MACHINE LEARNING

ASSIGNMENT – 1

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: b) 4

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: 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 the 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 the above

12. For clustering, we do not require-
Ans: a) Labeled data

13. How is cluster analysis calculated?
Ans:

I. By using KMeans Clusters:
   A. Random positioning of cluster focal points. Start with k centroids by putting them at random places.
   B. Compute distance from every data points and cluster them accordingly.
   C. Now adjust centroids so that they become center of gravity of the new formed clusters.
   D. Again recluster every points based on their distance from centroids.
   E. Again adjust the centroids like step no C.
   F. Recompute clusters and repeat this till data points stop changing clusters.

II. By Using Hierarchical Clustering analysis:
   A. Calculate the distances.
   B. Link the clusters.
   C. Choose a solution by selecting the right number of clusters.

14. How is cluster quality measured?

Ans: To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.The silhouette score falls within the range [-1, 1].

The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. The score of less than 0 means that data belonging to clusters may be wrong/incorrect.

15. What is cluster analysis and its types?

Ans: Clustering is the process of dividing the datasets into groups, consisting of similar data-points. Clustering is a type of unsupervised machine learning, which is used when you have unlabeled data. Clustering algorithms can be categorised into:

Centroid Clustering: This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you're a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

Density Clustering: Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are.

Distribution Clustering: Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to acluster.

Connectivity Clustering : Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.

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

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) 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?
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 : B) Alter table (B primary key)

11. What is data-warehouse?

Ans: Data warehouse is a relational database that is designed for query and analysis. It contains various heterogeneous types of data from multiple source. It usually contains historical data derived from transaction data, but it can include data from other sources. Historical data is the data kept over years and can be used for trend analysis, make future predictions and decision support.

A data warehouse integrates the data from one or more databases, so that analysis can be done to get results, such as the best performing school in a city. But constructing of warehouse can be expensive.

12. What is the difference between OLTP VS OLAP?

Ans :

OLAP (ONLINE ANALYTICAL PROCESSING)

1. Consists of historical data from various Databases.

2. It is subject oriented. Used for Data Mining, Analytics, Decision making, etc.

3. The data is used in planning, problem solving and decision making.

4. It reveals a snapshot of present business tasks.

5. Large amount of data is stored typically in TB, PB.

6. Relatively slow as the amount of data involved is large. Queries may take hours.

7. It only need backup from time to time as compared to OLTP.

8. This data is generally managed by CEO, MD, GM.

9. Only read and rarely write operation.

OLTP (ONLINE TRANSACTION PROCESSING)

1. Consists only operational current data.

2. It is application oriented. Used for business tasks.

3. The data is used to perform day to day fundamental operations.

4. It provides a multi-dimensional view of different business tasks.

5. The size of the data is relatively small as the historical data is archived. For ex MB, GB.

6. Very Fast as the queries operate on 5% of the data.

7. Backup and recovery process is maintained religiously.

8. This data is managed by clerks, managers.

9. Both read and write operations.

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

Ans : Main features of Data Warehouse are:

1. Subject Oriented: Databases organized based on major subjects like the customer, supplier, product etc..
2. Integrated: Databases constructed by integrating multiple different sources.
3. Time-Variant: Data is stored to provide information in a historical perspective.
4. Non-Volatile: Physically separate store of data for transfer from application data.

14. What is Star-Schema??

Ans: A star schema is a data warehousing architecture model where one fact table references multiple dimension tables, which, when viewed as a diagram, looks like a star with the fact table in the center and the dimension tables radiating from it. It is the simplest among the data warehousing schemas and is currently in wide use.

15. What do you mean by SETL?

Ans: Full form of SETL is "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

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 : 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. 10. Usually replacing the standard error by its estimated value does change the CLT.
b) False

7. 1. Which of the following testing is concerned with making decisions using data?
Ans : b) Hypothesis

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

10. What do you understand by the term Normal Distribution?
Ans : A standard normal distribution is a continuous probability distribution with density function $1/2\pi\sqrt{}\exp(-1/2(x^2))$. This has mean 0 and standard deviation 1. In general if X has a standard normal distribution then Y=μ+ has a normal distribution with mean μ and standard deviation σ.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans : There is no fixed rule to deal with missing data but one could use any of the heuristics mentioned below.

1. The most common way of dealing with missing data is to remove all rows with missing data if there are not too many rows with missing data.
2. If more than 50-60% of rows of a specific column are missing data, it is common to remove the column. The main problem with removing missing data thus, is that it could introduce substantial bias.
3. Imputation of data is also a common technique used to deal with missing data where the data is substituted with the best guess.
    a) Imputation with mean : Missing data is replaced by the mean of the column. This is a commonly used technique. However, this might not be appropriate if the data is not unimodal (for example suppose we fill missing value of weights, the mean of weights for males might be different from females and this might not be a unimodal distribution).
    b) Imputation with median : Missing data is replaced by the median of the column. A median is better than the mean when there are outliers, but once again, if the data is multi-model with multiple clusters, median might not work.
    c) Imputation with Mode: Missing data is replaced with mode of the column. This also leads to similar problems as the above two methods.
    d) Imputation with linear regression : With real valued data, this is another common technique. The missing value is replaced by performing linear regression based on the other feature values. This overcomes the problems with the above simpler forms of imputation.

I will recommend to follow KNN Imputer because KNNImputer is also a multivariate approach however it uses kNN which would average the features based on some distance metric (usually Euclidean). kNNImputer could

do just as good and could be faster I imagine. If you choose to iteratively run regression on a bunch of combinations of many columns then it will take a lot of time.

12. What is A/B testing?

13. Is mean imputation of missing data acceptable practice?

Ans: Missing data is replaced by the mean of the column. This is a commonly used technique. However, this might not be appropriate if the data is not unimodal (for example suppose we fill missing value of weights, the mean of weights for males might be different from females and this might not be a unimodal distribution).

14. What is linear regression in statistics?

Ans : Linear regression is a statistical technique where the score of a variable (criterion) Y is predicted from the score of a second (predictor) variable X. X is referred to as the predictor variable and Y as the criterion variable. In other words: Linear Regression is a field of study which emphasizes on the statistical relationship between two continuous variables known as Predictor and Response variables. (Note: when there are more than one predictor variables then it becomes multiple linear regression.)

15. What are the various branches of statistics?

Ans:

Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics
- Descriptive Statistic.

- Inferential Statistic.

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.

i. Measures of Central Tendency

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

ii. 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.

iii. 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.

iv. Mode

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

v. 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 determine 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