Important Question & Topics

> * Represents repeated questions more than one time no. of stars = no. of time repeated question.

Pg-1

> (I, II, III ----) Represents module no.

> All questions are PYQs.

> Solving questions according to module will be helpful

Subject : Data Warehouse & Mining

I □ Consider the following for a Hypermarket chain: Product, store, Time & Promotion.
- W.r.t. this business scenario, ans. the following questions.
- Clearly state any reasonable assumptions you make.

> Design a star schema. Whether the star schema is can be converted to snowflake schema ?

> Justify your answer & draw snowflake schema for datawarehouse (clearly mention fact table(s). Dimension table(s), their attributes & measure(s).

□ Define linear, non-linear & multiple regressions.
- Plan & regression model for ~~dis~~ disease development w.r.t. change in whether parameters.

I □ What is meant by metadata in context of Data warehouse ? Explain Diff. types of metadata stored in Data Warehouse. Illustrate suitable example. **

II □ Describe the various functionalities of Data mining as a step in the process of knowledge & Discovery. **

I □ In what way ETL cycle can be used in typical datawarehouse, explain with suitable instance.

- Explain ETL of datawarehousing in detail.

2023.11.04 22:54

I Every data structure in the data warehouse contains the time element. Why? **

Differentiate :

I i) OLTP V/s OLAP ****

I ii) Data Warehouse v/s Data Mart.

III Why naive Bayesian classification is called "naive"? Briefly outline the major ideas of naive Bayesian classification. **

Short note :

II i) Application of Data mining to Financial Analysis.

I ii) Factless Fact Table

I iii) Indexing OLAP data.

iv) Data Quality

III v) Decision Tree Based Classification Approach.

II Explain Data mining as a step in KDD. Give architecture of typical data mining system.

II Discuss different steps involved in Data-Preprocessing ***

IV Find clusters using k-means algorithm if we have several objects (4 types of medicine) & each object have two attributes or features as shown in the table below. The goal is to group these objects into k=2 group of medicine based on the two features (pH & weight index).

Object	Attribute 1(X) Weight index	Attribute 2 (Y) pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

2023.11.04 22:55

I The college wants to record the marks for the courses completed by students using the dimension: a) Course, b) Student, c) Time & a measure d) Aggregate marks. **

- Create a cube & describe the following OLAP operations

i) Rollup ii) Drill down iii) Slice iv) Dice

v) Pivot.

V A database has five transactions. Let minimum support = 30% & minimum Confidence = 70%

i) Find all frequent patterns using Apriori Algo.

ii) List strong association rules.

Transaction - Id	Items
A	1, 3, 4, 6
B	2, 3, 5, 7
C	1, 2, 3, 5, 8
D	2, 5, 9, 10
E	1, 4

Short note

I a) Data warehouse design strategies.

II b) Application of Data mining.

I c) Role of metadata

IV d) Hierarchical clustering

I e) Fact Constellation

II f) Data Visualization **

V g) Multidimensional & multilevel Association mining **

I h) ETL Process

V i) Discuss Association Rule mining & Apriori Algo.

2023.11.04 22:56

I □ Differentiate top-down & bottom-up approches for building data warehouse. Discuss the merits & ~~Demerits~~ Limitations of each approach **

□ Discuss:

II i) Architecture of a typical data mining system

II ii) Application & major issues in Data mining

I □ Consider a data warehouse for a hospital where there are three dimension *

a) Doctor b) Patient c) Time

- Consider two measures i) Count ii) Charge where charge is the fee that the doctor charges a patient for a visit.
- for the above example create a cube & Illustrate the following OLAP operations.

1) Rollup 2) Drill down 3) Slice 4) Dice

5) Pivot.

(Similar question with diff data is asked.)

III □ · Define : Classification

· Discuss issues in classification

I □ Discuss various OLAP Models & their ~~Arc~~ Architecture.

VI □ What is web structure mining? List the approaches used to structure the web pages to improve on the effectiveness of search engines & crawlers. Explain page rank technique in detail *

II □ Explain major issue in Data mining.

VI □ Explain web usage mining in detail

III □ What are the various methods for estimating classifiers accuracy?

III □ What are the various issues regarding Classification and Prediction?

2023.11.04 22:56

IV □ Use k-mean algorithm to create 3-cluster for the given set of values : {2, 3, 6, 8, 9, 12, 15, 18, 22} ***

I □ What are basic building blocks of Data Warehouse?

□ Differentiates

IV i) Agglomerative & Divisive clustering method

I ii) Star schema & Snowflake schema

I □ Design star schema for company company sales with three dimensions such as Location, Item & Time.

I □ • Differentiate b/w top-down & bottom-up approaches for building data warehouse.
• Discuss merits & limitations
• Explain the practical approach for designing a data warehouse. **

II □ Discuss issues in Data mining.

III □ • Explain Decision Tree based Classification Approach with example.
• Discuss metrics for evaluating Classifier Performance

VI □ What is web mining? Explain web structure mining & web usage mining in detail. **

II □ In real-world data, tuples with missing values for some attributes are common occurence.
• Describe various methods for handling this problem. **

V □ Explain market basket analysis with an example.

IV □ Describe k medoids algorithm.

II □ Discuss different types of attributes.

VI □ Elaborate on the taxonomy of web mining.

2023.11.04 22:57

- Describe the steps involved in the FP growth Algorithm, Compare & Contrast its steps with Apriori Algorithm.

- Explain how the accuracy of a classifier can be measured.
  - How bagging strategy helps improving the classifier accuracy?

- Solve

- Apply Apriori algorithm to find all frequent item sets from the following table:

  min support count = 2
  min confidence = 70%

TID	Item
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	1, 2, 3, 5

- A data base has five transactions. Let min confidence = 60%. min count support count = 2 &.
  - Find all frequent item sets using Apriori Algo.
  - List strong association rule

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5
500	1, 3, 5

2023.11.04 22:57

- Consider the data given below. Create adjacency matrix. Apply complete link algorithm to cluster the given data set & draw the dendogram.

	A	B	C	D	E
A	0	2	6	10	9
B	2	0	3	9	8
C	6	3	0	7	5
D	10	9	7	0	4
E	9	8	5	4	0

- Important Topics (Based on Past 5 to 6 year previous question papers)

Module - 1] High weightage
Module - 2]

Module - 3
- Decision Tree Induction
- Naive Bayesian Classification
- Evaluating Accuracy of a Classifier.

Module - 4
- K-means, K-medoids
- Hierarchical method (Agglomerative)
  ↳ Dendogram

Module 5
- Market Basket Analysis
- Frequent Patter Mining
- Apriori Algorithm
- Mining multilevel Association Rule & Mining multidimensional Association Rules.

2023.11.04 22:58

Module - 6

- Web Structure Mining - Page Rank, Web Usage mining
- Web Content Mining - &

2023.11.04 22:58