MACHINE LEARNING

Answers:

1. B) 4
2. D) 1,2 and 4
3. D) formulating the clustering problem
4. A) Euclidean Distance
5. B) Divisive clustering
6. D) All answers are correct
7. A) Divide the data points into groups
8. B) Unsupervised learning
9. A) K- Means clustering
10. A) K-means clustering algorithm
11. D) All of the above
12. A) Labeled data
13. How is cluster analysis calculated? – Cluster analysis works by portioning n objects into number of clusters in which each object belongs to the cluster with the nearest mean. The optimal no of clusters is calculated with max separation distance by selecting no of clusters at random and calculating the Euclidian from the identified centroids. Centroids are varied till best characteristics are arrived at. Optimal no of clusters can identified with methods like elbow method.
14. How is cluster quality measured? - There are various indexes for measuring cluster quality which shall be used based on aim of clustering i.e Between-cluster separation, Within-cluster homogeneity (low distances), Good representation of data by centroids and Within-cluster homogeneous distributional shape. Few are:
    A. Silhouette coefficient
    B. Adjusted Rand index
    C. Fowlkes-Mallows scores
    D. Calinski-Harabasz Index
    E. Davies-Bouldin Index etc.

15. What is cluster analysis and its types? – Clustering is a type of unsupervised machine learning method where data is divided into a number of clusters in such a manner that the data points belonging to a cluster have similar characteristics and different clusters are having separate characters. Common types are:

    A. Exclusive clustering
    B. Overlapping clustering
    C. Hierarchical clustering

WORKSHEET 1 SQL

1. A) Create & D) Alter
2. A) Update & B) Delete
3. B) Structured Query Language
4. B) Data Definition Language
5. A) Data Manipulation Language
6. C) Create Table A (B int,C float)
7. B) Alter Table A ADD COLUMN D float
8. B) Alter Table A Drop Column D
9. B) Alter Table A Alter Column D int
10. A) Alter Table A Add Constraint Primary Key B
11. What is data-warehouse? - A Data Warehouse is defined as a central repository where information is coming from one or more data sources.
12. What is the difference between OLTP VS OLAP? Both are online processing systems but 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? - There are three prominent data warehouse characteristics:
    A. Integrated: The way data is extracted and transformed is uniform, regardless of the original source.
    B. Time-variant: Data is organized via time-periods (weekly, monthly, annually, etc.).
    C. Non-volatile: A data warehouse is not updated in real-time. It is periodically updated via the uploading of data, protecting it from the influence of momentary change.
14. What is Star-Schema? - Star schema is the type of multidimensional model which is used for data warehouse. In star schema, the fact tables and the dimension tables are contained. In this schema fewer foreign-key join is used. This schema forms a star with fact table and dimension tables.
15. What do you mean by SETL? – Simple, Extract, Transformation, load is simple data integration based on the ETL (Extract, Transform, and Load) design pattern. Is a simple, flexible, cost-effective solution that can easily be extended later to help drive a digital transformation strategy for the organization.

STATISTICS WORKSHEET-1

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned

5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship
10. What do you understand by the term Normal Distribution? – Normal distribution means where data has no bias left or right, that is where mean = median = mode.
11. How do you handle missing data? – There are various methods to handle missing data, some are:
    A. Deletion – List wise deletion or Pair wise deletion.
    B. Mean/Median/Mode – Generalized or Similar case imputation
    C. Prediction model imputation
    D. KNN imputation
12. What imputation techniques do you recommend? – There is no one fit all methodology for imputation. Selection of appropriate technique depends on several factors like type of variable, nature & criticality of data etc.
    A. Mice,
    B. KNN,
    C. Missforest, and
    D. Fuzzy K-Means clustering etc are recommended.
13. What is A/B testing? - It is an analytical method for making decisions that estimates population parameters based on sample statistics, is basically statistical hypothesis testing. A process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.
14. Is mean imputation of missing data acceptable? Mean imputation is an easy to use technique but the problem of mean imputation is that mean imputation reduces variance of the data and has potential for introducing bias and hence reduces model accuracy, so mean imputation is not a recommended method and can be used if there is very little missing data then it will do relatively little harm, will be much easier to implement.