CS-203 (New): Data Mining and Data Warehousing

Question Bank

Chapter I – Introduction to Data Mining

(1) Explain basic data mining tasks with an example.

(2) Give details on data mining versus knowledge discovery in databases.

(3) Discuss data mining issues.

(4) What is data mining metrics?

(5) What are social implications of data mining?

(6) Give an overview of Applications of data mining.

Chapter II – Introduction to Data Warehousing

(1) Explain: data warehousing, OLAP.

(2) What do you mean by machine learning?

(3) Explain pattern matching.

(4) Suppose that a data ware house for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg_grade. When at the lower conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

(a) Draw a snowflake schema diagram for the data warehouse.

(b) Starting with the base cuboids (student, course, semester, instructor), what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student.

(5) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the

fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(a) Draw a star schema diagram for the data warehouse.

(b) Starting with the base cuboids (date, spectator, location, game), what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM_Place in 2004.

(6) A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind you answer.

(7) Discuss issues to consider during data integration.

(8) Write a note on the Architecture of Data Warehousing.

(9) Explain the need of Data Processing.

(10) Explain the terms related to Data Processing-

(a) Data Cleaning (b) Data Integration & Transformation

(c) Data Reduction

<u>Chapter III – Data Mining Techniques</u>

(1) Explain frequent item-set algorithm.

(2) Give an example for Apriori with transactions and explain Apriori-gen-algorithm.

(3) Explain sampling algorithm with an example.

(4) The Apriori algorithm uses prior knowledge of subset support properties.

(a) Prove that all nonempty subsets of a frequent item-set must also be frequent.

(b) Prove that the support of any nonempty subset s` of item-set s must be at least as great as the support of s.

(c) Given frequent item-set l and subset s of l, prove that the confidence of the rule "s` → (l-s`)" cannot be more than the confidence of "s → (l-s)", where s` is a subset of s

(d) A partitioning variation of Apriori subdivides the transactions of a database D into n non-overlapping partitions. Prove that any item-set that is frequent in D must be frequent in at least one partition of D.

(5) Define a frequent set. Define an association rule.

(6) Define a FP-tree. Discuss the method of computing a FP-tree.

Chapter IV – Classification and Prediction

(1) Briefly outline the major steps of decision tree classification.

(2) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning.

(3) Give a decision tree, you have the option of

(a) converting the decision tree to rules and the pruning the resulting rules. Or

(b) pruning the decision tree and them converting the pruned tree to rules.

What advantage does (a) have over (b)?

(4) Why is naïve Bayesian classification called "naïve"? Briefly outline the major ideas of naïve Bayesian classification.

(5) Explain with an example Bayesian classification.

(6) Explain decision tree-based algorithm.

(7) What do you mean by CART?

(8) Give an example for classification using prediction.

(9) Write a note on linear regression/non-linear regression.

(10) Some non-linear regression models can be converted to linear models by applying transformations to the predictor variables. Show how the non-linear regression equation $y = \alpha X^{B}$ can be converted to a linear regression equation solvable by the method of least squares.

Chapter V – Accuracy Measures

(1) Explain the following accuracy measures-

(a) Precision (b) F-measure (c) Confusion matrix

(d) Cross-validation (e) Bootstrap

Chapter VI – Software of Data mining and Applications of data mining

(1) Write a note on Applications of data mining.

Chapter VII – Clustering

(1) Give examples for different clustering attributes.

(2) Give an example for hierarchical algorithms.

(3) Give an example for K-means clustering.

(4) What is clustering? What are the different clustering techniques?

(5) Write a note on Hierarchical clustering.

Chapter VIII – Brief overview of advanced techniques

(1) What are the different types of web mining?

(2) How is web usage mining different from web structure mining and web content mining?

(3) How is text mining related to web mining? What are the techniques of text mining?

(4) How do you extract structures from unstructured text data? What features are extracted in this process?

(5) Which frequent item-set mining is suitable for text mining?

(6) What are the differences between mining techniques of structured data, semi-structured data and unstructured data?

(7) Write a note on Text mining.

(8) Explain Web mining.