|  |  |  |  |
| --- | --- | --- | --- |
| serial | Machine learning answers | Sql answers | Statistics answers |
| 1 | B | A,d | a |
| 2 | d | A,b,c | a |
| 3 | d | b | b |
| 4 | a | b | d |
| 5 | b | a | d |
| 6 | d | c | b |
| 7 | a | b | b |
| 8 | b | b | a |
| 9 | a | d | c |
| 10 | a | a |  |
| 11 | c |  |  |
| 12 | a |  |  |
| 13 |  |  |  |

MACHINE LEARNING

1 answer: steps to calculate cluster analysis

1: getting our data set, preferably scaled data for easy calculation.

2: Calculate the distance from each data point to the centre of a cluster.

3: calculate mean average of each data set.

4: calculate the distance between revised mean.

5: Graph and Summarize the Clusters

2 answer: cluster quality is measured by

- SSE: sum of the square error from the items of each cluster.

-Inter cluster distance: sum of the square distance between each cluster centroid.

-Intra cluster distance for each cluster: sum of the square distance from the items of each cluster to its centroid.

- Maximum Radius: largest distance from an instance to its cluster centroid.

-Average Radius: sum of the largest distance from an instance to its cluster centroid divided by the number of clusters.

3 answer:   
 Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another.

types

-Hierarchical clustering,

-Partition clustering,

-Exclusive Clustering,

- Overlapping Clustering,

-Fuzzy clustering,

-Complete clustering.

SQL WORKSHEET

1 answer:

Data warehousing is the electronic storage of a large amount of information by a business or organization. A data warehouse is designed to run query and analysis on historical data derived from transactional sources for business intelligence and data mining purposes.

2 answer:

Difference between OLAP and OLTP

-Online Analytical Processing (OLAP) is a category of software tools that analyse data stored in a database whereas Online transaction processing (OLTP) supports transaction-oriented applications in a 3-tier architecture.

-OLAP creates a single platform for all type of business analysis needs which includes planning, budgeting, forecasting, and analysis while OLTP is useful to administer day to day transactions of an organization.

-OLAP is characterized by a large volume of data while OLTP is characterized by large numbers of short online transactions.

-In OLAP, data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database whereas OLTP uses traditional DBMS

3 answer:

characteristics of data warehouse

- Some data is denormalized for simplification and to improve performance.

-Large amounts of historical data are used.

-Queries often retrieve large amounts of data.

-Both planned and ad hoc queries are common.

-The data load is controlled.

4 answer:

Star schema

In data warehousing and business intelligence, a star schema is the simplest form of a dimensional model, in which data is organized into facts and dimensions. A fact is an event that is counted or measured, such as a sale or login. A dimension contains reference information about the fact, such as date, product, or customer. A star schema is diagrammed by surrounding each fact with its associated dimensions. The resulting diagram resembles a star.

5 answer:

SETL (set theory language is a very high-level programming language based on the mathematical theory of sets.

STATISTICS

Answer 1)

A normal distributionis a common probability distribution. It has a shape often referred to as a "bell curve."

Many everyday data sets typically follow a normal distribution: for example, the heights of adult humans, the scores on a test given to a large class, errors in measurements.

The normal distribution is always symmetrical about the mean.

answer2)

Missing data occurs when loss of data or errors, Missing data is handled by imputing nan values to categorical and mean values to numerical, sometimes we use mode for categorical depends upon the data,

Answer 3):

A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

Answer 4): mean imputation

-Bad practice in general

-If just estimating means: mean imputation preserves the mean of the observed data

-Leads to an underestimate of the standard deviation

-Distorts relationships between variables by “pulling” estimates of the correlation toward zero

Answer 5):

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression

Answer 6):

Two branches, *descriptive statistics* and *inferential statistics*, comprise the field of statistics.

#### Descriptive Statistics:

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

#### Inferential Statistics

The branch of statistics that analyses sample data to draw conclusions about a population.