STATISTICS WORKSHEET-1

Answers:

1) True

2)All of the mentioned

3) Modeling bounded count data

4) All of the mentioned

5)Poisson

6)False

7)Hypothesis

8)0

9)Outliers cannot conform to the regression relationship

10)A symmetrical data distribution, where most of the results lie near the mean. In graph form, normal distribution will appear as a bell curve

11)we need to first check whether the data missing completely at random, or not at random. Depending on this classification we can proceed with different methods to deal with missing data.

A. imputation with mean : Missing data is replaced by mean of the column

B. imputation with median: Missing data is replaced by median of the column

C. imputation with mode: Missing data is replaced by most repeated values of the column

D. imputation with linear regression : With real valued data, this is a common technique. The missing value is replaced by performing linear regression based on the other feature values.

12)It's a split testing, the process of comparing two variants of a page element, usually by testing user's response to variant A vs variant B and concluding which of the two variants is more effective.

13)it is a non-standard, but a fairly flexible imputation algorithm.

14)Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

15)there are two main branches of statistics

Inferential Statistic:- it is used to make inference and describe about the population. These statistics are more useful when it's not possible to examine each member of the population.

Descriptive statistics: it is used to get a brief summary of data in numerical or graphical form.

SQL WORKSHEET 1

Answers:

1) create & Alter

2)Update & Delete

3)Structured Query Language

4)Data Definition Language

5)Data Manipulation Language

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

7)Alter Table A ADD COLUMN D float

8)Alter Table A Drop Column D

9)Alter Table A Column D float to int

10)Alter Table A Add Primary key B

11)it is a type of data management system that is designed to enable and support BI activities, especially analytics. It helps to perform queries and analysis and often contain large amounts of historical data.

12)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.

13)Data load is controlled

Large amounts of historical data are used

Data is denormalised for simplification and to improve performance

14)it is a dimensional model, in which data is organised into facts and dimensions. A fact is an event that is counted or measured. A dimension contains reference information about the fact.

A star schema is diagramed by surrounding each fact with its associated dimensions. The resulting diagram resembles a star.

15)The name define as SET Language.

It is a very high level language with dynamic typing and dynamic data structures, based on the mathematical notion of set.

The language introduced a fundamentally new paradigm in programming in which sets, ordered sets and maps are the principal data structures and the programs are expressed in terms of set constructors, set operations, and predicates on sets.

MACHINE LEARNING WORKSHEET 1

Answers:

1)4

2)1,2 and 4

3)formulating the clustering problem

4)Euclidean distance

5)Divisive clustering

6)All answers are correct

7)Divide the data points into groups

8)Unsupervised learning

9)All of the above

10)K-means clustering algorithm

11)All of the above

12)Labeled data

13)calculate the distances, link the clusters, and choosing a solution by selecting the right number of clusters

14) To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

15)Cluster analysis is the task of grouping a set of data points in such a way that they can be characterised by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

A) Centroid Clustering:

In centroid cluster analysis we choose the number of clusters that we want to classify and A line is then drawn separating the data points into the clusters based on their proximity to the centroids.The algorithm will then reposition the centroid relative to all the points within each cluster.

B)Density Clustering:

It groups data points by how densely populated they are. The algorithm will select a random point then start measuring the distance between each point around it. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

C)Distribution clustering

It is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier a cluster.

D) Connectivity clustering

It recognises each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related.

The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.