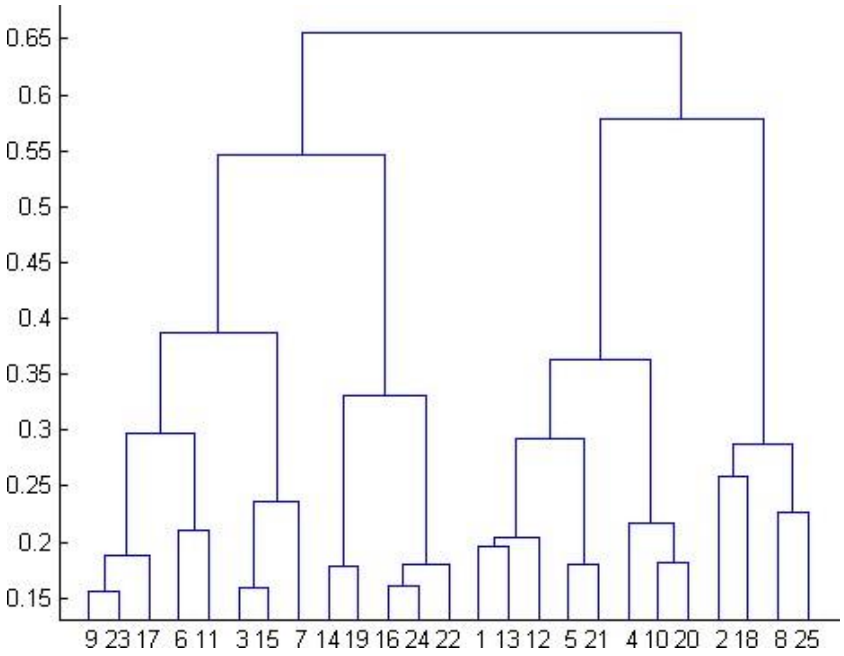
MACHINE LEARNING SOLUTION

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



a) 2

b) 4

c) 6

d) 8

ANSWER : (B) 4

2. In which of the following cases will KMeans clustering fail to give good results?
1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with nonconvex shapes

Options:

a) 1 and 2

b) 2 and 3

c) 2 and 4

d) 1, 2 and 4

ANSWER : (D) 1,2 AND 4

3. THE MOST IMPORTANT PART OF _____ IS SELECTING THE VARIABLES ON WHICH CLUSTERING IS BASED.
A) INTERPRETING AND PROFILING CLUSTERS
B**)** SELECTING A CLUSTERING PROCEDURE
C) ASSESSING THE VALIDITY OF CLUSTERING.
D) FORMULATING THE CLUSTERING PROBLEM
ANSWER : (D) FORMULATING THE CLUSTERING PROBLEM

4. THE MOST COMMONLY USED MEASURE OF SIMILARITY IS THE _____ OR ITS SQUARE.
A) EUCLIDEAN DISTANCE B) CITYBLOCK DISTANCE
C) CHEBYSHEV'S DISTANCE
D) MANHATTAN DISTANCE
ANSWER : (A) EUCLIDEAN DISTANCE

5._____ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

a) Nonhierarchical clustering

b) Divisive clustering

c) Agglomerative clustering

d) Kmeans clustering

ANSWER : (b) Divisive clustering

6. Which of the following is required by Kmeans clustering?

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

d) All answers are correct

ANSWER : (D) ALL ANSWERS ARE CORRECT

7. The goal of clustering is to

a) Divide the data points into groups

b) Classify the data point into different classes

c) Predict the output values of input data points

d) All of the above

ANSWER : (a) Divide the data points into groups

8. Clustering is a

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

ANSWER : (B) UNSUPERVISED LEARNING

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima? a) K Means clustering

b) Hierarchical clustering

c) Diverse clustering

d) All of the above

ANSWER : (D) ALL OF THE ABOVE

10. Which version of the clustering algorithm is most sensitive to outliers?

a) Kmeans clustering algorithm

b) Kmodes clustering algorithm

c) Kmedians clustering algorithm

d) None

ANSWER : (A) KMEANS CLUSTERING ALGORITHM

11. Which of the following is a bad characteristic of a dataset for clustering analysis

a) Data points with outliers

b) Data points with different densities

c) Data points with nonconvex shapes

d) All of the above

ANSWER : (d) All of the above

12. For clustering, we do not require

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

ANSWER : (a) Labeled data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated? ANS:

The hierarchical cluster analysis follows three basic steps:

i)calculated the distances
ii) linked the clusters
iii) choose a solution by selecting the right number of clusters

1.First, we have to select the variables upon which we base our clusters. In the dialog window we add the math, reading, and writing tests to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.
2.In the dialog box Statistics we can specify whether we want to output the proximity matrix (these are the distances calculated in the first step of the analysis) and the predicted cluster membership of the cases in our observations. Again, we leave all settings on default.
3.In the dialog box Plots we should add the Dendrogram. The Dendrogram will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.
4.The dialog box Method… allows us to specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.
5.For interval data, the most common is Square Euclidian Distance. It is based on the Euclidian Distance between two observations, which is the square root of the sum of squared distances. Since the Euclidian Distance is squared, it increases the importance of large distances, while weakening the importance of small distances.
6.If we have ordinal data (counts) we can select between ChiSquare or a standardized ChiSquare called PiSquare. For binary data, the Squared
Euclidean Distance is commonly used.
We have to choose the Cluster Method. Typically, choices are between groups linkage, nearest neighbor, furthest neighbor, and Ward's method. Single linkage works best with long chains of clusters, while complete linkage works best with dense blobs of clusters. Between groups linkage works with both cluster types. It is

recommended is to use single linkage first. Although single linkage tends to create chains of clusters, it helps in identifying outliers. After excluding these outliers, we can move onto Ward's method.

Ward's method uses the F value (like in ANOVA) to maximize the significance of differences between clusters.

A last consideration is standardization. If the variables have different scales and means we might want to standardize either to Z scores or by centering the scale. We can also transform the values to absolute values if we have a data set where this might be appropriate.

14. How is cluster quality measured?

ANSWER=

1.Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterised by minimal intra cluster distance and maximal inter cluster distance.

2.To measure the quality of clustering results, there are two kinds of validity indices: external indices and internal indices.

3.An external index is a measure of agreement between two partitions where the first partition is the a priori known clustering structure, and the second results from the clustering procedure.

3.Internal indices are used to measure the goodness of a clustering structure without external information.

4.For external indices, we evaluate the results of a clustering algorithm based on a known cluster structure of a data set (or cluster labels).

5.For internal indices, we evaluate the results using quantities and features inherent in the data set. The optimal number of clusters is usually determined based on an internal validity index.

6.Extrinsic Measures require ground truth labels. Examples are Adjusted Rand index, FowlkesMallows scores, Mutual information based scores, Homogeneity, Completeness and Vmeasure.

7.Intrinsic Measures does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, CalinskiHarabasz Index, DaviesBouldin Index etc.

15. What is cluster analysis and its types?

ANSWER=

1.Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

2.It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Types of cluster analysis:

->Hierarchical clustering: Also known as 'nesting clustering' as it also clusters to exist within bigger clusters to form a tree.

->Partition clustering: Its simply a division of the set of data objects into nonoverlapping clusters such that each objects is in exactly one subset.

->Exclusive Clustering: They assign each value to a single cluster.

->Overlapping Clustering: It is used to reflect the fact that an object can simultaneously belong to more than one group.

->Fuzzy clustering: Every objects belongs to every cluster with a membership weight that goes between 0:if it absolutely doesn't belong to cluster and 1:if it absolutely belongs to the cluster. Complete clustering: It perform a hierarchical clustering using a set of dissimilarities on 'n' objects that are being clustered. They tend to find compact clusters of an approaximately equal diagram

WORKSHEET SQL SOLUTION

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. Which of the following is/are DDL commands in SQL?

A) Create B) Update

C) Delete D) ALTER

ANSWER : (A) CREATE , (D) ALTER

2. Which of the following is/are DML commands in SQL? A) Update C) Select

A) Update B) Delete

C) Select D) Drop

ANSWER : (A) UPDATE , (B) DELETE

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

3. Full form of SQL is:

A) Strut querying language B) Structured Query Language

C) Simple Query Language D) None of them

ANSWER : (B) STRUCTURED QUERY LANGUAGE

4. Full form of DDL is:

A) Descriptive Designed Language B) Data Definition Language

C) Data Descriptive Language D) None of the above

ANSWER : (B) Data Definition Language

5. DML is:

A) Data Manipulation Language B) Data Management Language

C) Data Modeling Language D) None of them

ANSWER : (A) DATA MANIPULATION LANGUAGE

6. Which of the following statements can be used to create a table with column B int type and C float type?

A) Table A (B int, C float) B) Create A (b int, C float)

C) Create Table A (B int,C float) D) All of them

ANSWER : (C) CREATE TABLE A (B INT,C FLOAT)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?

A) Table A (D float) B) Alter Table A ADD COLUMN D float

C) Table A(B int, C float, D float) D) None of them

ANSWER : (B) ALTER TABLE A ADD COLUMN D FLOAT

8. Which of the following statements can be used to drop the column added in the above question?

A) Table A Drop D B) Alter Table A Drop Column D

C) Delete D from A D) None of them

ANSWER : (B) ALTER TABLE A DROP COLUMN D

9. Which of the following statements can be used to change the data type

(from float to int) of the column D of table A created in above questions?

A) Table A (D float int B) Alter Table A Alter Column D int

C) Alter Table A D float int D) Alter table A Column D float to int

ANSWER : (B) ALTER TABLE A ALTER COLUMN D INT

10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?

A) Alter Table A Add Constraint Primary Key B

B) Alter table (B primary key)

C) Alter Table A Add Primary key B

D) None of them

ANSWER : (C) ALTER TABLE A ADD PRIMARY KEY B

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What is datawarehouse?

ANSWER

1.A data warehouse is a large collection of business data used to help an organization make decisions.

2.One of the primary components in a SQL Server business intelligence (BI) solution is the data warehouse.

3.Indeed, the data warehouse is, in a sense, the glue that holds the system together. The warehouse acts as a central repository for heterogeneous data that is to be used for purposes of analysis and reporting.

3.A data warehouse consolidates, standardizes, and organizes data in order to support business decisions that are made through analysis and reporting.

4.The data might originate in RDBMSs such as SQL Server or Oracle, Excel spreadsheets, CSV files, directory services stores such as Active Directory, or other types of data stores, as is often the case in large enterprise networks.

5.The data warehouse must be able to store data from a variety of data sources in a way that lets tools such as SQL Server Analysis Services (SSAS) and SQL Server Reporting Services (SSRS) efficiently access the data. These tools are, in effect, indifferent to the original data sources and are concerned only with the reliability and viability of the data in the warehouse.

6.A data warehouse is sometimes considered to be a place for archiving data; however, that is not its purpose. Although historical data is stored in a data warehouse, only the historical range necessary to support analysis and reporting is retained there.

7.For example, if a business rule specifies that the warehouse must maintain two years worth of historical data, older data is offloaded to another system for archiving or is deleted, depending on the specified business requirements.

WHAT IS THE DIFFERENCE BETWEEN OLTP VS OLAP?

ANSWER

1.OLTP is an acronym for Online transaction processing. OLAP is an acronym for Online analytical processing.
2.OLTP is an online transactional system. It manages database modification. OLAP is an online analysis and data retrieving process.
3.OLTP is characterized by large numbers of short online transactions. OLAP is characterized by a large volume of data.
4.OLTP is an online database modifying system. OLAP is an online database query management system.
5.OLTP uses traditional DBMS. OLAP uses the data warehouse.
6.Tables in OLTP database are normalized. Tables in OLAP database are not normalized.
7.OLTP and its transactions are the sources of data. Different OLTP databases become the source of data for OLAP.

8.OLTP database must maintain data integrity constraint. OLAP database does not get frequently modified. Hence, data integrity is not an issue.
9.OLTP response time is in milliseconds. OLAP response time is in seconds to minutes.
10.The data in the OLTP database is always detailed and organized. The data in OLAP process might not be organized.
11.Queries in OLTP process are standardized and simple. OLAP consists Complex queries involving aggregations.
12.OLTP needs Complete backup of the data combined with incremental backups. OLAP only need a backup from time to time. Backup is not important compared to OLTP.
13.OLTP is used by Data critical users like clerk, DBA & Data Base professionals. OLAP is Used by Data knowledge users like workers, managers, and CEO.
14.OLTP is designed for real time business operations. OLAP is designed for analysis of business measures by category and attributes.
15.OLTP kind of Database users allows thousands of users. OLAP kind of Database allows only hundreds of users.
16.OLTP is designed to have fast response time, low data redundancy and is normalized. In OLAP a data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database.
17.An example of OLTP system is ATM center. Any Datawarehouse system is an OLAP system.

13. What are the various characteristics of datawarehouse?

ANSWER

Here is the list of some of the characteristics of data warehousing:

1.Subject oriented: A data warehouse is subject oriented, as it provides information a topic rather than the ongoing operations of organizations. Such issues may be inventory, promotion, storage, etc. Never does a data warehouse concentrate on the current processes. Instead, it emphasized modeling and analyzing decisionmaking data. It also provides a simple and succinct description of the particular subject by excluding details that would not be useful in helping the decision process.

2.Integrated: Integration in Data Warehouse means establishing a standard unit of measurement from the different databases for all the similar data. The data must also get stored in a simple and universally acceptable manner within the Data

Warehouse. Through combining data from various sources such as a mainframe, relational databases, flat files, etc., a data warehouse is created. It must also keep the naming conventions, format, and coding consistent. Such an application assists in robust data analysis. Consistency must be maintained in naming conventions, measurements of characteristics, specification of encoding, etc.

1.Timevariant: Compared to operating systems, the time horizon for the data warehouse is quite extensive. The data collected in a data warehouse is acknowledged over a given period and provides historical information. It contains a temporal element, either explicitly or implicitly. One such location in the record key system where Data Warehouse data shows time variation is. Each primary key contained with the DW should have an element of time either implicitly or explicitly. Just like the day, the month of the week, etc.

Nonvolatile: Also, the data warehouse is nonvolatile, meaning that prior data will not be erased when new data are entered into it. Data is readonly, only updated regularly. It also assists in analyzing historical data and in understanding what and when it happened. The transaction process, recovery, and competitiveness control mechanisms are not required. In the Data Warehouse environment, activities such as deleting, updating, and inserting that are performed in an operational application environment are omitted.

14. What is StarSchema??

ANSWER

1.Star schema is the fundamental schema among the data mart schema and it is simplest.
2.This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables.
3.The star schema is a necessary case of the snowflake schema. It is also efficient for handling basic queries.
4.It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star’s points.
5.In Star Schema, Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data.

6.Often, A Star Schema having multiple dimensions is termed as Centipede Schema. It is easy to handle a star schema which have dimensions of few attributes.
7.Advantages of Star Schema :o Simpler Queries: Join logic of star schema is quite cinch in compare to other join logic which are needed to fetch data from a transactional schema that is highly normalized.
Simplified Business Reporting Logic: In compared to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as asof reporting and periodoverperiod.
Feeding Cubes: Star schema is widely used by all OLAP systems to design OLAP cubes efficiently. In fact, major OLAP systems deliver a ROLAP mode of operation which can use a star schema as a source without designing a cube structure.
7.Disadvantages of Star Schema :o Data integrity is not enforced well since in a highly denormalized schema state.
Not flexible in terms if analytical needs as a normalized data model. OStar schemas don't reinforce manytomany relationships within business entities – at least not frequently.
Sales price, sale quantity, distant, speed, weight, and weight measurements are few examples of fact data in star schema.

15. What do you mean by SETL?

ANSWER

1.SETL (SET Language) is a very highlevel programming language based on the mathematical theory of sets.
2.It was originally developed by (Jack) Jacob T. Schwartz at the New York University (NYU) Courant Institute of Mathematical Sciences in the late 1960s.
3.SETL provides two basic aggregate data types: unordered sets, and latter also called tuples).
4.The elements of sets and tuples can be of any arbitrary type, including sets and tuples themselves. Maps are provided as sets of pairs (i.e., tuples of length 2) and can have arbitrary domain and range types.
5.Primitive operations in SETL include set membership, union, intersection, and power set construction, among others.

6.SETL provides quantified boolean expressions constructed using the universal and existential quantifiers of firstorder predicate logic.
7.SETL provides several iterators to produce a variety of loops over aggregate data structures.
8.SETL is an interpreted language with a syntax that is resembles C and in many

STATISTICS WORKSHEET SOLUTION

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

ANSWER : (A) TRUE

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

ANSWER : (A) CENTRAL LIMIT THEOREM

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count

data

c) Modeling contingency tables

d) All of the mentioned

ANSWER : (B) MODELING BOUNDED COUNT DATA

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is calledchi-squared distribution d) All of the mentioned

ANSWER : (D) ALL OF THE MENTIONED

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

ANSWER : (C) POISSON

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

ANSWER : (B) FALSE

7. 1. Which of the following testing is concerned with making decisions using data?
a)Probbility

b) Hypothesis

c) Causal

d) None of the mentioned

ANSWER : (B) HYPOTHESIS

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

ANSWER : (A) 0

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

ANSWER : (C) OUTLIERS CANNOT CONFORM TO THE REGRESSION RELATIONSHIP

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

ANSWER

1.Normal distribution is also known as Gaussian Distribution.

2.Normal distribution is a probability distribution that is symmetric about the mean i.e showing that data near the mean are more frequent in occurrence than data far from the mean.
3.In graph form, normal distribution will appear as a bell curve.
4.The normal distribution is the most common type of distribution assumed in statistical analyses. However Real life data rarely follows a perfect normal distribution.
5.The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/ one standard deviation of the mean, 95% are within +/ two standard deviations, and 99.7% are within + three standard deviations.
6.The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution.
7.The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail. positive skewness implies that the right tail of the distribution is longer than the left.
8.The kurtosis statistic measures the thickness of the tail ends of a distribution in relation to the tails of the normal distribution.
9.The normal distribution model is motivated by the Central Limit Theorem.

11. How do you handle missing data? What imputation techniques do you recommend?

ANSWER

1. I will interpret that why there are missing values in our dataset. For instance some people hesitate to put down the information for eample Men hesitate to put down the salary and Women hesitate to put down the age, Survey informations are not that valid, People may have died.
2.Then I will interpret what are the different types of missing data: o Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observe.
Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variable.

Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

3.In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model.

4.Very basic type of imputation technique is Computing the overall mean, it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. It is very fast, but has clear disadvantages. One disadvantage is that mean imputation reduces variance in the dataset.

5.There are other machine learning techniques like XGBoost and Random Forest for data imputation but personally my favorite imputation technique is KNN(K Nearest Neighbors) and alos it is widely used. In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors)

12. What is A/B testing?

ANSWER

1.A/B testing is basically used to compare two different products those comparison is made from the user input, whether the user is clicking on product 1 or user is clicking on product 2.

2.Technically we can say that A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

3.For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply

scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

4.In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

5.It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

6.A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and postmodification results to decide whether the changes are working as desired or not.

7.A/B testing doesn't work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

13. Is mean imputation of missing data acceptable practice?

ANSWER

1.Mean imputation is a very basic type of imputation technique.

2.It is the only tested function that takes no advantage of the time series characteristics or relationship between the variables.

3.Mean imputation reduces the variance of the imputed variables.

4.Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.

5.Mean imputation does not preserve relationships between variables such as correlations.

6.Thus for a professional Data Scientist mean imputation of missing data is not at all acceptable practice.

14. What is linear regression in statistics?

ANSWER

Linear regression is used to study the linear relationship between a dependent variable Y (blood pressure) and one or more independent variables X (age, weight, sex).

More genreally, a linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term, also called the intercept term.

There are two types of linear regression one is simple linear regression and other is multiple lineaar regeression.

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y . The variable we are basing our predictions on is called the predictor variable and is referred to as X .

When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the predictions of Y when plotted as a function of X form a straight line.

When there are multiple predictor variable, the prediction method is called multiple linear regresson. The statistcal formaula for multiple linear regresson is y = 0 + 1x1 + 2x2 + ... + pxp.

15. What are the various branches of statistics?

ANSWER

There are two branches of statistics they are:

i) Descriptive statistics
ii) Inferential statistics

Descriptive statistics: It organizes raw data into meaningful information.An house hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full. Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this particular branch “ descriptivestatistics". That is,it focuses on collecting,

summarizing and presenting set of data. For example, Industrial statistics, population statistics, trade statistics etc.

Inferential statistics: It analyses sample data to draw conclusion about population. Marketing research team of a company wants to know how far the people need a particular product manufactured by the company. There is one hundred thousand population in a particular city. It is bit difficult to go and ask all one hundred thousand people, due to time consumption and other factors. Hence, it takes a sample of 1000 people to draw conclusion for the whole population. That is making general statement from the study of particular cases or any treatment of data, which leads to prediction or inference concerning a larger group of data. For example, we want to have an idea about percentage of illiterates in a country. We take a sample from a population and the proportion of illiterates in the sample. That sample with the help of probability enables us to find the proportion to the original population.