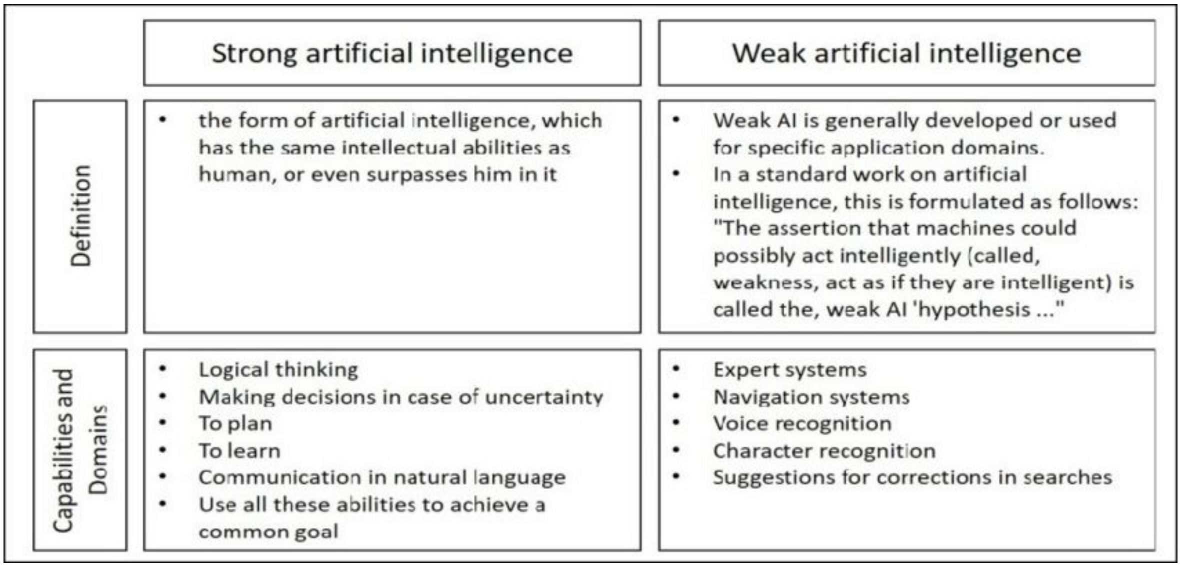
Some fundamental questions up to ML (CS-3035) mid-semester exam.
(but not limited to):

https://www.javatpoint.com/clustering-in-machine-learning

1. What is learning in computer science/machine learning?

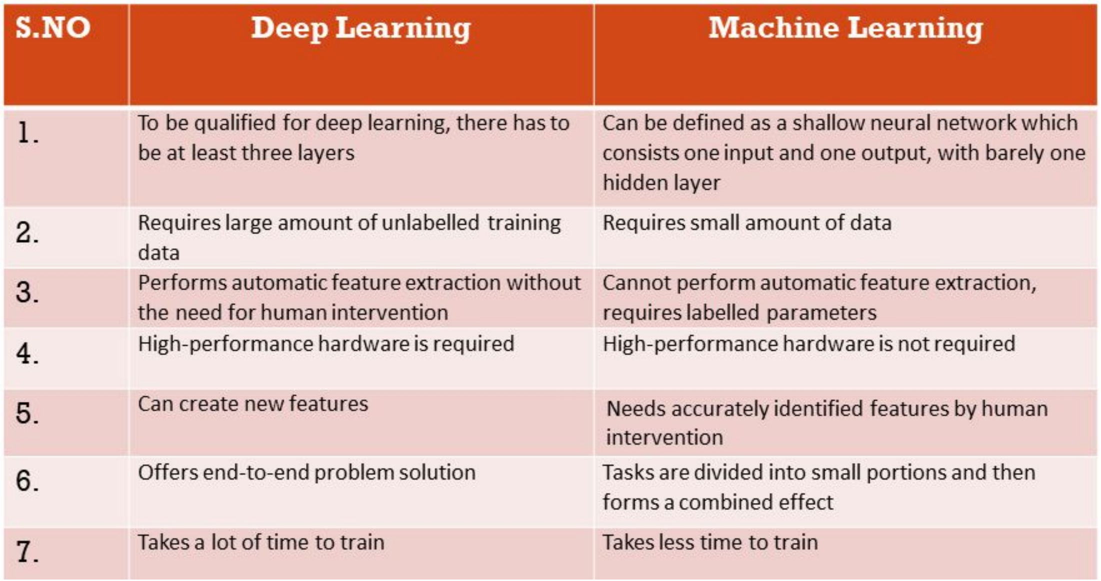
'Learning' in machine learning refers to the determination of model parameters using the given training dataset.

2. Differentiate between weak-AI and Strong-AI.



	Strong artificial intelligence	Weak artificial intelligence
Definition	• the form of artificial intelligence, which has the same intellectual abilities as human, or even surpasses him in it	• Weak AI is generally developed or used for specific application domains. • In a standard work on artificial intelligence, this is formulated as follows: "The assertion that machines could possibly act intelligently (called, weakness, act as if they are intelligent) is called the, weak AI 'hypothesis ..."
Capabilities and Domains	• Logical thinking • Making decisions in case of uncertainty • To plan • To learn • Communication in natural language • Use all these abilities to achieve a common goal	• Expert systems • Navigation systems • Voice recognition • Character recognition • Suggestions for corrections in searches

3. Differentiate between Machine Learning and Deep Learning.



S.NO	Deep Learning	Machine Learning
1.	To be qualified for deep learning, there has to be at least three layers	Can be defined as a shallow neural network which consists one input and one output, with barely one hidden layer
2.	Requires large amount of unlabelled training data	Requires small amount of data
3.	Performs automatic feature extraction without the need for human intervention	Cannot perform automatic feature extraction, requires labelled parameters
4.	High-performance hardware is required	High-performance hardware is not required
5.	Can create new features	Needs accurately identified features by human intervention
6.	Offers end-to-end problem solution	Tasks are divided into small portions and then forms a combined effect
7.	Takes a lot of time to train	Takes less time to train

4. Explain different types of learning using suitable real-world examples.

Learning Problems

- **1. Supervised Learning**

  An example of supervised learning is text classification problems. In this set of problems, the goal is to predict the class label of a given piece of text. One particularly popular topic in text classification is to predict the sentiment of a piece of text, like a tweet or a product review.

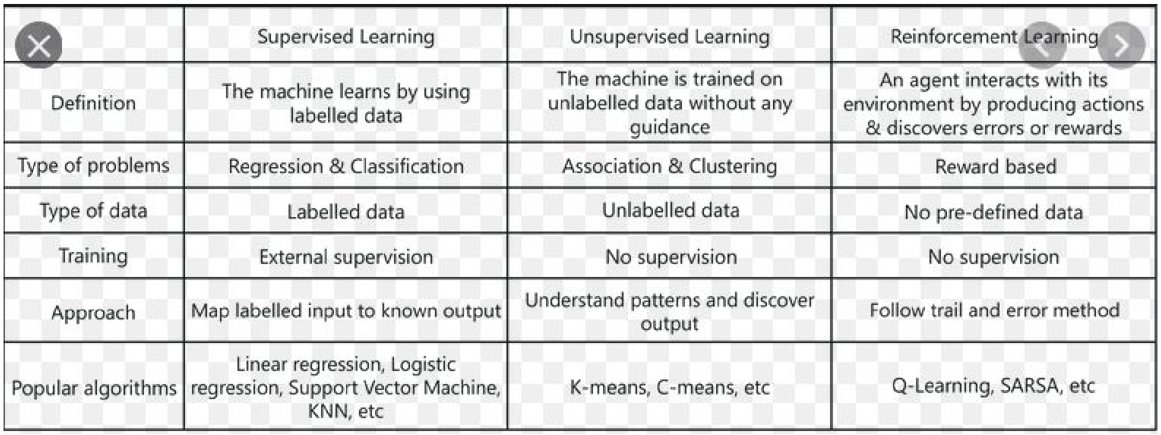
- **2. Unsupervised Learning**

  Finding customer segments Clustering is an unsupervised technique where the goal is to find natural groups or clusters in a feature space and interpret the input data. There are many different clustering algorithms. One common approach is to divide the data points in a way that each data point falls into a group that is similar to other data points in the same group based on a predefined similarity or distance metric in the feature space. Clustering is commonly used for determining customer segments in marketing data. Being able to determine different segments of customers helps marketing teams approach these customer segments in unique ways. (Think of features like gender, location, age, education, income bracket, and so on.)

  https://www.springboard.com/blog/lp-machine-learning-unsupervised-learning-supervised-learning/

- **3. Reinforcement Learning**

  The example of reinforcement learning is your cat is an agent that is exposed to the environment. The biggest characteristic of this method is that there is no supervisor, only a real number or reward signal.



	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labelled data	The machine is trained on unlabelled data without any guidance	An agent interacts with its environment by producing actions & discovers errors or rewards
Type of problems	Regression & Classification	Association & Clustering	Reward based
Type of data	Labelled data	Unlabelled data	No pre-defined data
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output	Follow trail and error method
Popular algorithms	Linear regression, Logistic regression, Support Vector Machine, KNN, etc	K-means, C-means, etc	Q-Learning, SARSA, etc

5. Differentiate between supervised learning and unsupervised learning.

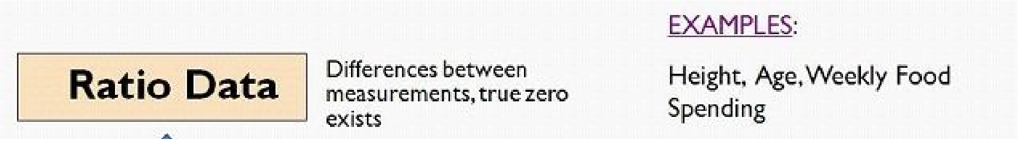
Unsupervised Learning	**Supervised Learning**
The machine is given huge sets of data that are not labelled as inputs to analyse.	The input is in the form of raw data that is labelled.
The machine needs to figure out the output on its own by identifying patterns in the raw data provided to it.	The machine is already fed with the required feature set to classify between inputs (hence the term 'supervised').
Divided into two types of problems – Association (where we want to find a set of rules that describe our data) and Clustering (where we want to find groups in our data).	Divided into two types of problems – Regression (outputs are real values) and Classification (outputs are categories).
K-means for clustering problems and Apriori algorithm for association rule learning problems.	Linear regression for regression problems, Random Forest for classification and regression problems, Support Vector Machines for classification problems.

6. Explain different data types used in modern machine learning paradigm with examples.

https://medium.com/swlh/data-types-in-statistics-used-for-machine-learning-5b4c24ae6036

7. What is the true zero point for a numeric data type? Explain with an example.

Absolute/true zero means that the **zero point** represents the absence of the property being measured



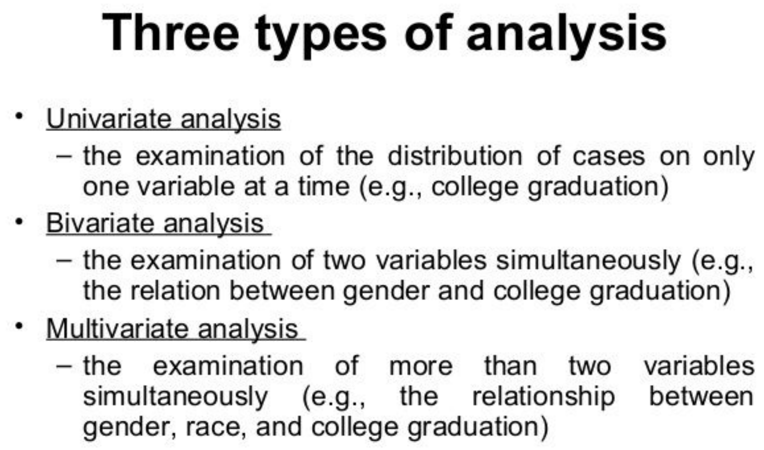
Ratio Data

Differences between measurements, true zero exists

EXAMPLES:

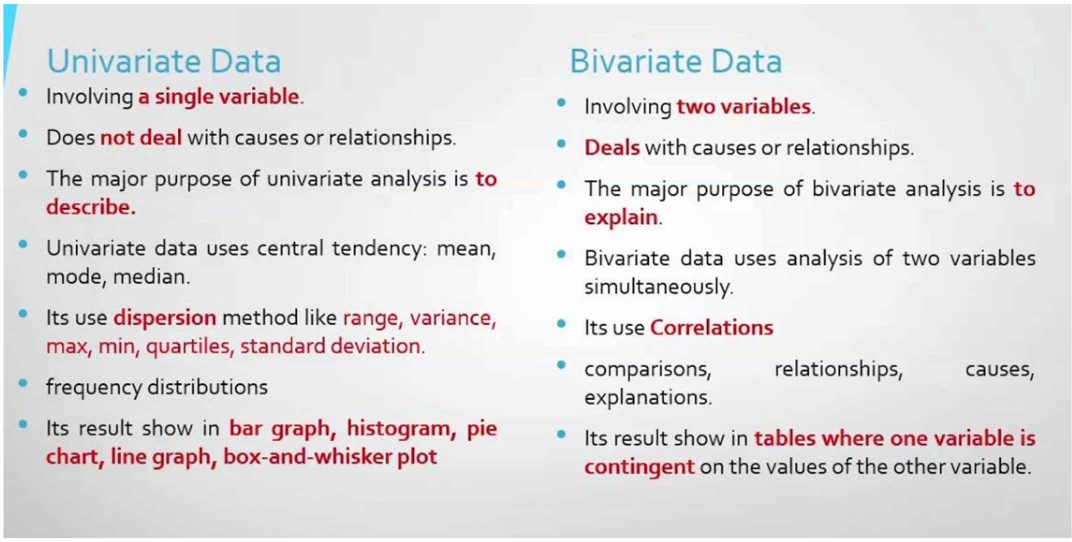
Height, Age, Weekly Food Spending

8. Differentiate between Univariate and Multivariate data analysis.



Three types of analysis

- Univariate analysis
  - the examination of the distribution of cases on only one variable at a time (e.g., college graduation)
- Bivariate analysis
  - the examination of two variables simultaneously (e.g., the relation between gender and college graduation)
- Multivariate analysis
  - the examination of more than two variables simultaneously (e.g., the relationship between gender, race, and college graduation)

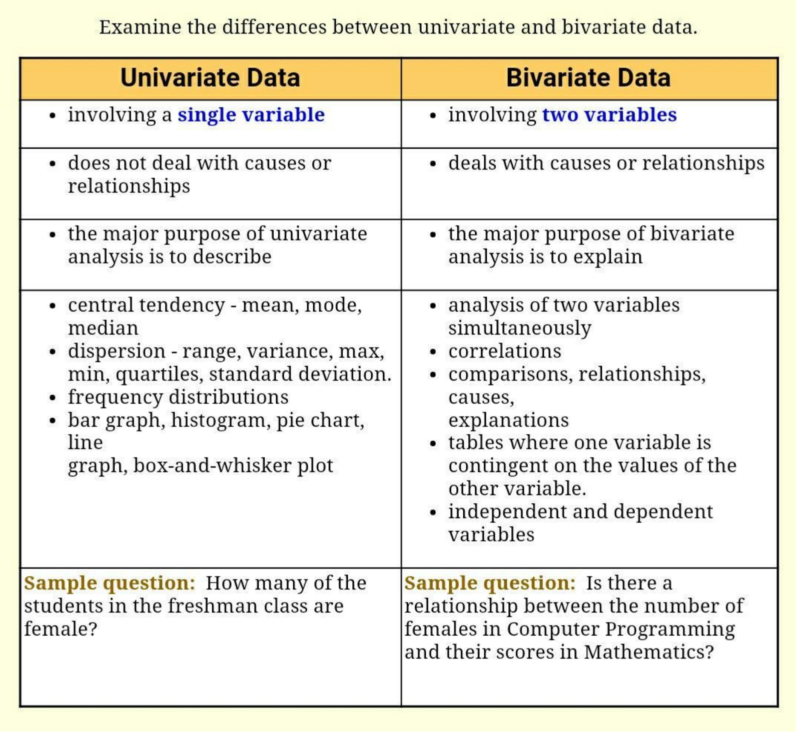


Univariate Data

- Involving **a single variable**.
- Does **not deal** with causes or relationships.
- The major purpose of univariate analysis is **to describe.**
- Univariate data uses central tendency: mean, mode, median.
- Its use **dispersion** method like range, variance, max, min, quartiles, standard deviation.
- frequency distributions
- Its result show in **bar graph, histogram, pie chart, line graph, box-and-whisker plot**

Bivariate Data

- Involving **two variables**.
- **Deals** with causes or relationships.
- The major purpose of bivariate analysis is **to explain**.
- Bivariate data uses analysis of two variables simultaneously.
- Its use **Correlations**
- comparisons, relationships, causes, explanations.
- Its result show in **tables where one variable is contingent** on the values of the other variable.



Examine the differences between univariate and bivariate data.

Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

9. What do you mean by central tendency? Explain it with suitable examples.

Central tendency means measuring the center or distribution of location of values of a data set. It gives an idea of the average value of the data in the data set and also an indication of how widely the values are spread in the data set.

Measures of Central Tendency

Generally, the central tendency of a dataset can be described using the following measures:

- **Mean (Average):** Represents the sum of all values in a dataset divided by the total number of the values.
- **Median:** The middle value in a dataset that is arranged in ascending order (from the smallest value to the largest value). If a dataset contains an even number of values, the median of the dataset is the mean of the two middle values.
- **Mode:** Defines the most frequently occurring value in a dataset. In some cases, a dataset may contain multiple modes while some datasets may not have any mode at all.

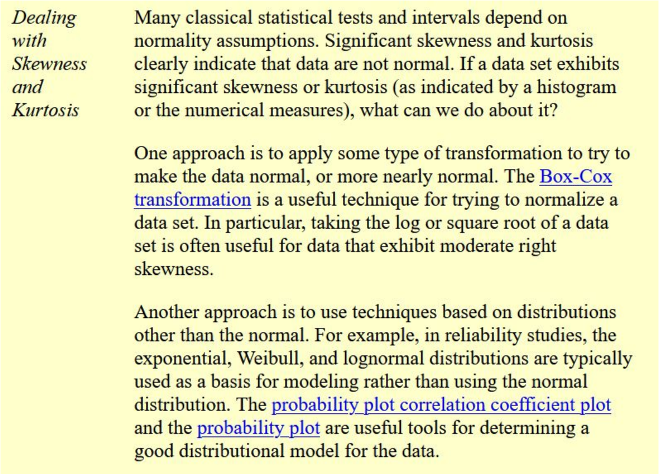
Some of the important examples of central tendency include mode, median, arithmetic mean and geometric mean, etc.

10. What is skewness? Explain at least one remedy for it.

Skewness is a quantifiable measure of how distorted a data sample is from the normal distribution. In normal distribution, the data is represented graphically in a bell-shaped curve, where the mean (average) and mode (maximum value in the data set) are equal.
Remedy:Log Transform
Log transformation is most likely the first thing you should do to remove skewness from the predictor.It can be easily done via *Numpy*, just by calling the log() function on the desired column.
Remedy 2:Square Root Transform:The square root sometimes works great and sometimes isn't the best suitable option. In this case, I still expect the transformed distribution to look somewhat exponential, but just due to taking a square root the range of the variable will be smaller.

11. What is kurtosis? Discuss at least one solution for it

Kurtosis is a measure of whether the data **are** heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high **kurtosis** tend to have heavy tails, or outliers. Data sets with low **kurtosis** tend to have light tails, or lack of outliers. A uniform distribution **would** be the extreme case.



Dealing with Skewness and Kurtosis

Many classical statistical tests and intervals depend on normality assumptions. Significant skewness and kurtosis clearly indicate that data are not normal. If a data set exhibits significant skewness or kurtosis (as indicated by a histogram or the numerical measures), what can we do about it?

One approach is to apply some type of transformation to try to make the data normal, or more nearly normal. The Box-Cox transformation is a useful technique for trying to normalize a data set. In particular, taking the log or square root of a data set is often useful for data that exhibit moderate right skewness.

Another approach is to use techniques based on distributions other than the normal. For example, in reliability studies, the exponential, Weibull, and lognormal distributions are typically used as a basis for modeling rather than using the normal distribution. The probability plot correlation coefficient plot and the probability plot are useful tools for determining a good distributional model for the data.

12. Explain the similarity and dissimilarity between Normal distribution and Student's T-test.

Dissimilarities:

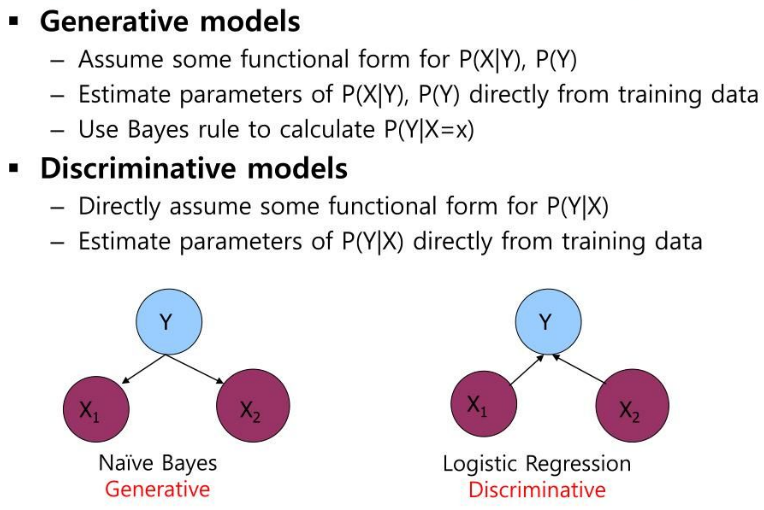
1. Standard from the comparison of normal distribution and t-distribution density function. The density function of the t distribution has a thicker tail than the standard normal distribution.
2. in addition to all t-distribution will give better results than normal distribution whenever we have less number of data points(<30 in general).

3. The difference is that the t distribution is leptokurtic, and so has higher kurtosis than the normal distribution. That means that, for a t and a normal with the same mean and variance, data from the t distribution have a tendency to appear either closer to the mean or farther from the mean than typical normal data, with a more sudden transition in between. And that means that the probability of obtaining values very far from the mean is larger than in the normal distribution.

Similarities:

1. Like the normal distribution, the t-distribution is symmetric. If you think about folding it in half at the mean, each side will be the same
2. Like a standard normal distribution (or z-distribution), the t-distribution has a mean of zero. The normal distribution assumes that the population standard deviation is known.

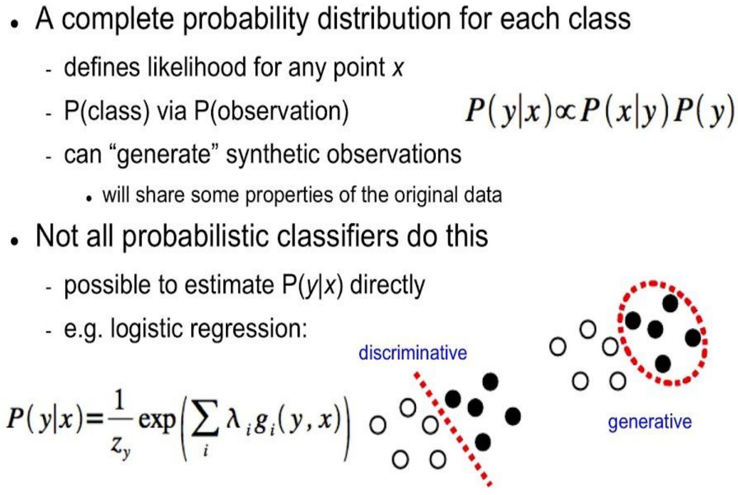
13. Explain discriminative and generative learning models with suitable Examples.



- **Generative models**
  - Assume some functional form for P(X|Y), P(Y)
  - Estimate parameters of P(X|Y), P(Y) directly from training data
  - Use Bayes rule to calculate P(Y|X=x)
- **Discriminative models**
  - Directly assume some functional form for P(Y|X)
  - Estimate parameters of P(Y|X) directly from training data

Naïve Bayes
Generative

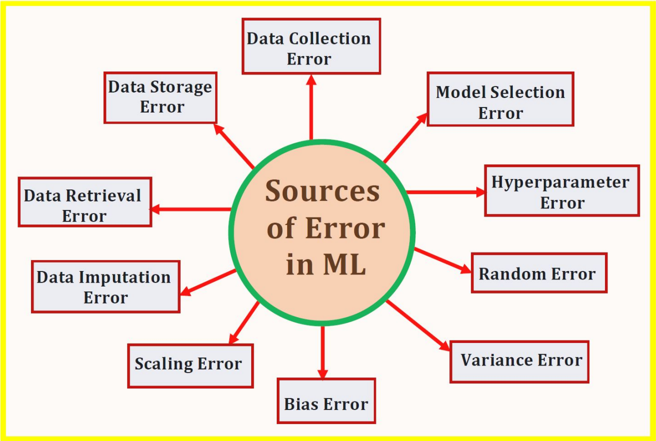
Logistic Regression
Discriminative



- A complete probability distribution for each class
  - defines likelihood for any point *x*
  - P(class) via P(observation) $P(y|x) \propto P(x|y)P(y)$
  - can "generate" synthetic observations
    - will share some properties of the original data
- Not all probabilistic classifiers do this
  - possible to estimate P(*y*|*x*) directly
  - e.g. logistic regression:

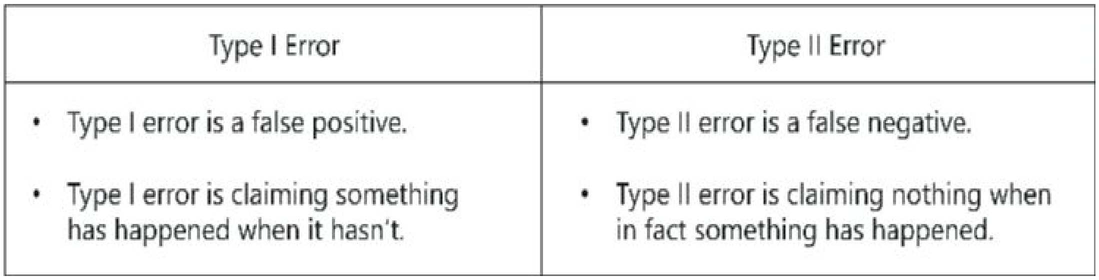
$$P(y|x)=\frac{1}{z_y}\exp\left(\sum_i \lambda_i g_i(y,x)\right)$$

14. What are the different types of errors used in machine learning?



https://medium.com/towards-artificial-intelligence/12-common-errors-in-machine-learning-729cb9d0952a

15. Differentiate between Type-I and Type-II errors.



Type I Error	Type II Error
• Type I error is a false positive. • Type I error is claiming something has happened when it hasn't.	• Type II error is a false negative. • Type II error is claiming nothing when in fact something has happened.

16. Explain different evaluation metrics used in classification.

https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

17. What are the properties of a distance metric?

Distance metrics play an important role in machine learning. They provide a strong foundation for several machine learning algorithms like k-nearest neighbors for supervised learning and k-means clustering for unsupervised learning. Different distance metrics are chosen depending upon the type of the data. Distance Metrics is used to know the input data pattern in order to make any Data Based decision. A good distance metric helps in improving the performance of Classification, Clustering and Information Retrieval process significantly.

18. Discuss different types of distance metrics using suitable expressions.

There are 4 types of Distance Metrics in Machine Learning.

1. Euclidean Distance
2. Manhattan Distance
3. Minkowski Distance
4. Hamming Distance

Euclidean Distance -

Euclidean Distance represents the shortest distance between two points

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

Manhattan Distance -

Manhattan Distance is the sum of absolute differences between points across all the dimensions.

$$d = |p_1 - q_1| + |p_2 - q_2|$$

Minkowski Distance

Minkowski Distance is the generalized form of Euclidean and Manhattan Distance.

$$D = \left(\sum_{i=1}^{n} |p_i - q_i|^p \right)^{1/p}$$

Hamming Distance

Hamming Distance measures the similarity between two strings of the same length. The Hamming Distance between two strings of the same length is the number of positions at which the corresponding characters are different.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

19. Explain the Minkowski Distance is a generalization of the Manhattan and Euclidean Distance metrics.

Minkowski Distance is the generalized form of Euclidean and Manhattan Distance.

Minkowski Distance calculates the distance between two points. It is a generalization of the Euclidean and Manhattan distance measures and adds a parameter, called the "*order*" or "*p*", that allows different distance measures to be calculated.

The Minkowski distance measure is calculated as follows:

$$D = \left(\sum_{i=1}^{n} |p_i - q_i|^p \right)^{1/p}$$

Where “*p*” is the order parameter.

When p is set to 1, the calculation is the same as the Manhattan distance. When p is set to 2, it is the same as the Euclidean distance.

- *p=1*: Manhattan distance.
  - *p=2*: Euclidean distance.

Intermediate values provide a controlled balance between the two measures.

It is common to use Minkowski distance when implementing a machine learning algorithm that uses distance measures as it gives control over the type of distance measure used for real-valued vectors via a hyperparameter “*p*” that can be tuned.

20. Explain any two bounded distance metrics with examples.

21. Explain the Voronoi diagram used in KNN?

https://www.youtube.com/watch?v=PGy1rATkViA

22. What is KNN classification? Why is it called lazy learning?

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.

K-NN is a lazy learner because it doesn’t learn a discriminative function from the training data but memorizes the training dataset instead. There is no training time in K-NN. The prediction step in K-NN is expensive. Each time we want to make a prediction, K-NN is searching for the nearest neighbors in the entire training set. An eager learner has a model fitting or training step. A lazy learner does not have a training phase.

23. Explain the KNN algorithm with a small hand crafted dataset and demonstrate its working principle.

1. Load the desired data.
2. Choose the value of k.
3. For getting the class which is to be predicted, repeat starting from 1 to the total number of training points we have.
4. The next step is to calculate the distance between the data point whose class is to be predicted and all the training data points. Euclidean distance can be used here.

5. Arrange the distances in non-decreasing order.
6. Assume the positive value of k and filtering k lowest values from the sorted list.
7. We have top k top distances.
8. Let ka represent the points that belong to the ath class among k points.
9. If ka>kb then put x in the class.

https://www.analyticssteps.com/blogs/how-does-k-nearest-neighbor-works-machine-learning-classification-problem

<Demonstration ke liye 50 chaap do>

24. Explain the advantages and disadvantages of KNN algorithm.

Some Advantages of KNN

- Quick calculation time
- Simple algorithm – to interpret
- Versatile – useful for regression and classification
- High accuracy – you do not need to compare with better-supervised learning models
- No assumptions about data – no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.

Some Disadvantages of KNN

- Accuracy depends on the quality of the data
- With large data, the prediction stage might be slow
- Sensitive to the scale of the data and irrelevant features
- Require high memory – need to store all of the training data
- Given that it stores all of the training, it can be computationally expensive

25. How many types of clustering are available in ML? Explain each type with examples.

Types of Clustering Methods :

The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

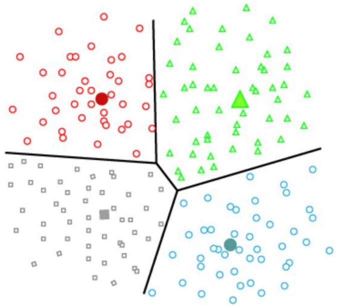
1. **Partitioning Clustering**
2. **Density-Based Clustering**

3. **Distribution Model-Based Clustering**
4. **Hierarchical Clustering**
5. **Fuzzy Clustering**

Partitioning Clustering

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the K-Means Clustering algorithm.

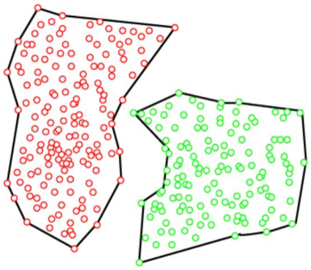
In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid**.**



Density-Based Clustering

The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

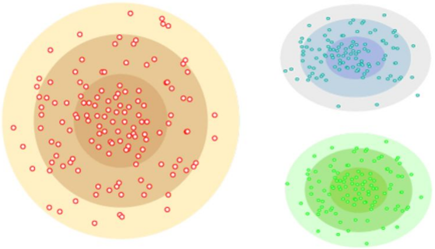
These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.



Distribution Model-Based Clustering

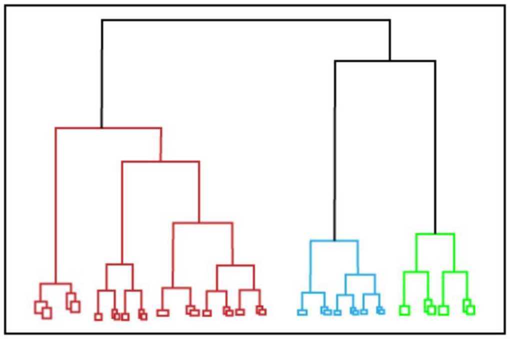
In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly Gaussian Distribution.

The example of this type is the Expectation-Maximization Clustering algorithm that uses Gaussian Mixture Models (GMM).



Hierarchical Clustering

Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm.**



Fuzzy Clustering

Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster. Each dataset has a set of membership coefficients, which depend on the degree of membership to be in a cluster. Fuzzy C-means algorithm is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm.

26. Explain the KMeans algorithm with a small hand crafted dataset and demonstrate its working principle.

The k-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (k) of clusters.
Initially k number of so called centroids are chosen. A centroid is a data point (imaginary or real) at the center of a cluster. In Praat each centroid is an existing data point in the given input data set, picked at random, such that all centroids are unique (that is, for all centroids ci and cj, ci ≠ cj). These centroids are used to train a kNN classifier. The resulting classifier is used to classify (using k = 1) the data and thereby produce an initial randomized set of clusters. Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The process of classification and centroid adjustment is repeated until the values of the centroids stabilize. The final centroids will be used to produce the final classification/clustering of the input data, effectively turning the set of initially anonymous data points into a set of data points, each with a class identity.

<Demonstration ke liye 49 chaap do>

27. Explain the advantages and disadvantages of KMeans clustering.

K-Means Advantages :

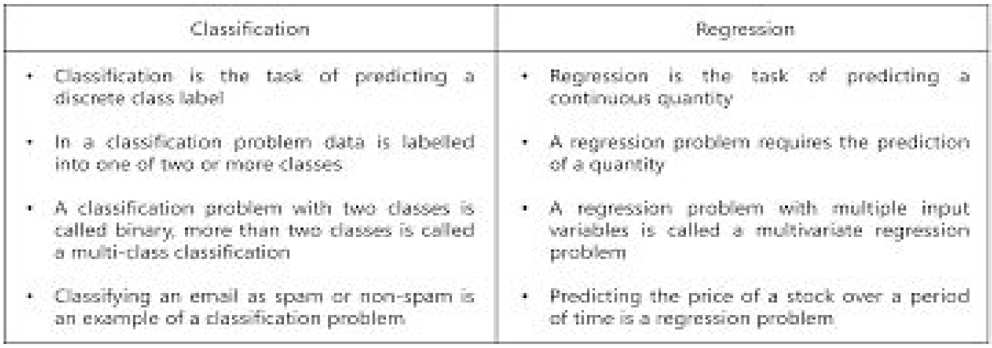
1) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.

2) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

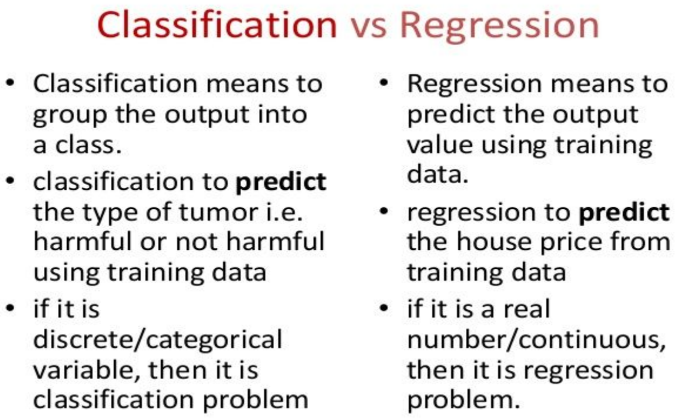
K-Means Disadvantages :

1) Difficult to predict K-Value.
2) With global cluster, it didn't work well.
3) Different initial partitions can result in different final clusters.
4) It does not work well with clusters (in the original data) of Different size and Different density

28. Differentiate between classification and regression.



Classification	Regression
• Classification is the task of predicting a discrete class label	• Regression is the task of predicting a continuous quantity
• In a classification problem data is labelled into one of two or more classes	• A regression problem requires the prediction of a quantity
• A classification problem with two classes is called binary, more than two classes is called a multi-class classification	• A regression problem with multiple input variables is called a multivariate regression problem
• Classifying an email as spam or non-spam is an example of a classification problem	• Predicting the price of a stock over a period of time is a regression problem



Classification vs Regression

- Classification means to group the output into a class.
- classification to **predict** the type of tumor i.e. harmful or not harmful using training data
- if it is discrete/categorical variable, then it is classification problem

- Regression means to predict the output value using training data.
- regression to **predict** the house price from training data
- if it is a real number/continuous, then it is regression problem.

29. Why do we call the Linear Regression a linear model?

Linear regression is called linear because you model your output variable (lets call it f(x)) as a linear combination of inputs and weights (lets call them x and w respectively).

30. Derive the cost function of Linear Regression using step by step Explanation.

31. What are the assumptions in Linear Regression? Also mention the solutions for them.

The Four Assumptions of Linear Regression

- Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
- Independence: The residuals are independent. ...
- Homoscedasticity: The residuals have constant variance at every level of x. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html
- Normality: The residuals of the model are normally distributed.

how to solve the assumptions:

https://towardsdatascience.com/assumptions-of-linear-regression-5d87c347140

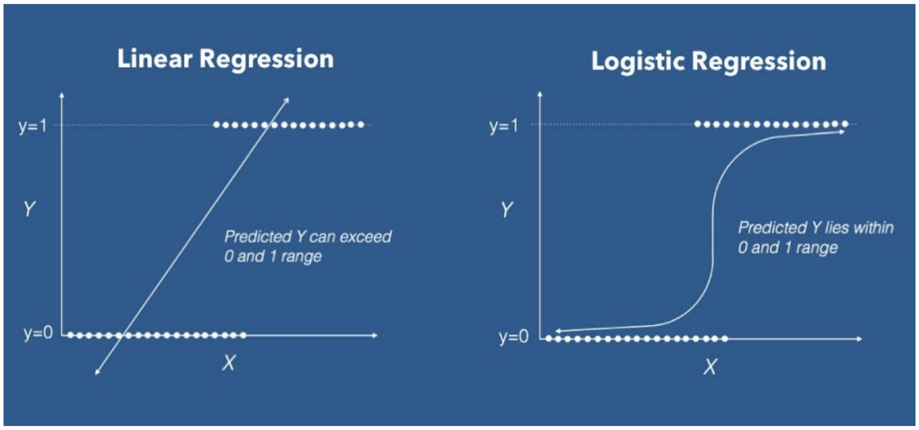
32. What is Polynomial Regression?

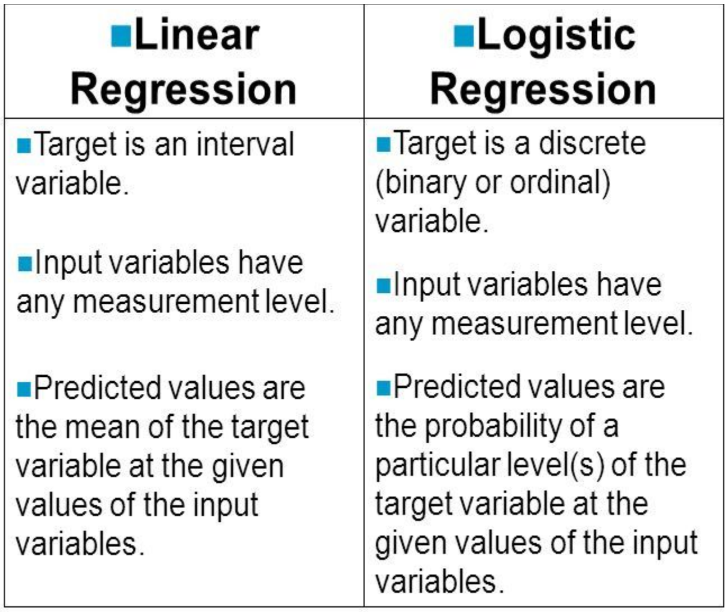
- Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

  $y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+...... b_nx_1^n$

- It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
- It is a linear model with some modification in order to increase the accuracy.
- The dataset used in Polynomial regression for training is of non-linear nature.
- It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.

33. Differentiate between Linear Regression and Logistic Regression.





■Linear Regression	■Logistic Regression
■Target is an interval variable. ■Input variables have any measurement level. ■Predicted values are the mean of the target variable at the given values of the input variables.	■Target is a discrete (binary or ordinal) variable. ■Input variables have any measurement level. ■Predicted values are the probability of a particular level(s) of the target variable at the given values of the input variables.

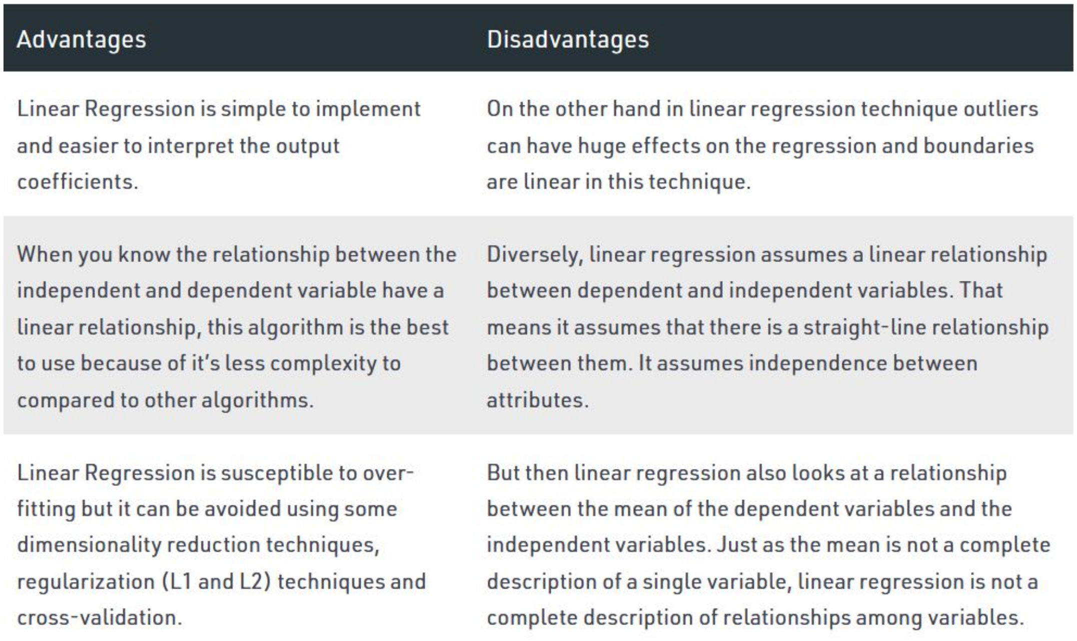
34. Justify the name “Logistic Regression”.

Logistic Regression is one of the basic and popular algorithm to solve a classification problem. It is **named** as '**Logistic Regression**', because it's underlying technique is quite the same as Linear **Regression**. The term “**Logistic**” is taken from the **Logit** function that is used in this method of classification.

35. Why do we use a logistic function in Logistic Regression?

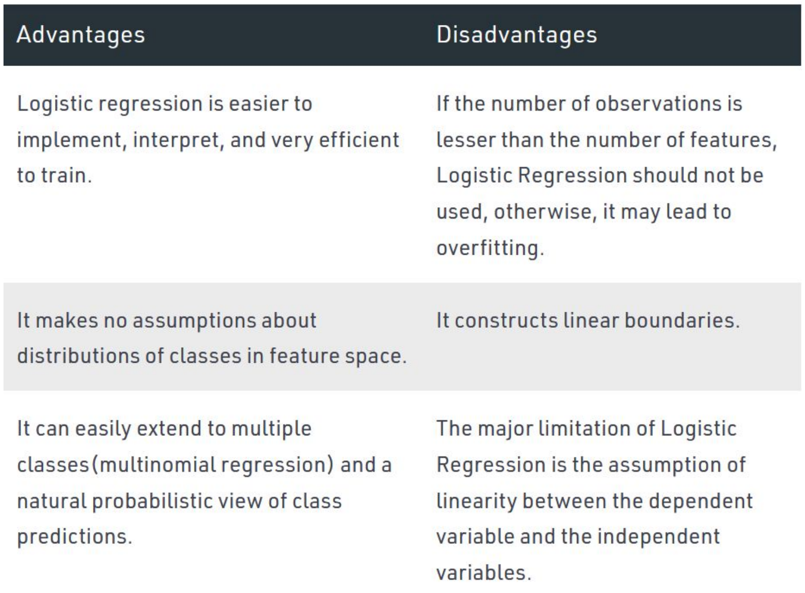
36. Derive the log-loss cost function of Logistic Regression using step by step explanation.

37. What are the advantages and disadvantages/drawbacks of Linear Regression algorithm?

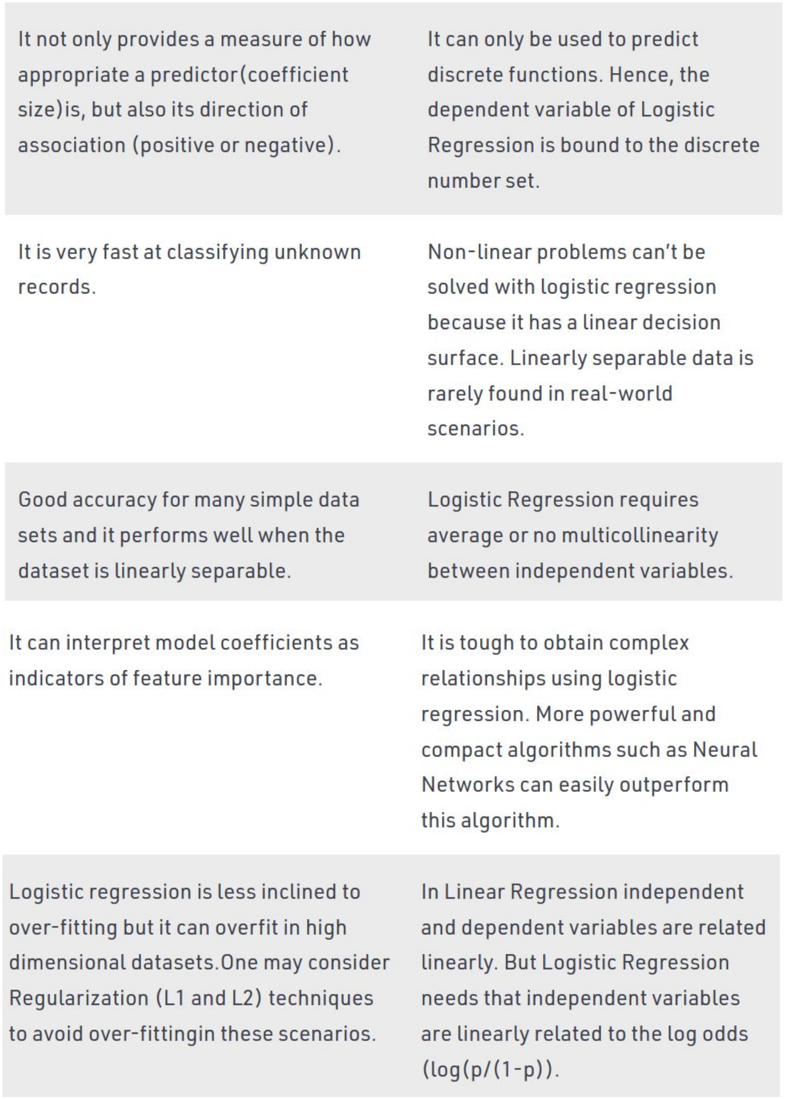


Advantages	Disadvantages
Linear Regression is simple to implement and easier to interpret the output coefficients.	On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity to compared to other algorithms.	Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.	But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

38. What are the merits and demerits of Logistic Regression model if any?



Advantages	Disadvantages
Logistic regression is easier to implement, interpret, and very efficient to train.	If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
It makes no assumptions about distributions of classes in feature space.	It constructs linear boundaries.
It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.	The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.



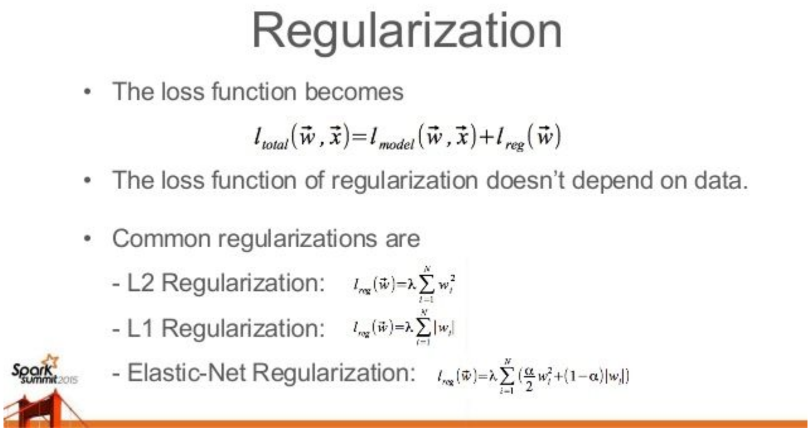
It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).	It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
It is very fast at classifying unknown records.	Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.	Logistic Regression requires average or no multicollinearity between independent variables.
It can interpret model coefficients as indicators of feature importance.	It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.
Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.One may consider Regularization (L1 and L2) techniques to avoid over-fittingin these scenarios.	In Linear Regression independent and dependent variables are related linearly. But Logistic Regression needs that independent variables are linearly related to the log odds (log(p/(1-p)).

39. What is regularization?

Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

The commonly used regularisation techniques are :

1. L1 regularisation
2. L2 regularisation
3. Dropout regularisation



Regularization

- The loss function becomes

$$l_{total}(\vec{w}, \vec{x}) = l_{model}(\vec{w}, \vec{x}) + l_{reg}(\vec{w})$$

- The loss function of regularization doesn't depend on data.
- Common regularizations are
  - L2 Regularization: $l_{reg}(\vec{w}) = \lambda \sum_{i=1}^{N} w_i^2$
  - L1 Regularization: $l_{reg}(\vec{w}) = \lambda \sum_{i=1}^{N} |w_i|$
  - Elastic-Net Regularization: $l_{reg}(\vec{w}) = \lambda \sum_{i=1}^{N} (\frac{\alpha}{2} w_i^2 + (1-\alpha)|w_i|)$

Spark summit 2015

40. Explain different types of regularization in ML using appropriate Examples.

L1 Regularization or Lasso Regularization

L1 Regularization or Lasso Regularization adds a penalty to the error function. The penalty is the sum of the **absolute** values of weights.

$$Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + p\sum_{i=1}^{n}|w_i|)$$

p is the tuning parameter which decides how much we want to penalize the model.

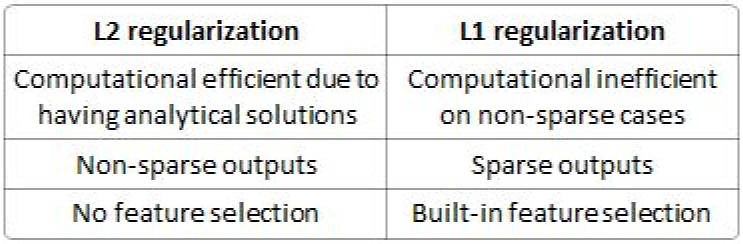
L2 Regularization or Ridge Regularization

L2 Regularization or Ridge Regularization also adds a penalty to the error function. But the penalty here is the sum of the **squared** values of weights.

$$Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + p\sum_{i=1}^{n}(w_i)^2)$$

Similar to L1, in L2 also, p is the tuning parameter which decides how much we want to penalize the model.

41. Differentiate between Lasso (L1) and Ridge (L2) regularization.



L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection

42. What is Elastic-Net regularization?

In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods.The elastic net method overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method which uses a penalty function based on
‖ β ‖ 1 = ∑ j = 1 p | β j |

43. Explain the penalty term used in the Linear Regression for regularization of the model.

44. Explain how the penalty term influences the Logistic Regression for regularization of the model.

45. What is the Stochastic Gradient Descent Algorithm?

In Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. In Gradient Descent, there is a term called “batch” which denotes the total number of samples from a dataset that is used for calculating the gradient for each iteration. In typical Gradient Descent optimization, like Batch Gradient Descent, the batch is taken to be the whole dataset. Although, using the whole dataset is really useful for getting to the minima in a less noisy and less random manner, but the problem arises when our datasets gets big.

SGD ALGORITHM:-

```
$for\ i\ in\ range\,(m):$
$\theta_j = \theta_j - \alpha\,(\hat{y}^i - y^i)\,x_j^i$
```

46. What is the Least Square Method?

The **least squares method** is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. **Least squares** regression is used to predict the behavior of dependent variables.

47. Explain the OLS and Gauss-Markov Theorem.

Ordinary Least Squares regression (OLS) is more commonly named linear regression (simple or multiple depending on the number of explanatory variables).

In the case of a model with p explanatory variables, the OLS regression model writes:

$Y = \beta_0 + \Sigma_{j=1..p} \beta_j X_j + \varepsilon$

where Y is the dependent variable, β_0, is the intercept of the model, X_j corresponds to the j_{th} explanatory variable of the model (j= 1 to p), and e is the random error with expectation 0 and variance σ^2.

In the case where there are n observations, the estimation of the predicted value of the dependent variable Y for the i_{th} observation is given by:

$y_i = \beta_0 + \Sigma_{j=1..p} \beta_j X_{ij}$

The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values. This minimization leads to the following estimators of the parameters of the model:

[$\beta = (X'DX)_{-1} X' Dy$ $\sigma^2 = 1/(W - p^*) \Sigma_{i=1..n} w_i(y_i - y_i)$] where β is the vector of the estimators of the β_i parameters, X is the matrix of the explanatory variables preceded by a vector of 1s, y is the vector of the n observed values of the dependent variable, p* is the number of explanatory variables to which we add 1 if the intercept is not fixed, w_i is the weight of the i_{th} observation, and W is the sum of the w_i weights, and D is a matrix with the w_i weights on its diagonal.

Gauss-Markov:

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.

Gauss Markov Assumptions

There are five Gauss Markov assumptions (also called *conditions*):

1. **Linearity: the parameters we are estimating using the OLS method must be themselves linear.**
2. **Random: our data must have been randomly sampled from the population.**
3. **Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.**
4. **Exogeneity: the regressors aren't correlated with the error term.**
5. **Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.**

Purpose of the Assumptions

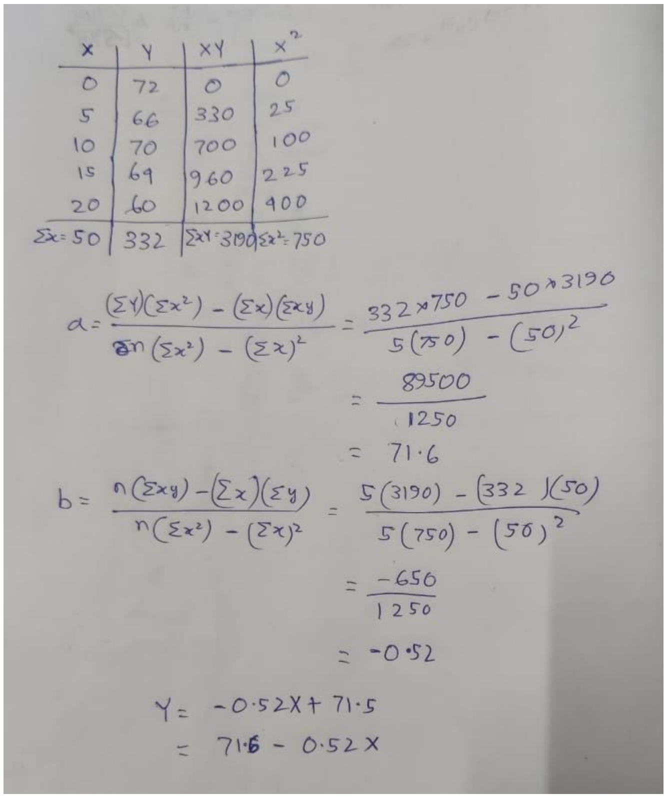
The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

Difference between OLS and GLS: https://stats.stackexchange.com/questions/155031/how-to-determine-if-gls-improves-on-ols

48. Find the equation of linear regression line using following data points: (0,72),(5,66), (10,70), (15,64), (20,60)



X	Y	XY	X^2
0	72	0	0
5	66	330	25
10	70	700	100
15	64	960	225
20	60	1200	400
$\Sigma x = 50$	332	$\Sigma xy = 3190$	$\Sigma x^2 = 750$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{332 \times 750 - 50 \times 3190}{5(750) - (50)^2}$$

$$= \frac{89500}{1250}$$

$$= 71.6$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{5(3190) - (332)(50)}{5(750) - (50)^2}$$

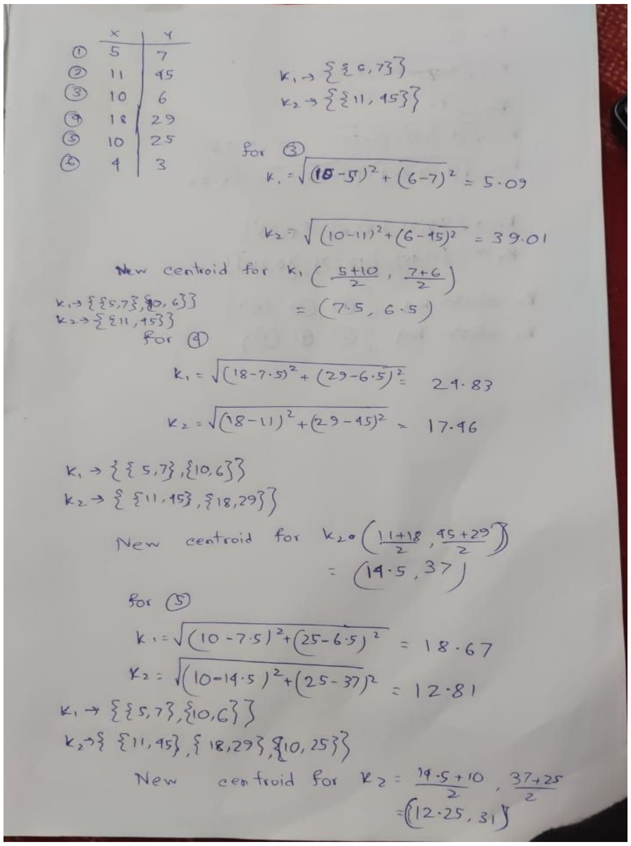
$$= \frac{-650}{1250}$$

$$= -0.52$$

$$Y = -0.52X + 71.5$$

$$= 71.6 - 0.52X$$

49. Using K-mean Clustering Algorithm, Cluster the following data points: (5,7), (11,45),(10, 6), (18,29), (10,25), (4,3); where K=2 and Euclidean Distance.



	x	y
①	5	7
②	11	45
③	10	6
④	18	29
⑤	10	25
⑥	4	3

$k_1 \rightarrow \{\{5,7\}\}$

$k_2 \rightarrow \{\{11,45\}\}$

for ③

$k_1 = \sqrt{(10-5)^2+(6-7)^2} = 5.09$

$k_2 = \sqrt{(10-11)^2+(6-45)^2} = 39.01$

New centroid for $k_1 \left(\frac{5+10}{2}, \frac{7+6}{2}\right)$

$= (7.5, 6.5)$

$k_1 \rightarrow \{\{5,7\},\{10,6\}\}$

$k_2 \rightarrow \{\{11,45\}\}$

for ④

$k_1 = \sqrt{(18-7.5)^2+(29-6.5)^2} = 24.83$

$k_2 = \sqrt{(18-11)^2+(29-45)^2} = 17.46$

$k_1 \rightarrow \{\{5,7\},\{10,6\}\}$

$k_2 \rightarrow \{\{11,45\},\{18,29\}\}$

New centroid for $k_2 \left(\frac{11+18}{2}, \frac{45+29}{2}\right)$

$= (14.5, 37)$

for ⑤

$k_1 = \sqrt{(10-7.5)^2+(25-6.5)^2} = 18.67$

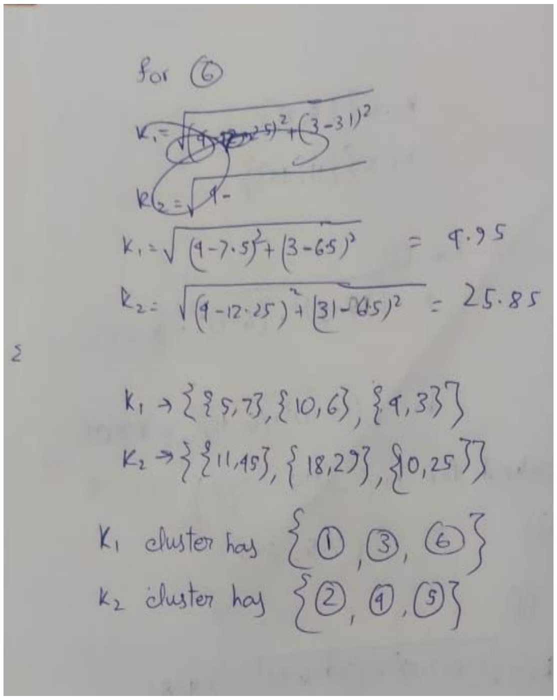
$k_2 = \sqrt{(10-14.5)^2+(25-37)^2} = 12.81$

$k_1 \rightarrow \{\{5,7\},\{10,6\}\}$

$k_2 \rightarrow \{\{11,45\},\{18,29\},\{10,25\}\}$

New centroid for $k_2 = \frac{14.5+10}{2}, \frac{37+25}{2}$

$= (12.25, 31)$



for ⑥

~~$k_1 = \sqrt{(4-12.25)^2+(3-31)^2}$~~

~~$k_2 = \sqrt{4-}$~~

$k_1 = \sqrt{(4-7.5)^2+(3-6.5)^2} = 4.95$

$k_2 = \sqrt{(4-12.25)^2+(31-6.5)^2} = 25.85$

$k_1 \rightarrow \{\{5,7\},\{10,6\},\{4,3\}\}$

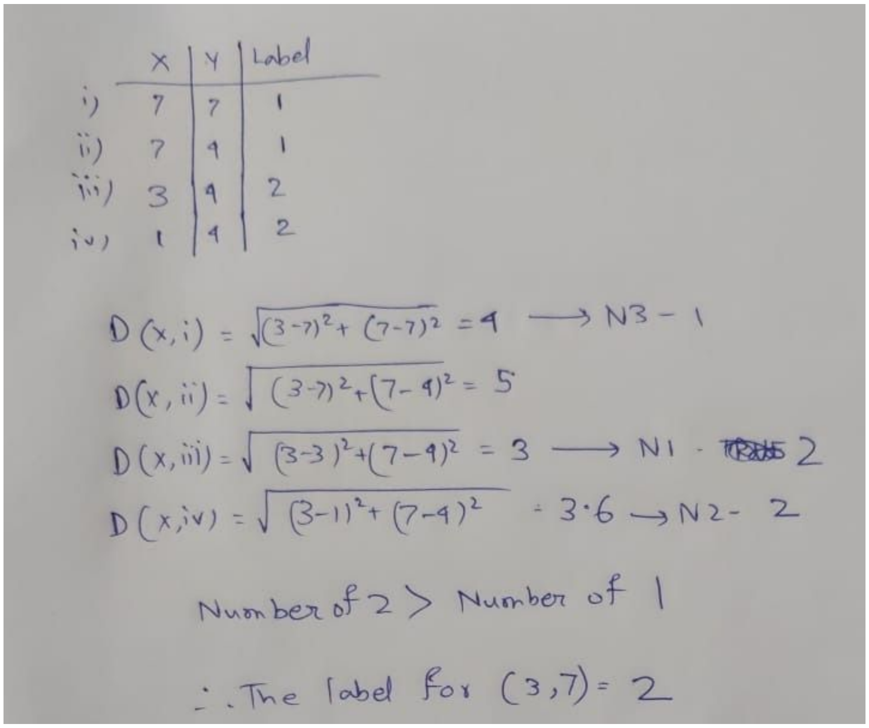
$k_2 \rightarrow \{\{11,45\},\{18,29\},\{10,25\}\}$

k_1 cluster has {①, ③, ⑥}

k_2 cluster has {②, ④, ⑤}

50. Using KNN algorithm and the given data set, predict the label of the test data point (3,7), where K=3 and Euclidean distance.

X	Y	Label
7	7	1
7	4	1
3	4	2
1	4	2



	X	Y	Label
i)	7	7	1
ii)	7	4	1
iii)	3	4	2
iv)	1	4	2

$D(x,i) = \sqrt{(3-7)^2 + (7-7)^2} = 4 \longrightarrow N3 - 1$

$D(x,ii) = \sqrt{(3-7)^2 + (7-4)^2} = 5$

$D(x,iii) = \sqrt{(3-3)^2 + (7-4)^2} = 3 \longrightarrow N1 \cdot 2$

$D(x,iv) = \sqrt{(3-1)^2 + (7-4)^2} = 3{\cdot}6 \longrightarrow N2 - 2$

Number of 2 > Number of 1

$\therefore$ The label for (3,7) = 2