Worksheet 2 Machine Learning

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:
   Ans. (B) 4
2. In which of the following cases will K-Means clustering fail to give good results?
   Ans. (D) 1,2 and 4 (Data points with outliers, Data points with different densities, Data points with non-convex shapes).
3. The most important part of is selecting the variables on which clustering is based.
   Ans. (D) Formulating the clustering problem
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 above
10. Which version of the clustering algorithm is most sensitive to outliers?
    Ans. (A) K-means clustering algorithm
11. Which of the following is a bad characteristic of a dataset for clustering analysis?
    Ans. (D) All of above
12. For clustering, we do not require-
    Ans. (A) Labeled data
13. How is cluster analysis calculated?
    Ans. It is basically a type of unsupervised learning method. Clustering is the task of dividing the population or data points into

a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Some methods are:

a. Density-Based Methods

b. Hierarchical Based Methods

14. How is cluster quality measured?

Ans. Measures for Quality of Cluster is done as If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations.

15. What is cluster analysis and its types?

Ans. Cluster analysis is a multivariate data mining technique whose goal is to groups objects (e.g., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Types are:

a. **Hierarchical Cluster Analysis**- In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster

b. **Centroid-based Clustering**- In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set.

c. **Distribution-based Clustering-** It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster.

d. **Density-based Clustering-** In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters

Worksheet 2 SQL

1. Which of the following is/are DDL commands in SQL?
   Ans. (A) Create & (D) Alter
2. Which of the following is/are DML commands in SQL?
   Ans. (A) Update (B) Delete & (C) Select
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. (B) Data Management 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. (B) Alter Table A Drop Column 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. (B) Alter Table A Alter Column D 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. (A) Alter Table A Add Constraint Primary Key B
11. What is data-warehouse?
    Ans. It is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning.
12. What is the difference between OLTP VS OLAP?
    Ans. OLAP (*online analytical processing*) and OLTP (online transactional processing) the main difference between them is OLAP is analytical in nature, and OLTP is transactional.

Worksheet 2 SQL

OLAP are designed for multidimensional analysis of data in a data warehouse, which contains both historical and transactional data. For e.g., financial analysis, budgeting, and forecast planning

OLTP is designed to support transaction-oriented applications by processing recent transactions as quickly and accurately as possible. For e.g., ATMs, online bookings, reservation systems.

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

Ans. The various characteristics of data-warehouse are:

a. Subject-oriented
b. Time-variant
c. Integrated
d. Persistent and non-volatile

14. What is Star-Schema?

Ans. It is the explicit data warehouse schema. It is known as star schema because the entity-relationship diagram of this schemas 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.

15. What do you mean by SETL?

Ans. SETL (Set Language) is an interpreted language with a syntax that is loosely C-like and in many cases similar to Perl. For example, variables types are determined automatically by their last assignment and every statement is terminated by a semicolon.

Worksheet 2 Statistics

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 id 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. (B) Modeling bounded count data
4. Point out the correct statement.
   Ans. (D) All of the mentioned
5. ______ random variables are used to model rates.
   Ans. (C) Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.
   Ans. (B) False
7. Which of the following testing is concerned with making decisions using data?
   Ans. (B) Hypothesis
8. 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
10. What do you understand by the term Normal Distribution?
    Ans. It is the type of distribution in which mean is equals to zero and the standard deviation is 1. In normal distribution there are three types skewed right, skewed left and symmetric distribution. It is also called as "Z- distribution".
11. How do you handle missing data? What imputation techniques do you recommend?
    Ans. We can handle missing data in these ways like to fill the missing data with 0 and we can also fill the missing data by taking mean or median. Mean or Median imputation technique been recommended. Some techniques also be used like fillna , front fill, back fill, etc.

Worksheet 2 Statistics

12. What is A/B testing?

Ans. It is also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. Is mean imputation of missing data acceptable practice?

Ans. Yes, mean imputation of missing data acceptable practice imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size.

14. What is linear regression in statistics?

Ans. Linear regression is used to predict the future growth or results. It shows the linear relationship between two variables. In linear regression, there are two kinds of variables are: the dependent variable and the independent variable. Formulae that it used is "y=mx+C"

15. What are the various branches of statistics?

Ans. The two main branches of statistics are:

Descriptive Statistics
Inferential Statistics

In the case of descriptive statistics, the data or collection of data is described in summary. But in the case of inferential stats, it is used to explain the descriptive one.