ASSIGNMENT-1

MACHINE LEARNING WORKSHEET:-

ANSWERS:-

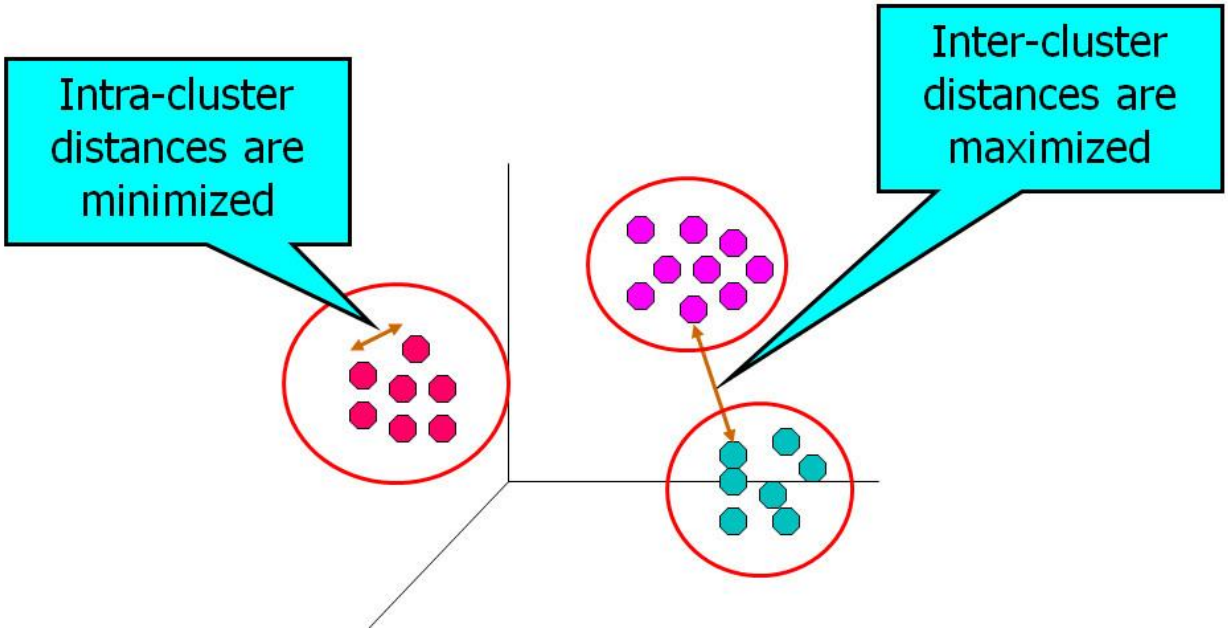
1. (b)
2. (d)
3. (d)
4. (a)
5. (b)
6. (d)
7. (a)
8. (b)
9. (d)
10. (a)
11. (d)
12. (a)
13. For instance, by varying k from 1 to 10 **clusters**. For each k, **calculate** the total within-**cluster** sum of square (wss). Plot the curve of wss according to the number of **clusters** k. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of **clusters**.
14. To **measure** a **cluster's** fitness within a **clustering**, we can compute the average silhouette coefficient value of all objects in the **cluster**. To **measure** the **quality** of a **clustering**, we can use the average silhouette coefficient value of all objects in the data set.
15. **Definition of Cluster Analysis:-**

- It groups the similar data in same group.
- The goal of this procedure is that the objects in a group are similar to one another and are different from the objects in other groups.
- Greater the similarity within a group and greater difference between the groups, more distinct the clustering.
- Cluster analysis provides a potential relationship and constructs systematic structure in large number of variables and observations.

Main objectives of clustering are:

1. Intra-cluster distance is minimized.

2. Intra-cluster distance is maximized.



Types of cluster analysis:-

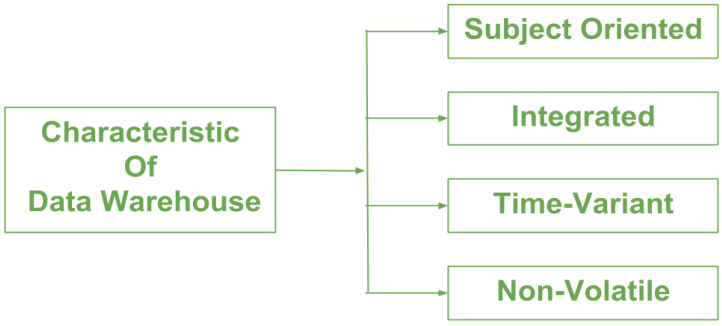
1. **Hierarchical clustering:** Also known as 'nesting clustering' as it also clusters to exist within bigger clusters to form a tree.
2. **Partition clustering:** It's simply a division of the set of data objects into non-overlapping clusters such that each object is in exactly one subset.
3. **Exclusive Clustering:** They assign each value to a single cluster.
4. **Overlapping Clustering:** It is used to reflect the fact that an object can simultaneously belong to more than one group.
5. **Fuzzy clustering:** Every object 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.
6. **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 approximately equal diameter.

SQL WORKSHEET:-

ANSWERS-

1. a, d
2. b, c
3. b
4. b
5. a
6. ..
7. ..
8. ..
9. …
10. (a)
11. **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.
12. **Difference between OLAP & OLTP:-**
    - Online Analytical Processing (OLAP) is a category of software tools that analyze 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.

13. Characteristic of Data Ware-house:-



a) **Subject-oriented –**

A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations. It can be achieved on specific theme. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc.

b) **Integrated**-

It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various data warehouse in shared and generally granted manner.

c) **Time-Variant**-

In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated.

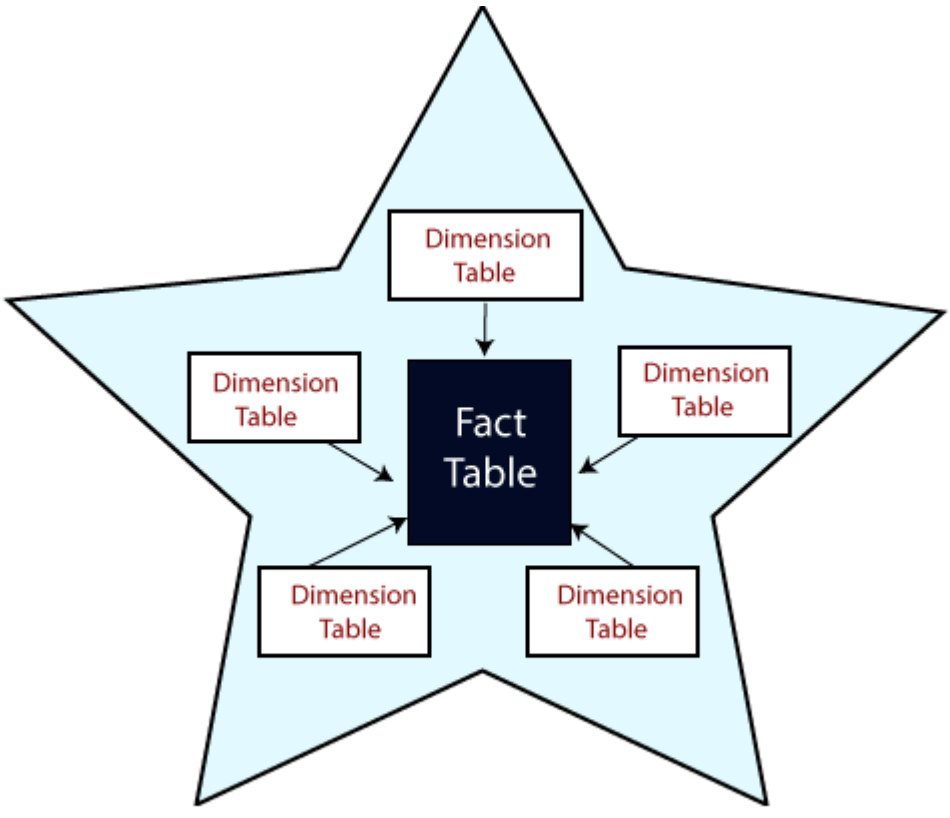
d) **Non-Volatile**-

As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is

inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantities on logical business. It evaluates the analysis within the technologies of warehouse.

14. A star schema is the elementary form of a dimensional model, in which data are organized into **facts** and **dimensions**. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model. The star schema is the explicit data warehouse schema. It is known as **star schema** because the entity-relationship diagram of this schema simulates a star, with points, diverge from a central table. The center of the schema consists of a large fact table, and the points of the star are the dimension tables.



Star Schema

15. Set Theory as a Language (or Set Language), SETL is a high-level programming language that's based on the mathematical theory of sets. It was developed in the early 1970's by mathematician Professor J. Schwartz. SETL is an interpreted language with a syntax that is resembles C and in many cases similar to Perl. In SETL every statement is terminated by a semicolon. Variable names are case-insensitive and are automatically determined by their last assignment.

STATISTICS WORKSHEET:-

ANSWERS-

1. (a)
2. (a)
3. (a)
4. (d)
5. (c)
6. (b)
7. (b)
8. (b)
9. (c)
10. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

11. Method to handle missing data:-

a) Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.

b) Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). ...

c) Random Forest

12. A/B testing (also known as split testing or bucket 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.

13. It is a non-standard, but a fairly flexible imputation algorithm. It uses Random Forest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

14. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which

variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15. Branches of Statistics

(a) Descriptive Statistics:-

Descriptive statistics is the first part of statistics that deals with the collection of data. People seem it too easy, but it is not that easy. The statisticians need to be aware of the designing and experiments. They also need to choose the right focus group and avoid biases. In contrast, Descriptive statistics are used in use to do various kinds of analysis on different studies.

Descriptive statistics have two parts:-

- Central tendency measures
- Variability measures

To help understand the analyzed data, the tendency measures and variability measures use tables, general discussions, and charts.

Measures of Central Tendency

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

- **Mean**

Mean is a conventional method used to describe the central tendency. Typically, to calculate the average of values, count all values, and then divide them with the number of available values.

- **Median**

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

- **Mode**

The mode is the frequently occurring value in the given data set.

- **Measures of Variability**

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

(b) Inferential Statistics:-

The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Inference statistics often speak in terms of probability by using descriptive statistics. Besides, these techniques are used primarily by a statistician for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Most predictions of the future and generalization on a population study of a smaller specimen are in the scope of the inference statistics. Besides, most of the social sciences experiments deal with the study of a small sample population that helps determines the behavior of the community.

Designing a real experiment, the researcher can bring conclusions relevant to his study. When making conclusions, it should be cautious not to draw wrongly or biased

Different types of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis