**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 dendrogram:

Ans: c) 6

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

Ans: d) 1, 2 and 4

3. The most important part of is selecting the variables on which clustering is based.

Ans: a) interpreting and profiling clusters

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

Ans: 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.

Ans: b) Divisive clustering

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

Ans: d) All answers are correct

7. The goal of clustering is to-

Ans: a) Divide the data points into groups

8. Clustering is a-

Ans: b) Unsupervised learning

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

Ans: d) All of the above

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

Ans: b) K-modes clustering algorithm

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

Ans: d) All of the above

12. For clustering, we do not require-

Ans: 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:

1) **calculate** the distances,

2) link the **clusters**, and

3) choose a solution by selecting the right number of **clusters**. First, we have to select the variables upon which we base our **clusters**

14. How is cluster quality measured?

Ans: 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. What is cluster analysis and its types?

Ans: **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.

These **types** are:

Centroid **Clustering**,

Density **Clustering**

 Distribution **Clustering**, and

Connectivity **Clustering**

**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?

Ans: A) Create

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

And: A) Update

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

3. Full form of SQL is:

Ans: B) Structured Query Language

4. Full form of DDL is:

Ans: B) Data Definition Language

5. DML is:

Ans: : 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?

Ans: 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?

Ans: 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?

Ans: A) Table A Drop 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?

Ans: 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?

Ans: C) Alter Table A Add Primary key B

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

11. What is data-warehouse?

Ans: A **Data Warehousing** is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

12. What is the difference between OLTP VS OLAP?

Ans:OLTP and OLAP both are the online processing systems. OLTP is a transactional processing while OLAP is an analytical processing system. ... The basic difference between OLTP and OLAP is that OLTP is an online database modifying system, whereas, OLAP is an online database query answering system.

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

Ans: Data warehouse can be controlled when the user has a shared way of explaining the trends that are introduced as specific subject. Below are major **characteristics** of data warehouse:

1. **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.  
     
   A data warehouse never put emphasis only current operations. Instead, it focuses on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.
2. **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.  
     
   A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benefits in effective analysis of data. Reliability in naming conventions, column scaling, encoding structure etc. should be confirmed. Integration of data warehouse handles various subject related warehouse.
3. **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.
4. **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 quantity on logical business. It evaluates the analysis within the technologies of warehouse.  
     
   In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:
   * Data Loading
   * Data Access

14. What is Star-Schema??

Ans:  the **star schema** is the simplest style of [data mart](https://en.wikipedia.org/wiki/Data_mart) [schema](https://en.wikipedia.org/wiki/Logical_schema) and is the approach most widely used to develop data warehouses and dimensional data marts.[[1]](https://en.wikipedia.org/wiki/Star_schema#cite_note-1) The star schema consists of one or more [fact tables](https://en.wikipedia.org/wiki/Fact_table) referencing any number of [dimension tables](https://en.wikipedia.org/wiki/Dimension_(data_warehouse)). The star schema is an important special case of the [snowflake schema](https://en.wikipedia.org/wiki/Snowflake_schema), and is more effective for handling simpler queries.

15. What do you mean by SETL?

Ans: Short for **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.

**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.

Ans: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?

Ans:a) Central Limit Theorem

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

Ans:d) All of the mentioned

4. Point out the correct statement.

Ans:d) All of the mentioned

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

Ans:c) Poisson

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

Ans:b) False

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

Ans:a) Probability

8. 4. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data.

Ans:a) 0

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

Ans: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 What do you understand by the term Normal Distribution?

Ans: **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. How do you handle missing data? What imputation techniques do you recommend?

**Techniques for Handling the Missing Data**

1. Listwise or case deletion.
2. Pairwise deletion.
3. Mean substitution.
4. Regression imputation.
5. Last observation carried forward.
6. Maximum likelihood.
7. Expectation-Maximization.
8. Multiple imputation.

12. What is A/B testing?

Ans: **A/B testing (also known as split testing) is a process of showing two variants of the same web page to different segments of website visitors at the same time and comparing which variant drives more conversions.**

13. Is mean imputation of missing data acceptable practice?

Ans: It is a non-standard, but a fairly flexible **imputation** algorithm. It uses RandomForest 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. What is linear regression in statistics?

Ans: **linear regression** is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modelling the relationship between a [scalar](https://en.wikipedia.org/wiki/Scalar_(mathematics)) response and one or more explanatory variables (also known as [dependent and independent variables](https://en.wikipedia.org/wiki/Dependent_and_independent_variables)). The case of one explanatory variable is called [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression); for more than one, the process is called **multiple linear regression**.[[1]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-Freedman09-1) This term is distinct from [multivariate linear regression](https://en.wikipedia.org/wiki/Multivariate_linear_regression), where multiple [correlated](https://en.wikipedia.org/wiki/Correlation_and_dependence) dependent variables are predicted, rather than a single scalar variable

15. What are the various branches of statistics?

Ans: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific **analysis** of data and both are equally important for the student of statistics.