DWM

1. What is a data warehouse and why is a data warehouse used for analytics?
2. What is etl process why is it used
3. What are types of olap
4. Difference between data lake and data warehouse and data marts
5. What is difference between ROLAP and MOLAP
6. Difference between OLTP and OLAP
7. Example based question - *check q 27*
8. What is web mining and the three types of web mining

Web mining is the process of using data mining techniques to automatically discover and extract information from web documents and services. There are three main types of web mining:

Web content mining: This involves extracting useful information and knowledge from web documents, such as HTML pages and PDF files.

Web structure mining: This involves analyzing the hyperlinks and other structural elements of the web to discover patterns and relationships.

Web usage mining: This involves analyzing the web usage data of visitors to a website, such as the pages they visit, the time they spend on each page, and the actions they take, in order to understand their behavior and preferences.

Web mining can be used for a variety of purposes, such as improving search engine results, personalizing web content, and detecting fraud or malicious activity.

1. Difference between web structure and web content

Web structure refers to the way that web pages are linked to each other and organized into a hierarchical structure. This includes the hyperlinks that connect one page to another, as well as the overall organization of the web into directories, categories, and other structures. Web structure mining involves analyzing the structure of the web to discover patterns and relationships.

Web content, on the other hand, refers to the actual content of web pages, such as the text, images, and other media that make up a web page. Web content mining involves extracting useful information and knowledge from this content.

In general, web structure is concerned with the organization and connectivity of the web, while web content is concerned with the information and knowledge contained within web pages.

1. How to choose seed points for K means

Random Selection,Centroid/Mean,Mode,Mediode,Dbscan,Kmeans++(probabilistic approach),Manually select(done only if you have knowledge)

1. What is “append” in the loading process of ETL?? Update ig
2. How DBSCAN is different from traditional clustering

It is robust to outliers , clusters are created by dense regions and high density is region is separated from low density region, can detect noise and no need to specify number of clusters like in k-means and can create cluster in large spatial data

1. Sequential pattern mining|

Sequential pattern mining is a data mining technique that involves discovering patterns in sequential data. Sequential data is a type of data that is ordered in time, such as a sequence of events or a sequence of transactions.

Sequential pattern mining involves identifying patterns in the data that occur in a specific order, such as the sequence of items that are purchased in a transaction or the sequence of events that occur in a log file. These patterns can provide valuable insights into the data and help to uncover trends and relationships that may not be immediately apparent.

For example, sequential pattern mining can be used to identify common sequences of events in a log file, such as the sequence of pages that are visited on a website, or the sequence of actions that are performed in a mobile app. This information can be used to improve the user experience, optimize the performance of the system, and detect anomalies or suspicious activity.

Overall, sequential pattern mining is a powerful tool for analyzing sequential data and uncovering valuable insights that can be used to improve a wide range of applications.

1. What is spatial data –

used to store geographical data like coordinates, latitude or longitude

Spatial data is a type of data that describes the location and shape of objects in a spatial (or geographical) context. It is often used to represent real-world objects, such as roads, buildings, and land masses, and their relationships to each other.

1. What is more dangerous, false positive or false negative? (False Negative)|`

Because it predicts cancerous patient as non cancerous which has higher cost than false positive which predicts non cancerous patient as cancerous.

1. What is the nature of outcomes for classifications and precision models?

Discrete or continuous

1. What is the correlation coefficient?

Chi sq

the correlation coefficient is a useful tool for understanding the relationship between two variables and for evaluating the strength of that relationship.

The correlation coefficient is a statistical measure that indicates the strength and direction of a linear relationship between two variables. It is a value between -1 and 1, where -1 indicates a perfect negative linear relationship (i.e. as one variable increases, the other decreases), 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship (i.e. as one variable increases, the other also increases).

The chi-square test is a statistical test that is used to determine whether there is a significant difference between the observed frequency of a particular event and the expected frequency of that event. It is commonly used in hypothesis testing to evaluate the relationship between different variables, and to determine whether those variables are independent of each other.Chi square use to determine goodness of model too.

1. What is a confusion matrix?|

matrix used for evaluating the performance of a classification model TP TN FP FN

Tool for analysing how well your classifier can recognize tuples of different classes

1. What is conditional independency?explain with an example.|
2. Diff btw classification and prediction|
3. Way to calc number of clusters
4. Selecting attribute subset used to reduce data
5. What is a hypercube **if the number of dimensions is greater than 3**.
6. What are drawbacks of k means
7. What are attributes
8. Frequent Itemset
9. Example vala Question - RIP (for eg create info packet for university for payroll analysis)
10. How to overcome drawback of k means
11. Error Metrics
12. What is a cross validation dataset?
13. What are factless tables? Explain with example.
14. Classification vs clustering
15. Star schema vs snowflake
16. Top down vs bottom up
17. Slice vs Dice (OLAP)
18. Cardinality of Star Scheme one to many
19. What is data visualisation?
20. What is ensemble models and explain bagging and boosting
21. Which is sequential and parallel in bagging and boosting
22. What are extrinsic and intrinsic methods in clustering evaluation
23. What is overfitting
24. What is tree pruning
25. Explain ROC curve
26. Voting method
27. Quantile Quantile plots
28. What is a box plot?
29. What is regression? Types of regression.
30. WHAT IS CLARANS😭😭
31. What normal form are dimension tables in, in snowflake schema? **3NF**
32. Q-Q PLOT - Determine whether two samples are from the same population.
33. Q-PLOT
34. What is entropy - represents the level of randomness in the data
35. What is boxplot - graph that gives you a good indication of how the values in the data are spread out.
36. What are different ways to calculate distance?Hamming Distance. Euclidean Distance.Manhattan Distance.Minkowski Distance. Which is used in K medoid? Euclidean
37. How to improve efficiency of Apriori- hashing, sampling, transaction reduction, partitioning
38. What is imbalanced data and how to handle it
39. What NF is snowflake schema 3nf
40. Types of crawlers? focused , incremental, distributed, parallel. Difference between Traditional Crawler and Focused Crawler.|
41. Page Rank comes under what type of web mining? Structure mining
42. What is pruning? Deleting of a child node from the branch
43. Drawbacks of Apriori cannot be used for large volumes of data since the number of scan increases.
44. Hierarchical clustering types and its termination condition
45. Adaboost|
46. Boosting methods - **adaboost & gradient boosting**|
47. Slowly changing dimensions = **a dimension that stores and manages both current and historical data over time in a data warehouse**.
48. Explain data warehouse
49. What is dendogram
50. Strategic information | Strategic information is required for an enterprise to decide the business strategies and establish the goals for the business
51. What is information package diagram (ipd)| **defines the relationships between subject matter and key performance measures**.
52. What is semi-additive attribute - **measures that have a different way of aggregation over time**
53. What is apriori and list methods to improve it - is an algorithm to generate association rules. Hash based technique, transaction reduction, sampling, partitioning
54. What is pre-pruning
55. What is true positive, explain with example
56. Explain precision, recall, specificity. Which is useful| Precision – how many of the positively classified were relevant. sensitivity/recall – how good a test is at detecting the positives.
57. Slice and dice operations in OLAP
58. What is an iceberg query? Queries that have group by and having clauses
59. What is data reduction? Reducing the data by decreasing its quantity but maintaining the quality.
60. Explain different techniques of data reduction in brief - data cube aggregation, dimension reduction : stepwise forward, step wise backward, data compression, discretization and concept hierarchies.
61. What is the difference between subject-oriented and application-oriented data warehouses?**Same difference b/w database and data warehouse**
62. Explain the need of a data warehouse in brief.
63. Explain what you know about clustering.
64. What other parameter is used to find frequent itemsets other than support and confidence. (Lift)
65. What is junk dimension table
66. Features of data warehouse
67. Can more than one concept from concept hierarchy exist in a single fact table, give example
68. Data mining tasks = Prediction, classification, association, clustering
69. Techniques to improve efficiency of apriori= hash based technique, sampling, reduction, partitioning
70. What is confidence and support
71. Spatial Data
72. How can you overcome the disadv of K-Means? Distance measure used in it
73. Tree Pruning
74. Business intelligence vs data mining
75. Information delivery component
76. Types of report generated by info delivery
77. Granularity
78. Steps in preprocessing
79. HITs hypertext induced topic search
80. Explain multilevel association mining in detail
81. Define support and its role in Apriori
82. Is confidence important to find frequent itemset?
83. What is the use of confidence?