MATRUSRI ENGINEERING COLLEGE

Computer Science and Engineering

B.E. VI Semester — Data Mining (PC602CSU23)

UNIT-WISE QUESTION BANK

UNIT I — Introduction to Data Mining & Getting to Know Your Data

Syllabus topics: *Introduction to Mining and Data Mining; kinds of data, patterns, technologies and applications; major issues in data mining; data objects and attribute types; basic statistical descriptions of data; measuring data similarity and dissimilarity.*

A. Previous Exam Questions

1. What is the purpose of data mining? Give examples of specific applications of data mining. *[Sept/Oct 2023, Q1(a), 2 Marks]*

2. Differentiate between data similarity and data dissimilarity measures. *[Sept/Oct 2023, Q1(b), 2 Marks]*

3. Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of these three issues, discuss how the assessment of data quality can depend on the intended use of the data, giving examples. Propose two other dimensions of data quality. *[Sept/Oct 2023, Q2(b), 14 Marks]*

4. Describe the steps involved in data mining when viewed as a process of knowledge discovery. *[Sept/Oct 2023, Q7(a), 7 Marks]*

5. How is a data warehouse different from a database? How are they similar? *[Aug/Sept 2024, Q1(a), 2 Marks]*

6. Compare predictive and descriptive data mining techniques. *[Aug/Sept 2024, Q1(g), 2 Marks]*

7. Discuss the major issues in data mining. *[Aug/Sept 2024, Q2(a), 7 Marks]*

8. Describe the various phases in the knowledge discovery process with a neat diagram. *[Aug/Sept 2024, Q2(b), 7 Marks]*

9. Is it that data mining needs machine learning, but machine learning doesn't necessarily need data mining? Comment. *[Feb/March 2024, Q1(a), 2 Marks]*

10. Find the cosine similarity between two term-frequency vectors x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) and y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1). *[Feb/March 2024, Q1(f), 2 Marks]*

11. Define the data mining functionalities: association, classification, regression, clustering, and outlier analysis. *[Feb/March 2024, Q2(a), 7 Marks]*

12. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (i) Compute the Euclidean distance. (ii) Compute the Manhattan distance. (iii) Compute the Minkowski distance using q = 3. (iv) Compute the supremum distance. *[Feb/March 2024, Q2(b), 7 Marks]*

13. Examine a suitable data mining algorithm/task for each of the following scenarios: is this an apple or an orange? Is it sunny or overcast? Detecting credit card fraud. How is this organised (grouping)? Determining the relationship between hours of study and exam results. What should I do next? *[Feb/March 2024, Q7(a), 7 Marks]*

14. The age values for a set of data tuples are given in increasing order: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Give the five-number summary of the data and show a boxplot of the data. *[Feb/March 2024, Q7(b), 7 Marks]*

B. Additional Practice Questions

15. Define data mining. Explain how it relates to the overall Knowledge Discovery in Databases (KDD) process.

16. Explain, with a neat diagram, the architecture of a typical data mining system.

17. What kinds of data can be mined? Discuss with examples: relational databases, data warehouses, transactional data, and other kinds of data (time-series, sequence, graph, spatial, text and multimedia data).

18. What kinds of patterns can be mined? Briefly explain characterization/discrimination, frequent pattern mining, classification, cluster analysis, and outlier analysis.

19. List and briefly explain the major issues in data mining under the categories: mining methodology, user interaction, efficiency and scalability, and diversity of data types.

20. Explain the different types of attributes: nominal, binary, ordinal, and numeric, with an example of each.

21. Explain the basic statistical descriptions used to study data: measures of central tendency (mean, median, mode) and measures of data dispersion (range, quartiles, variance, standard deviation).

22. What is a boxplot? Explain how it visually represents the five-number summary of a dataset.

23. Explain proximity measures for nominal and binary attributes, distinguishing between symmetric and asymmetric binary variables with an example.

24. Derive the Euclidean distance and Manhattan distance as special cases of the Minkowski distance.

25. What is cosine similarity? In what situations is it preferred over Euclidean distance, and why?

26. Explain, with examples, common data visualization techniques (histograms, scatter plots, quantile plots) used for exploratory data analysis.

27. Compare a database system and a data warehouse in terms of purpose, schema design, and typical usage (OLTP vs OLAP).

UNIT II — Data Preprocessing & Mining Frequent Patterns, Associations and Correlations

Syllabus topics: *Data preprocessing overview, data cleaning, data reduction; basic concepts and methods of frequent pattern mining; frequent itemset mining methods (Apriori, FP-Growth); pattern evaluation methods.*

A. Previous Exam Questions

1. How does the support measure differ from the confidence measure in association rule mining? *[Sept/Oct 2023, Q1(c), 2 Marks]*

2. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. *[Sept/Oct 2023, Q2(a), 7 Marks]*

3. A database has five transactions with min_sup = 60% and min_conf = 80%: T100{M,O,N,K,E,Y}, T200{D,O,N,K,E,Y}, T300{M,A,K,E}, T400{M,U,C,K,Y}, T500{C,O,O,K,I,E}. (i) Find all frequent itemsets using Apriori and FP-Growth respectively; compare the efficiency of the two mining processes. (ii) List all the strong association rules matching the metarule $\forall x \in$ transaction, buys(X, item1) $\wedge$ buys(X, item2) $\Rightarrow$ buys(X, item3) [s, c]. *[Sept/Oct 2023, Q3, 14 Marks]*

4. What is the Apriori property? Explain with an example. *[Aug/Sept 2024, Q1(b), 2 Marks]*

5. Find the frequent itemsets and strong association rules for a 12-transaction database using the Apriori algorithm, with support = 30% and confidence = 40%. *[Aug/Sept 2024, Q3(a), 7 Marks]*

6. Illustrate with an example how to calculate the support and confidence of an association rule. *[Feb/March 2024, Q1(b), 2 Marks]*

7. Build a Frequent Pattern (FP) Tree for a given transaction database (T100–T400) with min_support = 2. *[Feb/March 2024, Q3(a), 7 Marks]*

8. A contingency table summarizes supermarket transaction data for hotdogs and hamburgers (hotdogs $\wedge$ hamburgers = 2000, ¬hotdogs $\wedge$ hamburgers = 500, hotdogs $\wedge$ ¬hamburgers = 1000, ¬hotdogs $\wedge$ ¬hamburgers = 1500). Compare the use of all_confidence, max_confidence and Kulczynski measures with lift on this data. *[Feb/March 2024, Q3(b), 7 Marks]*

B. Additional Practice Questions

9. Give an overview of data preprocessing. Why is it needed before applying data mining algorithms?

10. Explain the major tasks in data cleaning: handling missing values, smoothing noisy data, and resolving inconsistencies.

11. What is data reduction? Explain dimensionality reduction, numerosity reduction, and data compression with examples.

12. Explain data discretization and concept hierarchy generation for numeric data.

13. Define frequent itemset, closed frequent itemset, and maximal frequent itemset. Illustrate the differences with an example.

14. Explain the Apriori algorithm in detail, including the join step and prune step, with a worked example.

15. Explain how the FP-Growth algorithm mines frequent itemsets without candidate generation. Describe the construction of the FP-tree and the FP-growth mining process.

16. Compare the Apriori algorithm and the FP-Growth algorithm in terms of approach, memory usage, and scalability to large databases.

17. What are multilevel association rules and multidimensional association rules? Explain with examples.

18. Why is the support-confidence framework alone insufficient for identifying interesting association rules? Illustrate with the classic example of a misleading rule.

19. Explain the lift and chi-square measures for correlation analysis of association rules, and state when each is appropriate.

20. What is constraint-based (or pattern-directed) frequent pattern mining? Give examples of constraints that can be pushed into the mining process.

21. Explain the concept of pattern evaluation methods and why 'interestingness' measures beyond support and confidence are needed.

UNIT III — Classification

Syllabus topics: *Classification: basic concepts, decision tree induction, Bayes classification methods, rule-based classification, model evaluation and selection, techniques to improve classification accuracy.*

A. Previous Exam Questions

1. Define classification in the context of data mining. What is the primary goal of classification, and how does it differ from other data mining tasks? *[Sept/Oct 2023, Q1(d), 2 Marks]*

2. Write an algorithm for k-nearest neighbour classification given k (the number of neighbours) and n (the number of attributes describing each tuple). Explain with an example. *[Sept/Oct 2023, Q4, 14 Marks]*

3. Write formulas for input, output, and error calculation for a hidden layer node in a multilayer feed-forward neural network. *[Aug/Sept 2024, Q1(c), 2 Marks]*

4. What is classification? Explain briefly about the decision tree classifier. *[Aug/Sept 2024, Q4(a), 7 Marks]*

5. Why is naive Bayesian classification called 'naive'? Briefly outline the major ideas of naive Bayesian classification and explain Naive-Bayes classification. *[Aug/Sept 2024, Q4(b), 7 Marks]*

6. "Information gain is based on information theory" — Comment. *[Feb/March 2024, Q1(c), 2 Marks]*

7. Can a contrast be observed between classification and prediction? *[Feb/March 2024, Q1(g), 2 Marks]*

8. Demonstrate the use of the Decision Tree Induction Algorithm using a given set of class-labelled training tuples from a customer database (attributes: age, income, student, credit rating; class: buys_computer). *[Feb/March 2024, Q4(a), 14 Marks]*

9. Examine the various metrics for evaluating a classifier. *[Feb/March 2024, Q4(b), 7 Marks]*

B. Additional Practice Questions

10. Explain the general two-step process of building a classification model: learning (training) and classification (testing).

11. Explain the decision tree induction algorithm (ID3/C4.5 style) and describe the attribute selection measures: information gain, gain ratio, and Gini index.

12. What is tree pruning? Distinguish between pre-pruning and post-pruning approaches to avoid overfitting.

13. State Bayes' theorem. Explain how it forms the basis of Bayesian classification.

14. Derive the Naive Bayes classification formula assuming class-conditional independence, and classify a sample tuple using a worked numeric example.

15. Explain rule-based classification. Describe how classification rules can be extracted from a decision tree.

16. Explain the sequential covering algorithm for direct rule extraction, and describe rule pruning.

17. What is a confusion matrix? Define and explain accuracy, precision, recall, F1-score, sensitivity and specificity for a classifier.

18. Explain the holdout method, k-fold cross-validation, and bootstrap method for evaluating classifier accuracy.

19. What is overfitting in classification, and what techniques can be used to avoid it?

20. Explain ensemble methods for improving classification accuracy: bagging, boosting (AdaBoost) and random forests.

21. Compare bagging and boosting techniques for ensemble classification, highlighting differences in how training samples are drawn, how base classifiers are combined, and their effect on bias and variance.

22. Explain the backpropagation algorithm used to train a multilayer feed-forward neural network.

23. Distinguish between lazy learners (e.g., k-NN) and eager learners (e.g., decision trees), giving the advantages and disadvantages of each.

24. Explain the ROC curve and AUC as measures of classifier performance.

UNIT IV — Cluster Analysis: Concepts and Methods

Syllabus topics: *Cluster analysis; partitioning methods; hierarchical methods (agglomerative vs divisive, distance measures); BIRCH; density-based methods (DBSCAN); grid-based methods (STING); evaluation of clustering.*

A. Previous Exam Questions

1. What is the purpose of evaluating clustering results? List at least two commonly used evaluation metrics to assess the quality of clustering outcomes. *[Sept/Oct 2023, Q1(e), 2 Marks]*

2. Briefly describe and give examples of each of the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. *[Sept/Oct 2023, Q5, 14 Marks]*

3. Write comparisons between the AGNES and DIANA algorithms for clustering. *[Aug/Sept 2024, Q1(d), 2 Marks]*

4. Suppose the data mining task is to cluster points (x, y) into three clusters: A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9), using Euclidean distance, with A1, B1 and C1 as initial cluster centres. Use the k-means algorithm to show (i) the three cluster centres after the first round of execution, and (ii) the final three clusters. *[Aug/Sept 2024, Q5(a), 7 Marks]*

5. Explain grid-based clustering methods. *[Aug/Sept 2024, Q5(b), 7 Marks]*

6. What are the key issues in hierarchical clustering? Explain. *[Aug/Sept 2024, Q7(b), 7 Marks]*

7. Enumerate the typical requirements of clustering in data mining. *[Feb/March 2024, Q1(d), 2 Marks]*

8. Compare the hierarchical clustering algorithms AGNES and DIANA. *[Feb/March 2024, Q5(a), 7 Marks]*

9. Inspect how we can assess the clustering tendency of a data set. *[Feb/March 2024, Q5(b), 7 Marks]*

B. Additional Practice Questions

10. Define cluster analysis. What are the major applications of clustering in real-world data mining?

11. Explain the general categorization of clustering methods: partitioning, hierarchical, density-based, grid-based, and model-based.

12. Explain the k-means partitioning algorithm step by step, and discuss its strengths and limitations.

13. Explain the k-medoids (PAM) algorithm and how it differs from k-means, particularly in handling outliers.

14. Distinguish between agglomerative and divisive hierarchical clustering approaches, with a diagram (dendrogram).

15. Explain the different inter-cluster distance measures used in hierarchical clustering: single linkage, complete linkage, average linkage, and centroid method.

16. Explain the BIRCH algorithm, including the concept of a Clustering Feature (CF) and CF-tree.

17. Explain the DBSCAN algorithm. Define core point, border point, and noise point, and explain the concepts of density-reachability and density-connectivity.

18. What are the advantages of density-based clustering methods (like DBSCAN) over partitioning methods for datasets with arbitrarily shaped clusters?

19. Explain the STING (Statistical Information Grid) grid-based clustering algorithm.

20. How is the appropriate number of clusters k determined in k-means clustering? Discuss the elbow method.

21. Explain the silhouette coefficient and other internal/external measures used to evaluate the quality of clustering results.

22. What is outlier analysis? Briefly explain statistical, distance-based, and density-based approaches to outlier detection.

UNIT V — Data Mining Trends and Research Frontiers

Syllabus topics: *Mining complex data types; other methodologies of data mining; data mining applications; data mining and society; data mining trends.*

A. Previous Exam Questions

1. What are some challenges and opportunities associated with mining complex data types, such as textual data or multimedia data? *[Sept/Oct 2023, Q1(f), 2 Marks]*

2. What is a recommender system? In what ways does it differ from a customer or product-based clustering system? How does it differ from a typical classification or predictive modelling system? Outline one method of collaborative filtering. Discuss why it works and what its limitations are in practice. *[Sept/Oct 2023, Q6, 14 Marks]*

3. What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuples)? *[Sept/Oct 2023, Q7(b), 7 Marks]*

4. Give examples of sequence data. *[Aug/Sept 2024, Q1(e), 2 Marks]*

5. Differentiate frequent subsequence and frequent substructure. *[Aug/Sept 2024, Q1(f), 2 Marks]*

6. What is sequential pattern mining? Explain the applications of sequential pattern mining with examples. *[Aug/Sept 2024, Q3(b), 7 Marks]*

7. List and explain the trends of data mining. *[Aug/Sept 2024, Q6(a), 7 Marks]*

8. How is data mining useful for society? Explain. *[Aug/Sept 2024, Q6(b), 7 Marks]*

9. Does ChatGPT (an AI/LLM tool) provide quick and accurate answers to data mining questions? Discuss. *[Feb/March 2024, Q1(e), 2 Marks]*

10. Outline the different complex data types which can be mined. *[Feb/March 2024, Q6(a), 7 Marks]*

11. Write a note on Data Mining and Society. *[Feb/March 2024, Q6(b), 7 Marks]*

B. Additional Practice Questions

12. Explain the key techniques used in mining text data, including text categorization and topic modelling.

13. What is multimedia data mining? Discuss the challenges in mining image, audio, and video data.

14. Explain spatial data mining and spatiotemporal data mining with examples of applications.

15. Explain graph mining and network data mining. Discuss frequent subgraph mining.

16. What is web mining? Explain its three categories: web content mining, web structure mining, and web usage mining.

17. What is data stream mining? Discuss the key challenges in mining continuously arriving, high-speed data streams.

18. Explain time-series data analysis and the key data mining tasks associated with it (trend analysis, similarity search, forecasting).

19. Discuss privacy-preserving data mining and why it is important.

20. Explain, with examples, the main application areas of data mining (e.g., finance, retail, healthcare, telecommunications, fraud detection, recommender systems).

21. Discuss the ethical and societal issues raised by widespread use of data mining, including privacy and bias/fairness concerns.

22. Discuss recent research trends in data mining, including its integration with machine learning, deep learning, and large language models.

23. Compare sequential pattern mining and frequent itemset (association rule) mining, highlighting how order/time is handled differently.