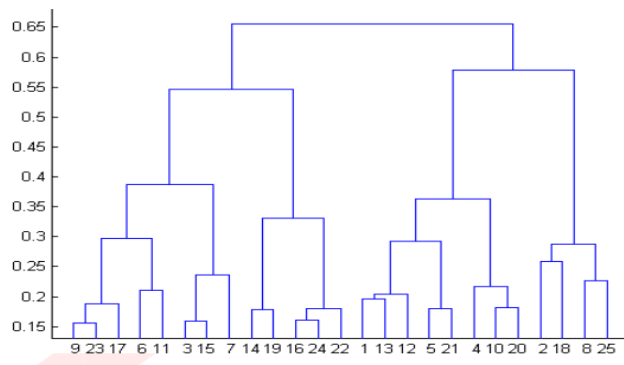


# ASSIGNMENT – 1

## Machine Learning

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

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



a) 2

**b) 4**

c) 6

d) 8

2. In which of the following cases will K-Means 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 non-convex shapes

Options:

a) 1 and 2

b) 2 and 3

c) 2 and 4

**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**

4. The most commonly used measure of similarity is the or its square.

**a) Euclidean distance**

b) city-block distance

c) Chebyshev's distance

d) Manhattan 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) Non-hierarchical clustering

**b) Divisive clustering**

c) Agglomerative clustering

d) K-means clustering

6. Which of the following is required by K-means clustering?

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

**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

8. Clustering is a

a) Supervised learning

**b) Unsupervised learning**

c) Reinforcement learning

d) None

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

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

**a) K-means clustering algorithm**

b) K-modes clustering algorithm

c) K-medians clustering algorithm

d) None

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 non-convex shapes

**d) All of the above**

12. For clustering, we do not require

**a) Labeled data**

b) Unlabeled data

c) Numerical data

d) Categorical data

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

13. How is cluster analysis calculated?

**Answer:**

The hierarchical cluster analysis follows three basics steps , first is calculate the distance, second is link the clusters and third is choose a solution by selecting the right number of clusters .First we have to select the variable upon which we base are clusters.

14. How is cluster quality measured?

**Answer:**

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

15. What is cluster analysis and its types?

**Answer:**

Cluster analysis is a multivariate data mining technique whose goal is to groups objects eg. Products respondents, or other entities based on a set of user selected characteristics or attributes.

# WORKSHEET 1 SQL

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

2. Which of the following is/are DML commands in SQL?

A) Update    B) Delete

C) Select    D) Drop

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

4. Full form of DDL is:

A) Descriptive Designed Language    B) Data Definition Language

C) Data Descriptive Language    D) None of the above.

5. DML is:

A) Data Manipulation Language    B) Data Management Language

C) Data Modeling Language    D) None of these

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

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

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

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

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

9. Which of the following statements can be used to change the data type (from float to int ) of the column

DoF 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**

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**

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

11. What is data-warehouse?

**Answer:**

**A data warehouse is a central repository of information that can be analyzed to make more informed decisions data flows into a data warehouse form transactional systems, relational databases, and other sources, typically on a regular Candance.**

12. What is the difference between OLTP VS OLAP?

**Answer:**

Difference b/w OLTP and OLAP

- Online analytical processing(OLAP) is a category of software tools that analyzed data stored in a data base, where else online transaction processing(OLTP) supports transactions oriented applications in a 3-tier architecture.
- OLAP creates a single platform for all types of business analysis needs which includes planning, budgeting, forecasting, and analysis, while OLTP is useful for administering day-to-day transaction of an organization.
- OLAP is characterized by a large volume of data, while OLTP is characterized by large number of short online transaction.
- In OLAP a data warehouse is created uniquely so that it can integrate different data source for building a consoliolated database, where as OLTP uses traditional DBMS.

13. What are the various characteristics of data-warehouse?

**Answer:**

Data warehouses are characterized by being:

These may include cloud, relational databases, flat files, structured and semi structured data, metadata and master data. The sources are combined in a manner that's consistent, reliable and ideally certifiable, providing a business with confidence in the data's quality.

14. What is Star-Schema?

**Answer:**

A star scheme is a multi-dimensional data model used to organize data in a database so that it is easy to understand and analyze.

15. What do you mean by SETL?

**Answer:**

SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets. It was originally developed by (Jack) Jacob J.

## STATISTICS WORKSHEET-1

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

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

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

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 called chi-squared distribution

**d) All of the mentioned**

5. \_\_\_\_\_ random variables are used to model rates.

- a) Empirical
- b) Binomial

**c) Poisson**

- d) All of the mentioned

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

- a) True

**b) False**

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

- a) Probability

**b) Hypothesis**

- c) Causal
- d) None of the mentioned

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

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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

**Answer:**

**A normal distribution is a type of continuous probability distribution in which most data points cluster towards the middle of the range, while the rest taper off symmetrically towards either extreme. The middle of the range is also known as the mean of the distribution.**

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

**Answer:**

**One way of handling missing values is the deletion of the rows or columns having null values. If any columns have more than half of the values as null then you can drop the entire column. In the same ways, rows can also be dropped if having one or more columns values as null. To select an imputation method, the one which the lowest over all error value is chosen. Even though ever common this method has two main short comings : One has to somewhat artificially mask observation which itself comes with problems such as choosing how many observations and which ones to mask.**

12. What is A/B testing?

**Answer:**

**A/B testing also known as split testing and bucket testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better.**

13. Is mean imputation of missing data acceptable practice?

**Answer:**

**Mean imputation is typically considered terrible practices since it ignores feature correlation .**

14. What is linear regression in statistics?

**Answer:**

**Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called in the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. The variable you are using to predict the other variable's value is called the independent variable.**

15. What are the various branches of statistics?



**Answer:**

There are three real branches of statistics :-

1. Data Collection
2. Descriptive Statistics
3. Inferential Statistics