30, TN=170. Compute Precision, Recall, F1-score, and Accuracy. (5 Marks)

7. **Case Study (15 Marks):** A retail chain wants to use data mining to (i) detect customer segments for targeted marketing, (ii) discover association rules for cross-selling, and (iii) predict customer churn. For each requirement: - Propose an appropriate data mining technique (clustering / association mining / classification), outline required data/features, preprocessing steps, evaluation metrics, and a short deployment plan (how models would be used in production). Include risks (privacy, data quality) and mitigation strategies.

8. Short notes (any two — 7 Marks each): (a) Ensemble methods (Bagging vs Boosting) — idea and when to prefer each.
(b) Dimensionality reduction using t-SNE and when to use it.
(c) Anomaly detection techniques — statistical vs machine learning based. (14 Marks total)