WORK SHEET SET 1

MACHINE LEARNING ASSIGNMENT 1

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

ANS : b) 4

1. 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.
a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem

ANS : 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

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.
a) Non-hierarchical clustering
b) Divisive clustering
c) Agglomerative clustering
d) K-means clustering

ANS : b) Divisive 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

ANS : 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

ANS : a) Divide the data points into groups

8. Clustering is a
a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None

ANS : b) Unsupervised learning

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

ANS : 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

ANS : a) K-means clustering algorithm

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

ANS : d) All of the above

12. For clustering, we do not require
a) Labeled data
b) Unlabeled data
c) Numerical data
d) Categorical data

ANS : a) Labeled data

13. How is cluster analysis calculated?

Cluster Analyses can be found in *Analyze/Classify…*. There are three methods for the cluster analysis:

- ✓ ***K-Means Cluster*,**
- ✓ ***Hierarchical Cluster*,**
- ✓ ***Two-Step Cluster*.**

K-means cluster is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from *1* (all cases in one cluster) to *n* (each case is an individual cluster). Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.

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.

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

14. How is cluster quality measured?

Measuring Clustering Quality

Suppose you have assessed the clustering tendency of a given data set. You may have also tried to predetermine the number of clusters in the set. You can now apply one or multiple clustering methods to obtain clustering of the data set. "How good is the clustering generated by a method, and how can we compare the clustering generated by different methods?"

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.

If ground truth is available, it can be used by extrinsic methods, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use intrinsic methods, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of "cluster labels." Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

Extrinsic Methods

When the ground truth is available, we can compare it with a clustering to assess the clustering. Thus, the core task in extrinsic methods is to assign a score, Q(C,Cg), to a clustering, C, given the ground truth, Cg. Whether an extrinsic method is effective largely depends on the measure, Q, it uses.

Intrinsic Methods

When the ground truth of a data set is not available, we have to use an intrinsic method to assess the clustering quality. In general, intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are. Many intrinsic methods have the advantage of a similarity metric between objects in the data set.

15. What is cluster analysis and its types?

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 techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

TYPES OF CLUSTER ANALYSIS

There are four basic types of cluster analysis used in data science.

i. Centroid Clustering,
ii. Density Clustering
iii. Distribution Clustering
iv. Connectivity Clustering.

i. 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.

The algorithm will start by randomly selecting centroids (cluster centers) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iteratations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats.

ii. 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. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

iii. 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.

These maps look a lot like targets at an archery range. In the event that a data point hits the bulls eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bulls eye represents lessening percentage or certainty.

Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to a cluster.

iv. 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 WORKSHEET 1

1. Which of the following is/are DDL commands in SQL?
A) Create
B) Update
C) Delete
D) ALTER

ANS : A)Create and D)alter

2. Which of the following is/are DML commands in SQL?
A) Update
B) Delete
C) Select
D) Drop

ANS : A) Update and B) Delete

3. Full form of SQL is:
A) Strut querying language
B) Structured Query Language
C) Simple Query Language
D) None of them

ANS : B) Structured Query Language

4. Full form of DDL is:
A) Descriptive Designed Language
B) Data Definition Language
C) Data Descriptive Language
D) None of the above.

ANS : Data Definition Language

5. DML is:
A) Data Manipulation Language
B) Data Management Language
C) Data Modeling Language
D) None of these

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

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?
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

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?
A) Table A Drop D
B) Alter Table A Drop Column D
C) Delete D from A
D) None of them

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?
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

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?
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

ANS : C) Alter Table A Add Primary key B

11. What is data-warehouse?

A **data warehouse** is a relational or multidimensional database that is designed for query and analysis. Data warehouses are not optimized for transaction processing, which is the domain of OLTP systems. Data warehouses usually consolidate historical and analytic data derived from multiple sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources.

A data warehouse usually stores many months or years of data to support historical analysis. The data in a data warehouse is typically loaded through an extraction, transformation, and loading (ETL) process from one or more data sources such as OLTP applications, mainframe applications, or external data providers.

Users of the data warehouse perform data analyses that are often time-related. Examples include consolidation of last year's sales figures, inventory analysis, and profit by product and by customer. More sophisticated analyses include trend analyses and data mining, which use existing data to forecast trends or predict futures. The data warehouse typically provides the foundation for a business intelligence environment.

12. What is the difference between OLTP VS OLAP?

OLTP	OLAP
✓ It is an online transactional system and manages database modification.	✓ It is an online data retrieving and data analysis system.
✓ It is a system that manages transaction-oriented applications on the internet for example, ATM.	✓ It is an online system that reports to multidimensional analytical queries like financial reporting, forecasting, etc.
✓ It is an online database modifying system	✓ It is an online database query answering system.
✓ OLTP has short transactions.	✓ OLAP has long transactions.
✓ The processing time of a transaction is comparatively less in OLTP.	✓ The processing time of a transaction is comparatively more in OLAP.
✓ Simpler queries.	✓ Complex queries.
✓ OLTP database must maintain data integrity constraint.	✓ OLAP database does not get frequently modified. Hence, data integrity is not affected.
✓ Tables in OLTP database are normalized (3NF).	✓ Tables in OLAP database are not normalized.
✓ OLTP and its transactions are the original source of data.	✓ Different OLTPs database becomes the source of data for OLAP.

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

The **key characteristics** of a data warehouse are as follows:

- Some data is denormalized for simplification and to improve performance
- Large amounts of historical data are used
- Queries often retrieve large amounts of data
- Both planned and ad hoc queries are common
- The data load is controlled

14. What is Star-Schema?

A star schema is the elementary form of a dimensional model, in which data are organized into **facts** and **dimensions**. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model. The star schema is the explicit data warehouse schema. It is known as **star schema** because the entity-relationship diagram of this schema 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?

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. 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 WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

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?
a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

ANS : a) Central Limit Theorem

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

ANS : b) Modeling bounded count data

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

ANS : d) All of the mentioned

5. _______ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

ANS : c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

ANS : b) False

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

ANS : b) Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the
original data.
a) 0
b) 5
c) 1
d) 10

ANS : a) 0

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

ANS : c) Outliers cannot conform to the regression relationship

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

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.

KEY TAKEAWAYS

- ✓ A normal distribution is the proper term for a probability bell curve.
- ✓ In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- ✓ Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- ✓ In reality, most pricing distributions are not perfectly normal.

Understanding Normal Distribution

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

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

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Imputation techniques:

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

i. There may not be enough observations with non-missing data to produce a reliable analysis

ii. In predictive analytic, missing data can prevent the predictions for those observations which have missing data

iii. External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2.Multivariate Imputation by Chained Equations (MICE):

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, Bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.

Secondly, as mentioned above, the algorithm treats different variables differently. So, all categorical variables should be treated as factor variables before implementing MICE.

You can also ignore some variables as predictors or skip a variable from being imputed using the MICE library in R. Additionally, the library also allows you to set a method of imputation discussed above depending upon the nature of the variable.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of over fitting. The extent of overfilling leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data.

For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, over fitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.

The MICE library in R also allows imputations by random forest by setting the method to “rf”. The authors of the MICE library have provided an example on how to implement the random forest method here.

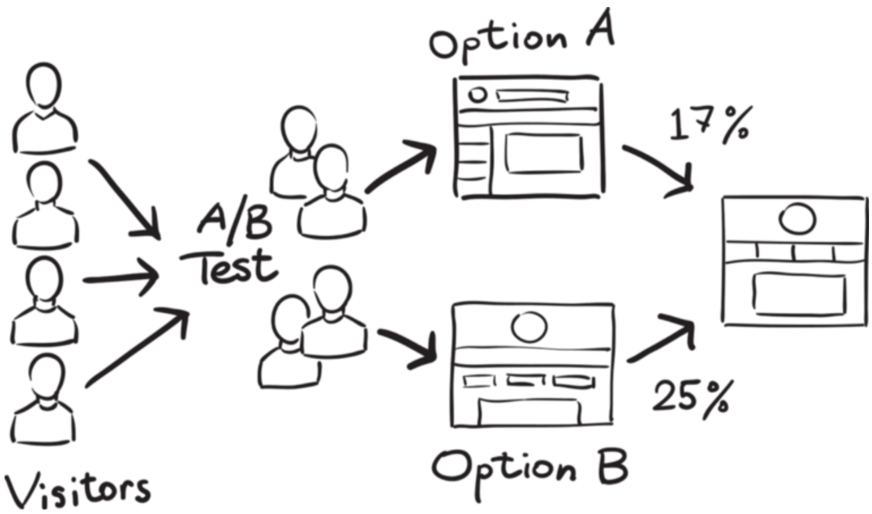
To sum up data imputations is tricky and should be done with care. It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. While using the above algorithms, predictor variables should be set up carefully to avoid confusion in the methods implemented during imputation. Finally, you can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

12. Is mean imputation of missing data acceptable practice?

It is a non-standard, but a fairly flexible imputation method when the dataset we want to use for Machine Learning contains missing data..The quick and easy workaround is to substitute a mean for numerical features and use a mode for categorical ones. Even better, someone might just insert 0's or discard the data and proceed to the training of the model.

Simply we can say Mean imputation is So simple. And yet, so dangerous.

Problem 1: Mean and mode ignore feature correlations

Let's have a look at a very simple example to visualize the problem. The following table have 3 variables: Age, Gender and Fitness Score. It shows a Fitness Score results (0–10) performed by people of different age and gender.

	Age	Gender	Fitness_Score
0	20	M	8
1	25	F	7
2	30	M	7
3	35	M	7
4	36	F	6
5	42	F	5
6	49	M	6
7	50	F	4
8	55	M	4
9	60	F	5
10	66	M	4
11	70	F	3
12	75	M	3
13	78	F	2

Table with correct, non-missing data

Now let's assume that some of the data in Fitness Score is actually missing, so that after using a mean imputation we can compare results using both tables.

	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

Mean Imputed

	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Mean Imputation of the Fitness_Score

Imputed values don't really make sense — in fact, they can have a negative effect on accuracy when training our ML model. For example, 78 year old women now has a Fitness Score of 5.1, which is typical for people aged between 42 and 60 years old. Mean imputation doesn't take into account a fact that Fitness Score is correlated to Age and Gender features. It only inserts 5.1, a mean of the Fitness Score, while ignoring potential feature correlations.

Problem 2:Mean reduces a variance of the data

Based on the previous example, variance of the real Fitness Score and of their mean imputed equivalent will differ. Figure below presents the variance of those two cases:

	Variance
Real Data	3.302198
Missing Data	1.300000

Fitness Score variance of the real and mean imputed data

As we can see, the variance was reduced (that big change is because the dataset is very small) after using the Mean Imputation. Going deeper into mathematics, a smaller variance leads to the narrower confidence interval in the probability distribution. This leads to nothing else than introducing a bias to our model.

14. What is linear regression in statistics?

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

i. If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
ii. If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or

to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

15. What are the various branches of statistics?

There are two main categories in Statistics, namely:

a. Descriptive Statistics
b. Inferential Statistics

A. <u>Descriptive Statistics:</u>

CONCEPT: The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES: The mean age of citizens who live in a certain geographical area, the mean length of all books about statistics, the variation in the time that visitors spent visiting a website.

INTERPRETATION: You are most likely to be familiar with this branch of statistics because many examples arise in everyday life. Descriptive statistics serves as the basis for analysis and discussion in fields as diverse as securities trading, the social sciences, government, the health sciences, and professional sports. Descriptive methods can seem deceptively easy to apply because they are often easily accessible in calculating and computing devices. However, this ease does not mean that descriptive methods are without their pitfalls.

B. <u>Inferential Statistics :</u>

CONCEPT: The branch of statistics that analyzes sample data to reach conclusions about a population.

EXAMPLE: A survey that sampled 1,264 women found that 45% of those polled considered friends or family as their most trusted shopping advisers and only 7% considered advertising as their most trusted shopping adviser. By using methods discussed in Section 6.4, you can use these statistics to draw conclusions about the population of all women.

INTERPRETATION: When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. This deeper level of analysis means that inferential statistical methods can be easily misapplied or misconstrued, and that many inferential methods require a calculating or computing device.