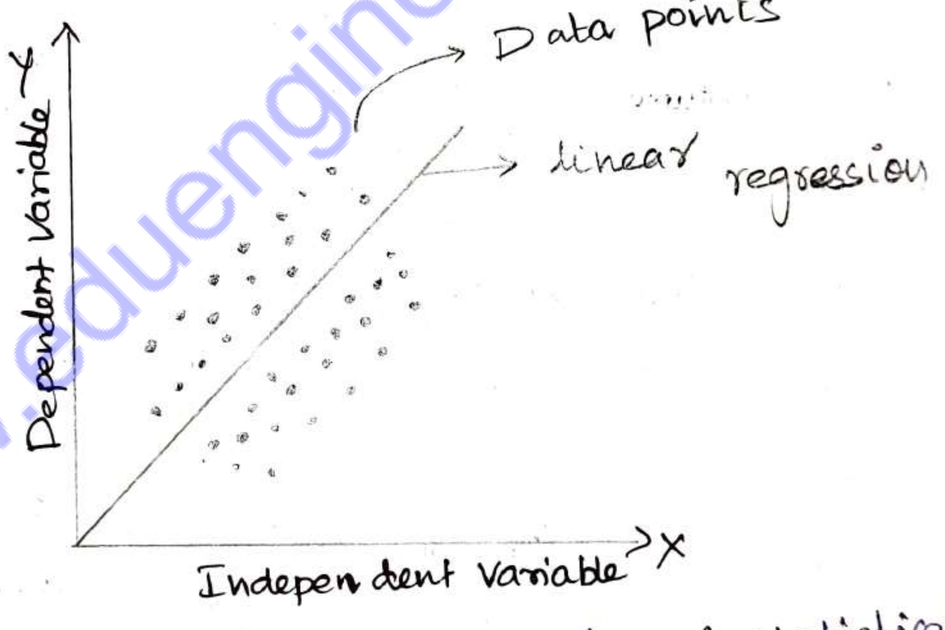
SUPERVISED LEARNING

REGRESSION

Regression find correlations between dependent and independent Variables. If the desired output consists of one or more continous variable then the task is called as regression

Regression algorithm help predict continous variable such as house prices, market trends, weather patterns oil and gas prices ect.



Regression analysis is a set of statistical model or methods used for the estimation relationship between variables and for modelling the future relationship between them

LINEAR REGRESSION MODELS :

* Linear regression is a statistical method that allows us to summarize and study relationship between two continuous quantitative variables.

* The objectives of a linear regression model is to find a relationship between the input variables and a target variable.

* One variable, denoted y is regared as the response, outcome or dependent variable

* The other, denoted x, is regared as the response, predictor explanatory or independent variable.

Regression models predict a continuous variable such as the sales made on the day or predict temparature of a city. Let's imagine that we fit a line with the training point that we have. If we want to add another data point but to fit it, we need to change existing model.

classification predicts categorial lables (classes), prediction model continuous-valued functions. classification is considered to be supervised learning.

REGRESSION LINE :

It gives the average relationship between the two variables in mathematical form.

For two variables x and y, there are always two lines of regression.

Regression line of x on y gives the best estimate for the value of x for any

$$X = a + by$$

Where

$a \rightarrow x$ - intercept

b = slope of the line

x = Dependent Variable

y = Independent Variable.

Regression line y on x

It gives the best estimate for the value of y for any specific give values of x

$$y = a + bx$$

Where

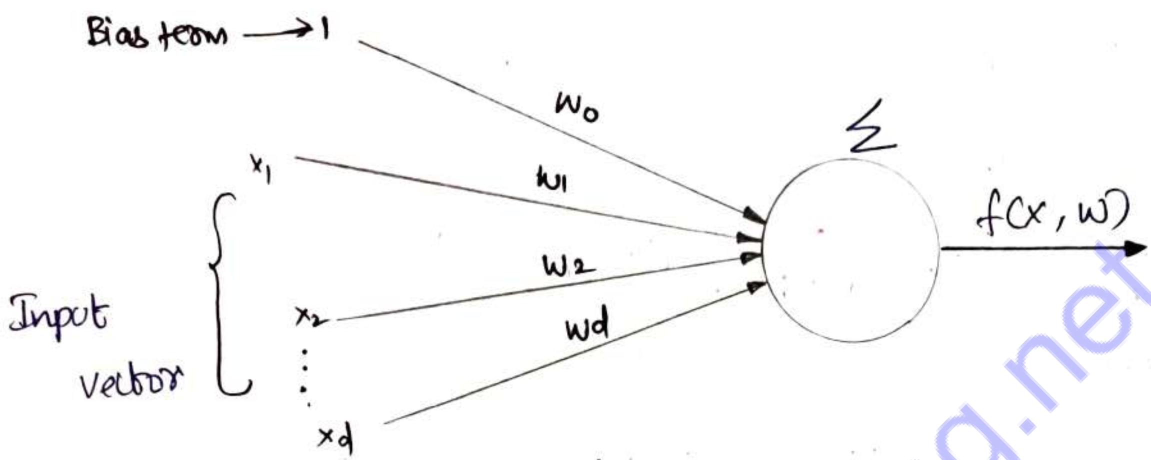
a = y - intercept

b = slope of the line

y = Dependent Variable

x = independen Variable

By using least square method we are able to construct a best fitting to scatter diagram points and then formolate a regression equation in the form of:



$$\hat{y} = a + bx$$

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

Regression analysis is the art and science of fitting straight lines to patterns of data. In linear regression model the variable of interest (dependent variable) is predicted from k other variables (independent variables using linear equation. If y denotes the the dependent variable and $x_1 \ldots x_k$ are the independent variables then the assumption is that the value of y at time t in the data sample is determined by the linear equation.

$$Y_1 = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \varepsilon_t$$

Where the betas are constants and the epsilons are independent and identically distributed normal random variables with mean zero.

At each split point the "error" between the predicted Value and the actual values is Squared to get a "Sum of Squared Errors (SSE)". The Split point errors across the variables are compared and the variable point yielding the lowest SSE is chosen as the root node slipt point. This process is recursevily continued.

Advantages:

Training a linear regression model is usually much faster than method such as neural networks.

Linear regression models are simple and require minimum memory to implement.

LEAST SQUARE:

The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data on the one hand and expected values on the other.

The least square criterion states that the sum of square of error is minimum.

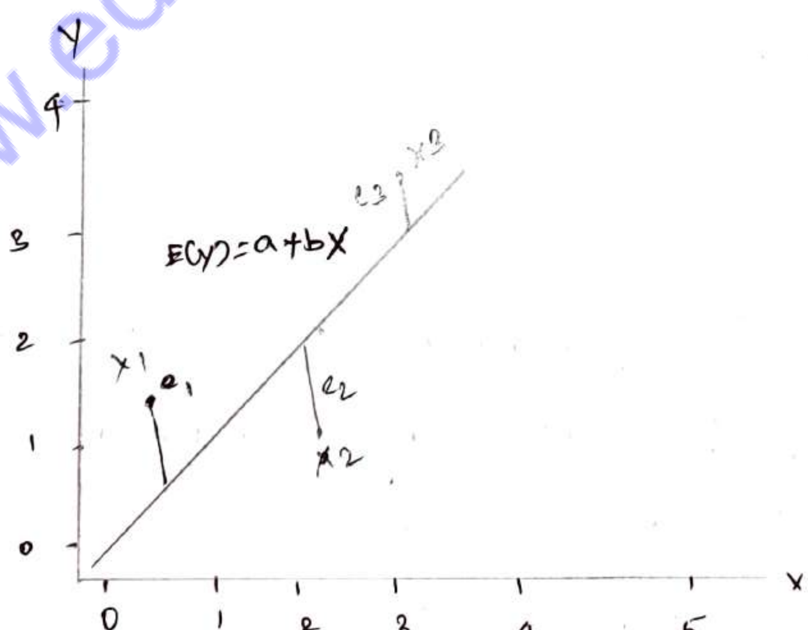
The least square Solutions yields $y(x)$ whose Sum to 1 but do not ensure outputs to be in the range $[0,1]$

How to draw such a line based on the data points observed? Suppose a imaginary line of $y = a + bx$

Imagine a vertical distance between the line and the data point $E = Y - E(Y)$

The error is the deviation of the data point from the imaginary line, regression line. Then what is the best values of a and b?

a & b that minimize sum of such errors.



Deviation does not have good properties for computation. Then why do we use square of deviation?

Let us get a & b that can minimize the sum of squared deviations. This method is called least squares.

Least square method minimizes the sum of squares of errors. Such as a & b are called least square estimators. i.e estimators of parameters α & β

The process of getting parameter estimators (eg. a & b) is called estimation. Least square method is the estimation method of ordinary least squares (OLS).

Disadvantages of least square

* Lack robustness to outliers
* Certain datasets unsuitable for least square classification.
* Decision boundary corresponds to Machine learning solution

MULTIPLE REGRESSION :

Regression analysis is used to predict the value of one or more responses from a set of predictors. It can also be used for estimate the linear association between the predictors and responses. predictors can be continuous or categorical or a mixture of both.

If the multiple independent variables affect the response variable, then the analysis call for a model different from that used for the single predictor valuable. In a situation where more than one independent factor (variable) affects the outcome of a process, a multiple regression model is used. This is referred to as multiple linear regression model or multivariate least square fitting.

$$Y_1 = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_r z_{jr} + \varepsilon_j$$

Where ε is the random error

$\beta_i, i = 0, 1 \ldots r$ are un-known regression co-efficient

Difference between Simple Regression and Multiple Regression

Simple Regression	Multiple Regression
One dependent variable Y predicted from one independent variable X	One dependent variable Y predicted from a set of independent variables $(x_1, x_2 \ldots x_k)$
One regression coefficient	One regression coefficient for each independent variable

BAYESIAN LINEAR REGRESSION:

* Bayesian linear regression allows a useful mechanism to deal with insufficient data or poor distributed data

* It allows user to put a prior on the coefficients and on the noise so that in the absence of data the priors can take over. A prior is a distribution on a parameter.

* If we could flip the coin an infinite number of times, inferring its bias would be easy by the law of large numbers.

* However what if we could only filp the coin a handful of times? Would we guess that a coin is baised if we saw three heads in three flips, an event that happens one out of eight times with unbiased coins? They overfit these data, inferring a coin bias of p=1

* Bayesian methods allows us to estimate model parameters to construct model forecasts and to conduct model comparisons. Bayesian learning algorithms can calculate explicit probabilities for hypotheses.

Bayesian classifiers use a simple idea that the training data are utilized to calculate an observed probability of each class based on feature values.

When the Bayesian classifier is used for unclassified data, it uses the observed probabilities to predict the most likely class for the new features.

Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.

Bayesian methods can accommodate hypotheses that make probabilistic predictions. New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

Uses of Bayesian classifiers:

* Used in text-based classification for finding spam or junk mail filtering
* Medical diagnosis
* Network security such as detecting illegal instruction.

Basic procedure for implementing Bayesian linear Regression :

Specify priors for the model parameters

Create a model mapping the training inputs to the training outputs.

Have a Markov chain Monte Carlo (MCMC) algorithm draw samples from the posterior distributions for the parameters.

Gradient Descent :

Gradient descent is a first-order optimization algorithm. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient of the function at the current point.

Gradient descent is popular for very large scale optimization problems because it is easy to implement, can handle black box functions and each iteration is cheap

The gradient will give the slope of the curve at that x and its direction will point to an increase in the function. so we can change x in the opposite direction to lower the function value.

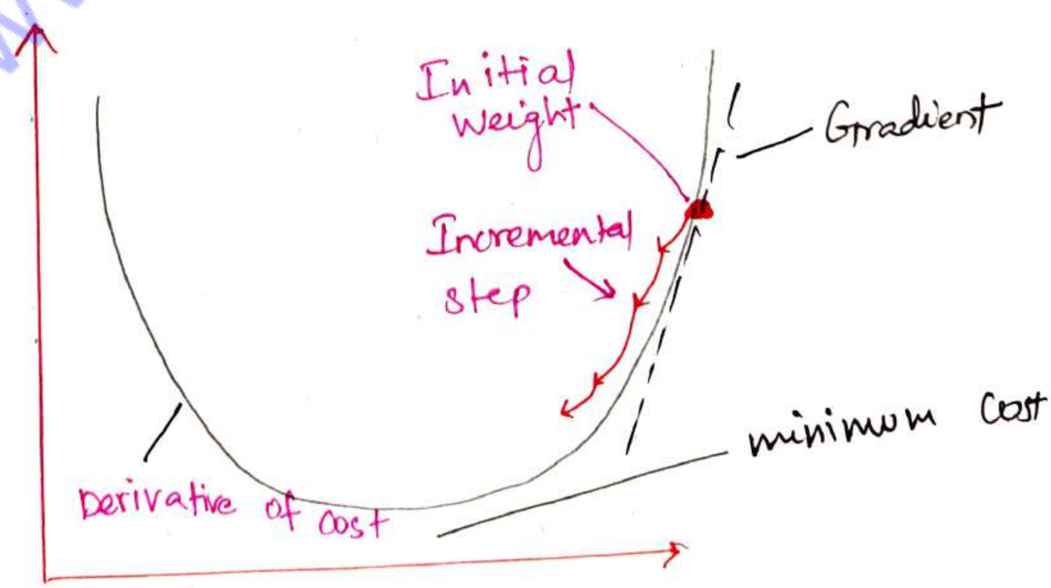
$$x_{k+1} = x_k - \lambda \nabla f(x_k)$$

The $\lambda > 0$ is a small number that forces the algorithm to make small jumps.

Limitation of Gradient Descent

Gradient descent is relatively slow close to the minimum : technically its asymptotic rate of convergence is inferior to many other methods.

For poorly conditioned convex problems gradient descent increasingly 'zigzags' as the gradient points nearly orthogonally to the shortest direction to a minimum point.



If we move towards a negative gradient or away from the gradient of the function at the current point it will give the local minimum of the function.

Whenever we move towards a positive gradient or towards the gradient of the function at the current point, we will get the local maximum of the function.

This entire procedure is known as Gradient ascent which also known as steepest descent. The main objective of using a gradient descent algorithm is to minimize the cost function using iteration.

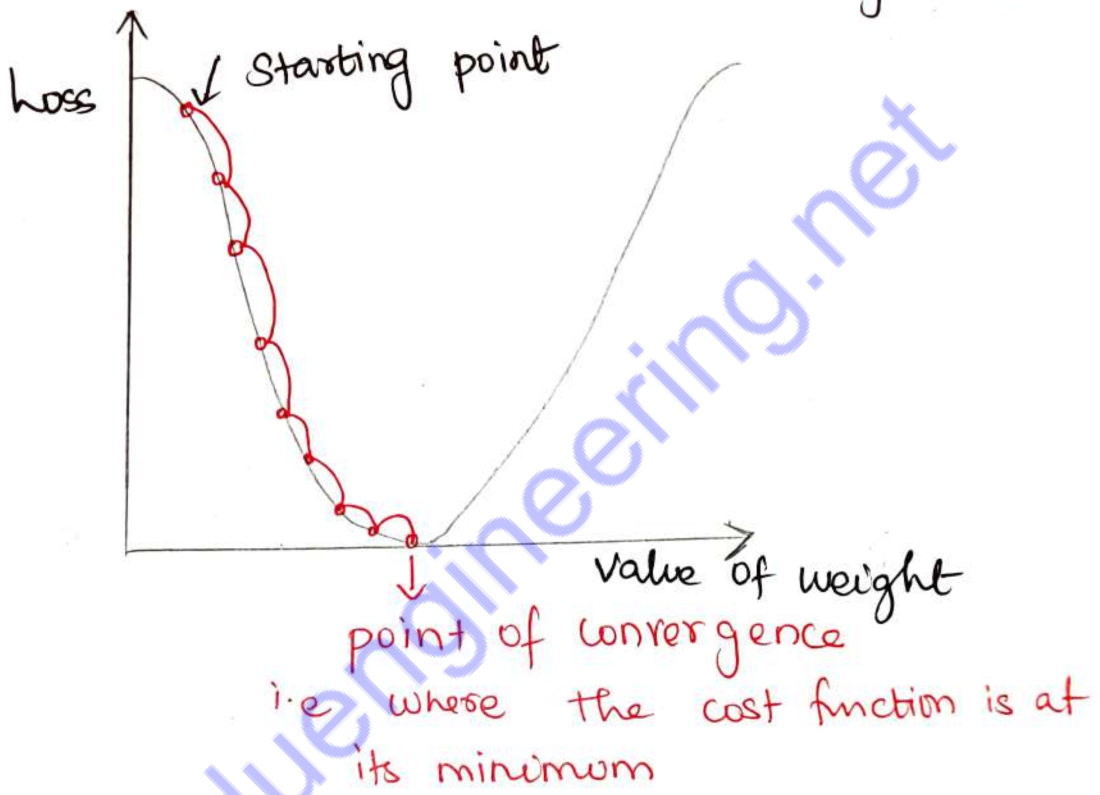
calculate the first order derivative of the function to compute the gradient or slope of the function.

Move away from the direction of the gradient which means slope increased from the current point by alpha times, where alpha is defined as learning Rate. It is a tuning parameter in the optimization process which helps to decide the length of the steps.

Working of Gradient Descent

$$y = mx + c$$

Where m represents the slope of the line and c represents the intercepts on the y-axis



The Slope becomes steeper at the starting point or arbitary point but whenever new parameters are generated then steepness gradually reduces and at the lowest point, it approaches the lowest point which is called a point of convergence.

Learning Rate: It is defined as the step size taken to reach the minimum or lowest point. This is typically a small value that is evaluated and updated based on the behavior of the cost function. If the learning rate is high it results in larger steps but it also leads to risks of overshooting the minimum.

At the same time a low learning rate shows the small step sizes which compromises overall efficiency but gives the advanage of more precision.

Types of Gradient Descent

1. Batch Gradient Descent
2. Stochastic gradient Descent
3. Mini Batch Gradient Descent

Batch Gradient Descent:

It is used to find the error for each point in the training set and update the model after evaluating all training examples. It is known as training epoch.

Stochastic gradient Descent

Stochastic gradient Descent is a type of gradient Descent that runs one training example per iteration. It is more efficient for large datasets.

Mini batch Gradient Descent

Mini batch gradient Descent is the combination of both batch gradient descent and stochastic gradient descent. It divides the training datasets into small batch sizes then performs the updates

LINEAR CLASSIFICATION MODELS :

A classification algorithm that makes its classification based on linear predictor function combining a set of weights with the feature vector.

It does classification decision based on the value of a linear combination of the characteristic Imagine that the linear classifier will merge into its weights all the characteristics that define a particular class.

Discriminative functions :

Linear Discriminant Analysis (LDA) is one of the commonly used dimensionality reduction techniques in machine learning to slove more than two-class classification problems. It is known as Normal Discriminant Analysis (NDA) or Discriminant Function Analysis (DFA)

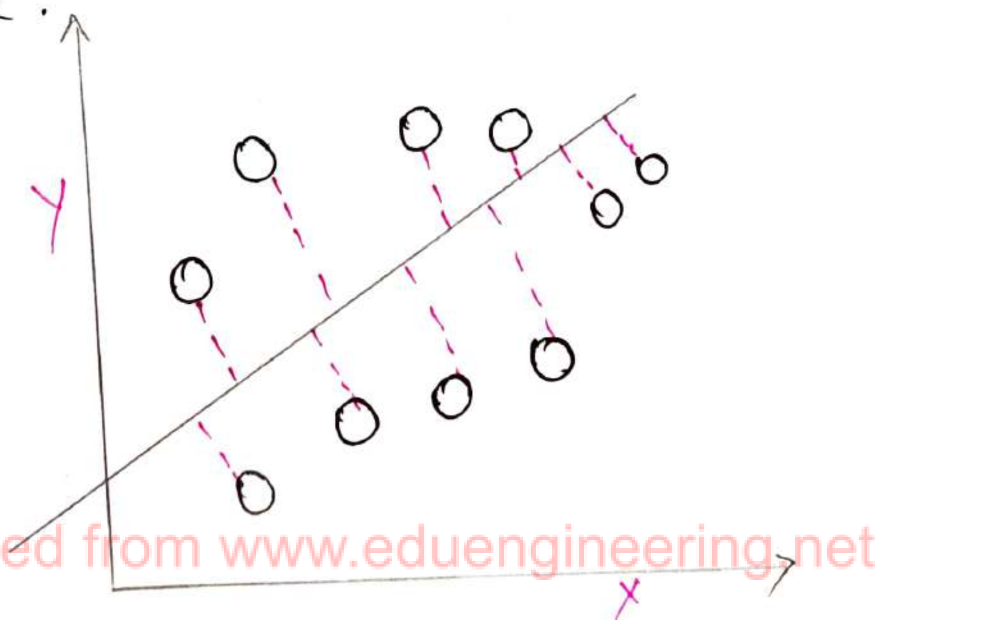
Linear Discriminant analysis is one of the most popular dimensionality reduction techniques used for Supervised classification problems in machine learning. It is also considered a pre-processing step for modeling differences in ML and application of

Linear Discriminant analysis is used as a dimensionality reduction technique in machine learning using which we can easily transform a 2-D and 3-D graph into a one dimensional plane

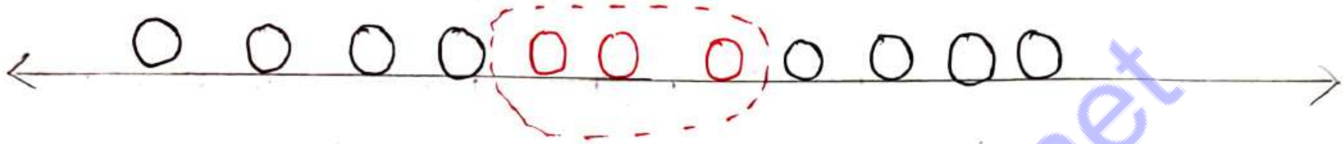
Let's consider an example where we have two classes in a 2-D plane having an x-y axis and we need to classify them efficiently. As we have already seen in the above example that LDA enable us to draw a straight line that can completely separate the two classes of data points.

Here LDA uses an x-y axis to create a new axis by separating them using a straight line and projecting data onto a new axis.

Hence we can maximize the separation between these classes and reduce the 2-D plane into one dimensional.



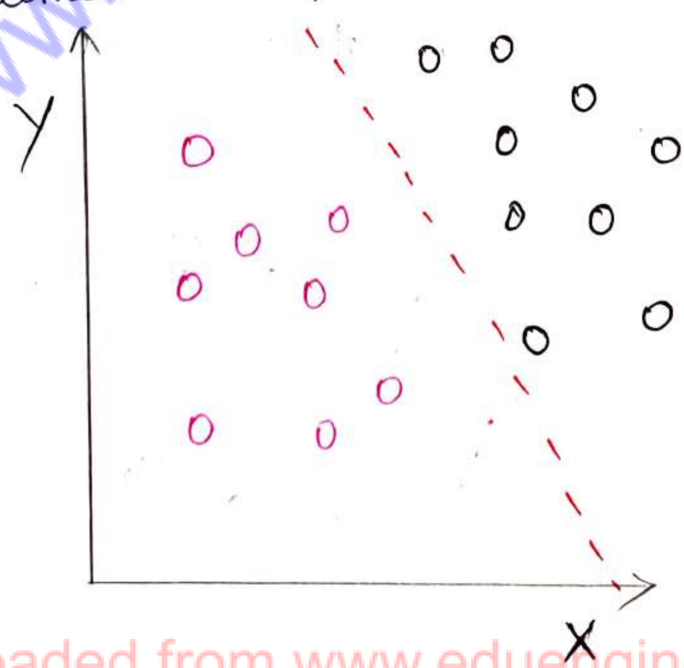
pattern classification. For eg, If we have two classes with multiple features and need to separate them efficiently. When we classify them using a single feature, Then it may show overlapping.



overlapping.

To overcome the overlapping issue in the classification process, we must increase the number of features regularly.

Eg:
Let's assume we have to classify two different classes having two sets of data points" in a 2-dimensional plane



To create a new axis, Linear Discriminant Analysis uses the following criteria:

⇒ It maximizes the distance between means of two classes.

⇒ It minimizes the variance within the individual class.

In other words we can say that the new axis will increase the separation between the data points of two classes and plot them onto the new axis.

LOGISTIC REGRESSION:

* Logistic regression is one of the most popular Machine learning algorithms, which comes under the Supervised learning technique.

* It is used for predicting the categorical dependent variable using a given set of Independent variables.

* Therefore the outcome must be a categorical or discrete value. It can be either yes or No, 0 or 1, true or flase ect. but instead of giving the exact value as 0 & 1 it gives the probabilistic values which lies between 0 & 1

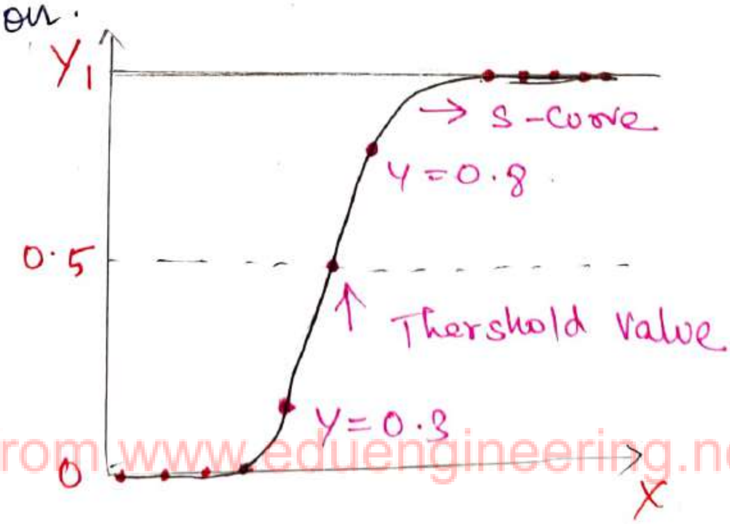
* Logistic regression is used for sloving the classification problems.

* In logistic regression, instead of fitting a regression line, we fit an 'S' shaped logistic function, which predicts two maximum values (0 and 1)

* In logistic regression instead of fitting a regression line, we fit an "S" shaped logistic function which predicts ~~two~~ two maximum values (0 and 1)

It is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets

It can be used to classify the observation using different types of data and can easily determine the most effective variables used for the classification.



LOGISTIC FUNCTION :

* The sigmoid function or logistic function used to map the predicted values to probabilities

* It maps any real value into another value within a range of 0 and 1.

* The value of the logistic regression must be between 0 and 1 which cannot go beyond this limit so it forms a curve like the "S" form.

* The dependent variable should be categorical in nature

* The independent variable should not have multi-collinearity.

Logistic Regression Equation :

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n$$

In logistic Regression y can be between 0 and 1 only so for this lets divide the above equation by $(1-y)$:

$$\frac{y}{1-y} \text{ ; } 0 \text{ for } y=0 \text{ and infinity for } y=1$$

* But we need range between -[infinity] to +[infinity] then take logarithm of the equation it will become.

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots b_nx_n$$

Type of Logistic Regression

Binomial:

In this regression there can be only two possible types of the dependent variables such as 0 or 1 pass or fail ect.

Multinomial:

In this regression there can be 3 or more possible unordered types of the dependent variable such as "cat" or "dog" or "sheep".

Ordinal: In this regression there can be 3 or more possible ordered types of dependent variables such as "low", "medium", or "high".

GENARATIVE MODEL:

* Generative models are a class of statistical models that generate new data instances.

* These models are used in unsupervised machine learning to perform tasks such as probability and likelihood estimation, modelling data points and distinguishing between classes using these probabilities.

* Generative models rely on the Bayes theorem to find the joint probability.

* Generative model describe how data is generated using probabilistic models.

* They predict $P(y/x)$, the probability of y given x, calculating the $P(x,y)$, the probability of x and y.

NAVIE BAYES CLASSIFIER:

* Navie Bayes algorithm is a supervised learning algorithm which is based on "Bayes theorem" and used for solving classification problems.

* It is mainly used in text classification that includes high dimensional training dataset.

It helps in building the fast machine learning algorithms or models that can make quick predictions.

* It is a probabilistic classifier which means it predicts on the basis of the probability of an object.

* Some popular examples of Naive Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

* It is called Naive because of it assumes that the occurence of a certain feature is independent of the occurence of other features. Such as if the fruit is identified on the bases of color, shape and taste, then red, spherical and sweet fruit is recognized as an apple.

* Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes Theorem:

It is also know as Bayes Rule or Bayes law which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes theorem is given as

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$

Where

P(A/B) is a posterior probability, probability of hypothesis A on the observed event B

P(B/A) is a likelihood probability, probability of the evidence given that the probability of the hypothesis is true

P(A) is a prior probability, probability of hypothesis before evidence.

P(B) is Marginal probability, probability of evidence.

Difference between generative and Discriminative Model

Generative model	Discriminative model
It generates new data	This discriminate between different kind of data instances
Generative model revolves around the distribution of a dataset to return a probability for a given example.	It makes predictions based on conditional probability and is either used for classification or regression

* Generative model capture the joint probability P(X, Y) or just p(X) if there are no labels.

* A generative model includes the distribution of the data itself, and tells you how likely a given example is

* Generative models are used in unsupervised machine learning to perform task such as probability a likelihood estimation.

* Eg: Gussains, Nawie Bayes

* Discriminative models capture the conditional probability p(Y/X)

* A discriminative model ignores the question of whether a given instance is likely and just tells you how likely a lable is apply to the instance.

* This model particularly used for Supervised learning.

* Eg: logistic Regression SVM

SUPPORT VECTOR MACHINE:

* Support Vector Machines (SVMs) are a set of supervised learning methods which learn from the dataset and used for classification.

* SVM is a classifier derived from statistical learning Theory by Vapnik and Chervonenkis.

Simply speaking we can think of an SVM model as representing the examples as points in space mapped so that each of examples of the separate classes are dividied by a gap that is wide as possible.

Example of Bad Decision Boundaries

SVM are primarily two-class classifiers with the distinct characteristic that they aim to find the optimal hyperplane such that the expected generalization error is minimized.

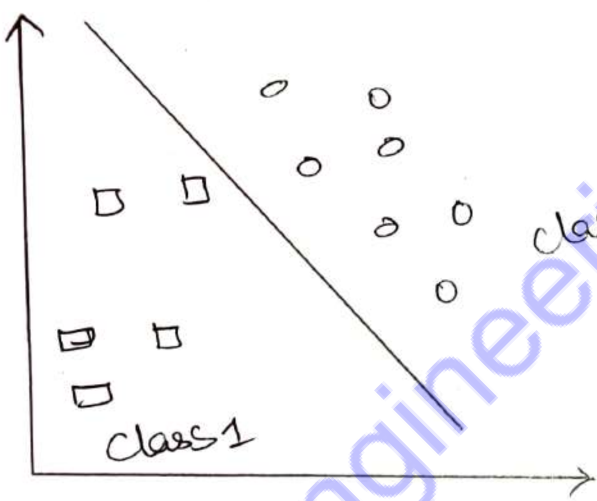
Instead of directly minimizing the empirical risk calculated from the training data SVMs perform structural risk minimization to achieve good generalization.

The empirical risk is the average loss of an estimator for a finite set of data drawn from p.

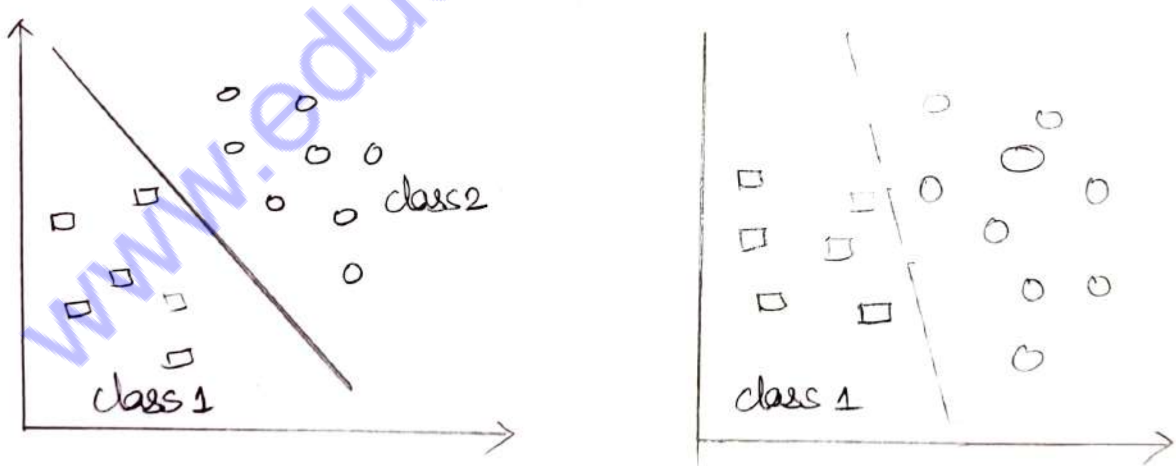
The idea of risk mimization is not only measure the performance of an estimator by its risk, but to actually search for the estimator that minimizes is over distribution p.

* It is a kind of large Margin classifier.

* It is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far away any point in training data.



Two class problem



Bad decision boundary of SVM

Given a set of training examples, each marked as belonging to one of two classes an SVM algorithm builds a model that predicts whether a new example falls into one class to the other

Because we dont know distribution p we instead minimize empirical risk over a training dataset drawn from p. This general learning technique is called empirical risk minimization.

Good Decision Boundary :

The decision boundary should be as far away from the data of both classes as possible.

If the data points lie very close to the boundary, the classifier may be consistent but is more likely to make errors on new instances from the distribution.

Hence we prefer classifiers that maximize the minimal distance of data points to the separator.

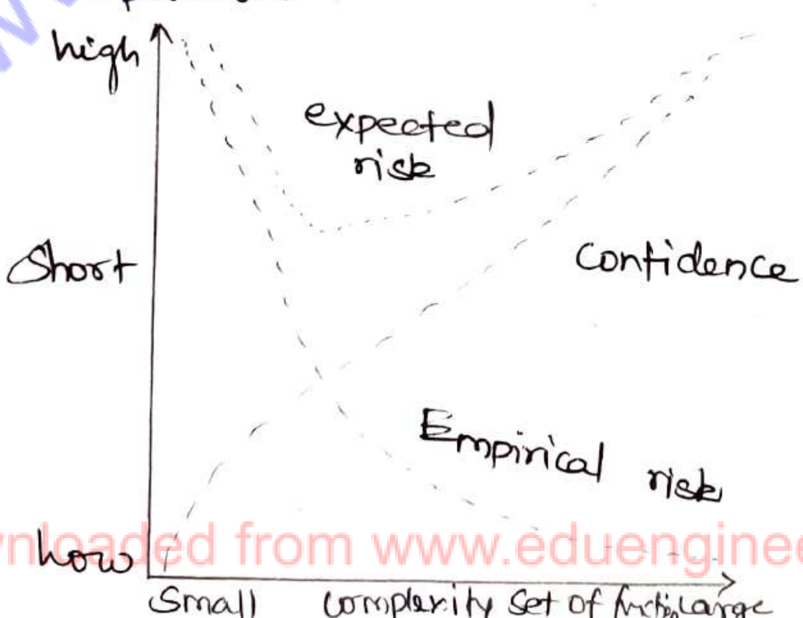
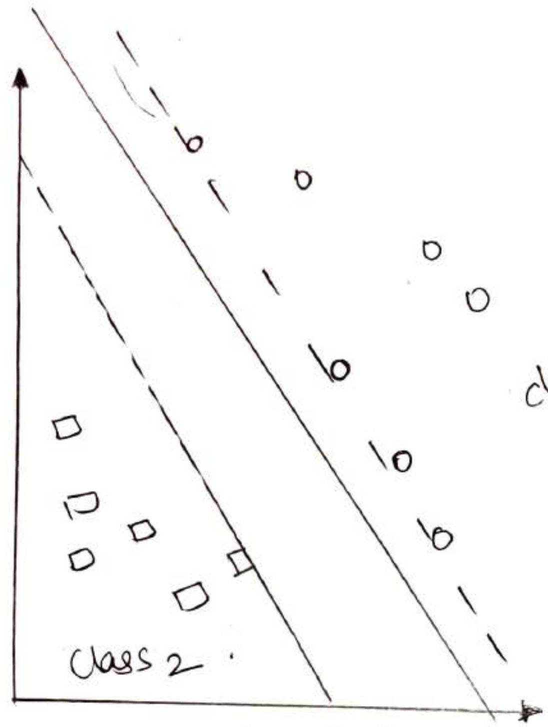


fig: Empirical Risk



Good decision bondary.

* The gap between data points and the Classifier bondary.

* The margin is the minimum distance of any Sample to the decision bondary.

* Margin of the Separator is distance between Support Vectors.

$$\text{Margin } (m) = \frac{2}{\|w\|}$$

* Maximal margin classifier is a classifier in the family F that maximizes the margin. Maximizing the margin is good according to intuition and PAC theory.

* Implies that only Support vectors matter other training examples are ignorable.

Key properties of SVM :

* Use a single hyperplane which subdivides the space into two half spaces one which is occupied by class 1 and the other by class 2.

* They maximize the margin of the decision boundary using quadratic optimization techniques, which find the optimal hyperplane

* Ability to handle large feature spaces

* Overfitting can be controlled by soft margin apporach

SVM Applications

SVM has been used successfully in many real word problems.

* Text (and hypertext) categorization.

* Image classification

* Bioinformatics (protein classification, cancer classification)

* Hand written character recognition

* Determination of SPAM email

Limitations of SVM :

* It is sensitive to noise.
* The biggest limitation of SVM lies in the choice of kernel.
* Another limitations is speed and size
* The optimal design for multiclass SVM classifier is also a research area.

SOFT MARGIN

For the very high dimensional problems common in text classification, sometimes ~~the~~ the data are linearly separable.

But in the general case they are not and even if they are, we might prefer a solution that better separates the bulk of the data while ignoring a few weird noise documents.

What if the training set is not linearly separable? slack variables can be added to allow misclassification of difficult or noisy examples resulting margin called soft.

A soft margin allows a few variables to cross into the margin or over the hyperplane.

DECISION TREE

* A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning

* In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions

* Each leaf node has a class label, determined by majority vote of training examples reaching that leaf.

* Each terminal node is a question on features. It branches out according to the answers.

* Decision tree learning is a method for approximating discrete valued target functions The learned function is represented by decision tree.

* A learned decision tree can also be re-represented as a set of if-then rules

* Decision tree learning is one of the most widely used and practical methods for inductive inference.

* It is robust to noisy data and capable of learning disjunctive expression.

* Decision tree learning method searches a completely expressive hypothesis

DECISION TREE REPRESENTATION

* Build a decision tree for classifying examples as positive & negative instance of concept.

* Each non leaf node has associated with it an attribute.

* Each leaf node has associated with it a classification (+ or −)

* Each arc node has associated with it one of the possible values of the attribute at the node from which the arc is directed

* Internal node denotes a test on an attribute. Branch represents an outcome of the test. Leaf node represents class labels or class distribution

A decision tree is a flow chart like structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distributions.

DECISION TREE ALGORITHM

To generate decision tree from the training tuple of data partition D

Input

Data partition (D)
Attribute List
Attribute Selection method

ALGORITHM :

⇒ Create a node (N)

⇒ If tuples in D are all same class then

⇒ Return node (N) as a leaf node labeled with the class C

⇒ If attribute list is empty then return N as a leaf node labeled with the majority class in D

⇒ Apply attribute selection method (D, attribute list) to find the 'best' splitting criterion.

⇒ Label node N with splitting attribute

⇒ If splitting attribute is discrete valued and multiway splits allowed.

⇒ Then attribute list → attribute list → splitting attribute

⇒ For (each outcome j of splitting criterion)

⇒ Let D_j be the set of data tuples in D satisfying outcome j

⇒ If D_j is empty then attach a leaf labeled with majority class in D to Node N;

⇒ Else attach the node returned by Generate Decision tree (D_j, attribute list) to Node N;

⇒ End of for loop

⇒ Return N

Decision tree generation consists of two phases one is tree construction & pruning.

In tree construction phase, all the training examples are at the root. partition examples recursively based on selected attributes

In tree purning phase, the identification & removal of branches that reflect noise or outliers

Advantages:

* Rules are simple and easy to understand
* Decision tree can handle both nominal & numerical attributes
* Decision tree are capable of handling datasets that may have errors.
* It has capable of handling datasets that may have missing values.
* Decision trees are considered to be a nonparametric method.
* Decision tree are self-explantory.

Disadvantages:

* Most of the algorithms require that the target attribute will have only discrete values
* Some problem are difficult to solve like XOR.
* Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.

RANDOM FORESTS :

Random forests is a famous system learning set of rules that belongs to the supervised getting to known method.

It may be used for both classification and regression issues in ML

It is based totally on the concept of ensemble studying that's a process of combining multiple classifiers to solve a complex problem and to enhance the overall performance of the model.

Random forest is a classifier that incorporates some of choice timber on diverse subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Random Forests Algorithm Working

Random forest works in two section first is to create the random woodland by combining N selection trees and second is to make predictions for each tree created inside the first segment.

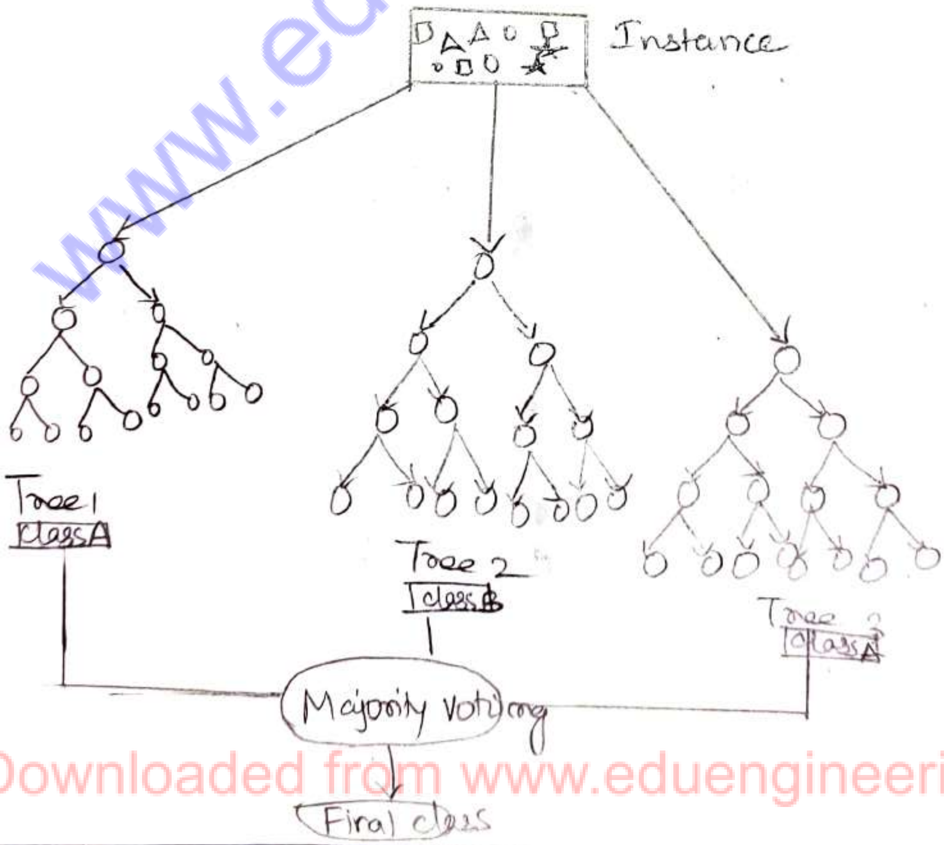
1 ⇒ Select random K statistics points from the Schooling set

2 ⇒ Build the selection trees associated with the selected information points.

3 ⇒ choose wide variety of N for selection trees associated with the selected information which we want to build.

4 ⇒ Repeat step 1 and 2

5 ⇒ For new factors locate the predictions of each choice tree and assign the new record factors to the category that wins most people votes.



Application of Random forest :

* Banking ⇒ Banking zone in general uses This algorithm for the identification of loan dangers

* Medicine : With the assistance of this set of roles, disorder traits and risks of the disorder may be recognized.

* Land use : We can perceive the areas of comparable land use with the aid of this algorithm.

* Marketing : Marketing tendencies can be recognized by the usage of this algorithm.

Advantages of Random Forest :

* It is capable of managing large database with high dimensionality.

* It enhances the accuracy of the version and forestalls the overfitting trouble.

Disadvantage :

* Although random forest can be used for both class + regression ~~respon~~ responsibilities it isn't extra appropriate for regression obligations.