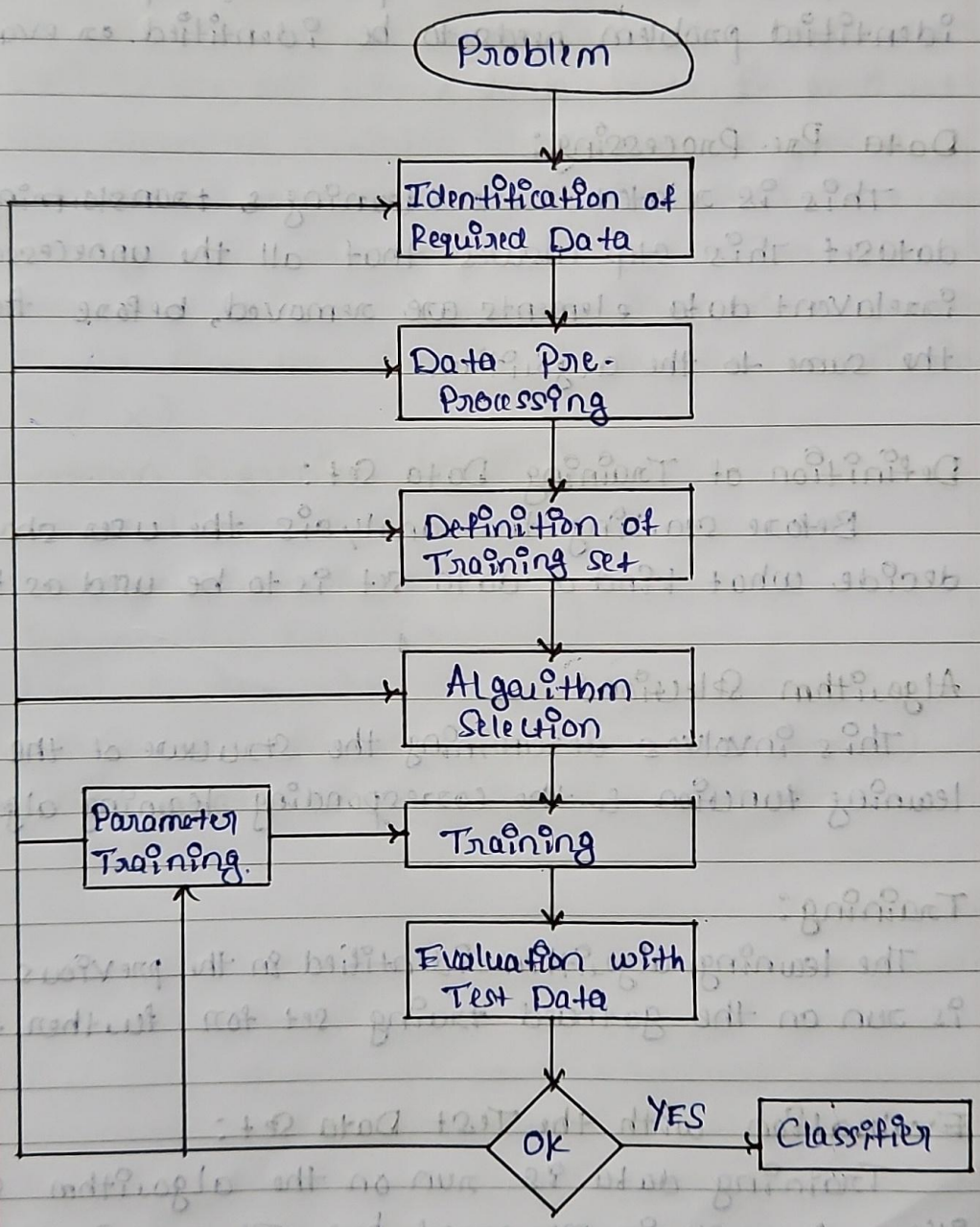
Chapter 3

classmate
Date 09/04/2023
Page

Steps in Classification Learning:

Classification Learning is a type of Supervised Learning in ML where the algorithm learns to predict the class or category of a given input based on labelled training data.

For example in email spam detection, the algorithm is trained on a labeled dataset where each email is labeled as spam or not spam based on its content & other features.



1 Problem identification:

Identifying the problem is the first step in the supervised learning model. The problem needs to be well-formed problem, i.e. a problem with well defined goals & benefits.

2 Identification of Required Data:

On the basis of problem identified above, the required data set that precisely represents the identified problem needs to be identified or evaluated.

3. Data Pre-Processing:

This is related to the cleaning & transforming the dataset. This step ensures that all the unnecessary & irrelevant data elements are removed, before feeding the same to the algorithm.

4 Definition of Training Data Set:

Before starting the analysis the user should decide what kind of data set is to be used as training set.

5 Algorithm Selection:

This involves determining the structure of the learning function & the corresponding learning algorithm.

6 Training:

The learning algorithm identified in the previous step is run on the gathered traing set for further fine tuning.

7 Evaluation with the Test Data Set:

Training data is run on the algorithm & it's performance is measured here. If a suitable result is not obtained further training of parameters may be required.

Q. Explain the Concept of Linear Regression:

Regression:

Regression is a type of supervised learning in ML where the algorithm learns to predict a continuous-output variable based on the input data. In Regression, the input data is a set of independent variables or predictors & the output variable is a continuous dependent variable.

Many problems related to prediction of numerical value can be solved using regression model. In the context of regression, the dependent variable (Y) is the one whose value is to be predicted. And the dependent variable depends on independent variable(s) or predictor(s). Regression is essentially finding a relationship or association b/w the dependent variable Y & the independent variable(s) X i.e find the function 'f' for the association $Y = f(X)$.

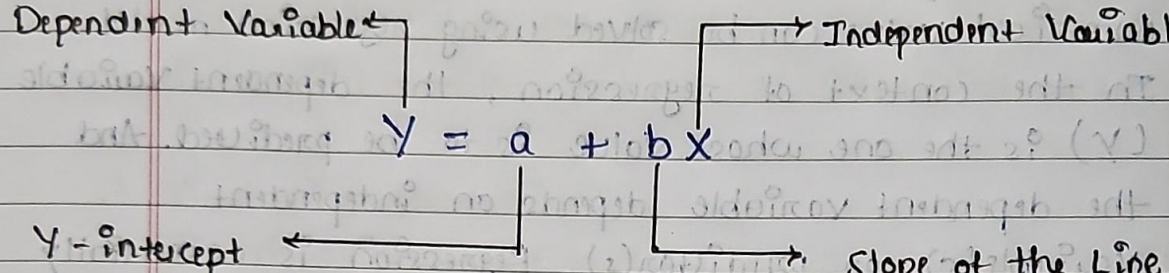
Common Regression Algorithms:

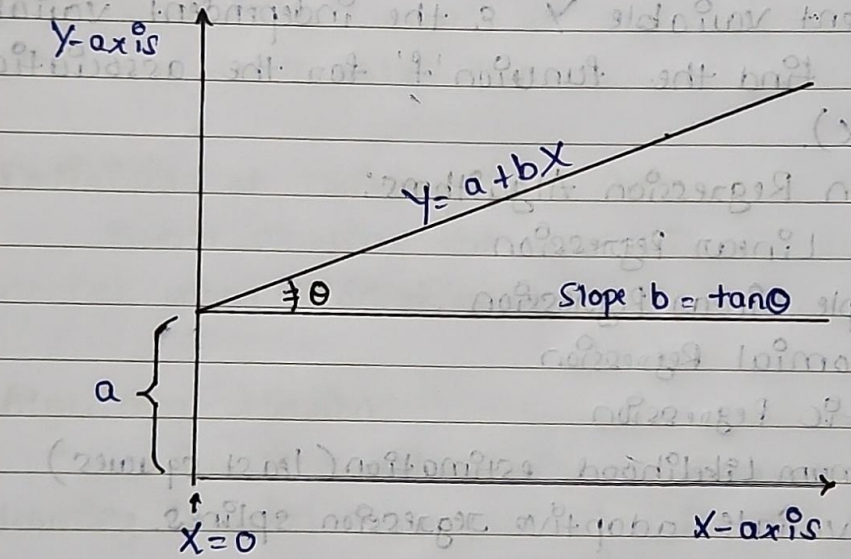
1. Simple Linear Regression
2. Multiple Linear Regression
3. Polynomial Regression
4. Logistic Regression
5. Maximum Likelihood estimation (least squares)
6. Multivariate adaptive regression splines.

Linear Regression:

As the name indicates, Simple Linear Regression is the simplest regression model which involves only one independent variable or predictor & one dependent variable or the Response variable.

This model assumes a Linear Relationship blw the dependent variable & the predictor or independent variable as shown below,





For Example:

If we take price of a property as the dependent variable & the Area of the property in Sq.m. as independent or predictor variable, then we can build a model using simple linear Regression

Price property = f (Area property).

Error in Simple Regression:

$$Y = (a + bX) + \varepsilon.$$

Assuming a linear association, we can reformulate the model,

Price Property = a + b. Areaproperty.

where, a & b are intercept & slope of the straight line respectively.

Straight lines can be defined in a slope-intercept form $Y = a + bX$,

The value of intercept indicates the value of Y when $X = 0$. It is known as 'the intercept' or 'Y intercept' because it specifies where the straight line crosses the vertical or Y-axis.

Multiple Linear Regression:

It aims to model the relationship between one dependent variable & one or more independent or predictor variables.

In the context of Simple Linear Regression we considered the Price of a property as the dependent variable & the Area of the property as the predictor variable. However, location, floor, num of years since purchase, etc are also important predictors which should not be ignored. Thus if we consider Price of property as dependent we can form the equation as;

$Price_{property} = f(Area_{property}, location, floor, Ageing)$

Hence,

Multiple Linear Regression with 'n' independent or predictor variable is,

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$$

Logistic Regression:

Logistic Regression is both classification & Regression technique depending on the scenario used.

It is a type of regression analysis used for predicting the outcome of a categorical dependent variable.

It is used to model the probability of a binary outcome based on one or more predictor variables. So here the dependent variable Y is binary 0 or 1 & independent variable X are continuous in nature.

The goal of logistic regression is to predict the likelihood that Y is equal to 1, i.e Probability that Y=1 rather than 0, with given certain values of X.

The logistic regression model works by transforming the Linear Regression equation using a logistic Function, which maps any value b/w $-\infty$ & $+\infty$ to a range b/w 0 & 1. The Logistic Function used in Logistic Regression is called as Sigmoid. & it's equation is as follows.

Natural Logarithm

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\ln\left(\frac{P}{1} - P\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

The logistic formula are stated in terms of the probability that Y=1 which is refered to as P. The probability that Y=0 is 1-P.

8. Explain the Types of Supervised Leaning algorithms with example?

1 Classification Algorithm :

In ML, classification Algorithm is a type of supervised leaning algorithm used to predict the class or category of a given input data point. It is a type of Algorithm that maps input data points to their corresponding output classes or categories.

The goal of a classification Algorithm is to create a model that can accurately classify a new & unseen data based on the patterns learned from the training data.

Types of Classification Algorithm:

1 K-Nearest Neighbour (KNN)
2 Decision Tree
3 Random Forest
4 Support Vector Machine (SVM)
5 Naive Bayes
6 Logistic Regression
7. Neural Networks

2. Regression Algorithm :

1 Linear Regression
2 Multiple Linear Regression
3 Polynomial Regression
4 Logistic Regression
5 Decision Tree Regression
6 Random Forest Regression
7 Support Vector Regression (SVR)
8 Bayesian Regression
9. Gaussian Process Regression

5. Explain the concept of KNN with an Example

KNN:

The KNN algorithm is a simple but extremely powerful classification algorithm. The name of the algorithm originates from the underlying philosophy of KNN i.e people having similar background or mind set tend to stay close to each other.

In the same way, as a part of KNN algorithm, the unknown & unlabelled data which comes for a prediction problem is judged on the basis of the training data set which are similar to the unknown element.

Working of KNN:

In KNN algorithm, the classification of new data point is based on the class of it's K nearest Neighbors. The value of K is choosen by the user & it determines the num of Neighbors that will be considered. The class of a new data point is assigned by taking a majority vote of the classes of its K Nearest Neighbors.

For example:

Let us consider a very simple student data set. It consist of 15 students studying in a class. Each of the students has been assigned a score on a scale of 10 on 2 performance parameter - 'Aptitude' & 'Communication'. Also a class value is assigned to each student based on the follow criteria.

1. Students having a good communication skills as well as good level of aptitude have been classified as 'Leader'
2. Students " " good communication skills but not so good level of aptitute " " 'Speaker'
3. Students " " not so good communication skill but a good level of aptitude have been " " 'Intel'

	Name	Aptitute	Communication	Class
Training Data	Sunil	2	5	Speaker
	Abhi	2	6	Speaker
	Varsha	8	10	Leader
Test Data →	Ram	7	3	Intel

When building a classification model, a part of the labelled data is retained as test data. the remaining portion of the input data is used to train the model - hence known as: Training data.

The motivation to retain a part of the data as test data is to evaluate the performance of the model. The performance of the classification model is measured by the number of correct classifications made by the model when applied to the unknown data set.

2 challenges in KNN:

1. What is the basis of this similarity or when we can say that 2 data elements are similar?

There are many measures of similarity the most common approach adopted by KNN to measure the similarity b/w 2 data elements is **Euclidean Distance**

Considering very simple data set having 2 features f_1 & f_2

The Euclidean Distance b/w 2 data elements d_1 & d_2 can be measured by

$$\text{Euclidian distance} = \sqrt{(f_{11} - f_{12})^2 + (f_{21} - f_{22})^2}$$

f_{11} is the value of feature f_1 for data ele d_1

f_{12} " " f_1 " d_2

f_{21} " " f_2 " d_1

f_{22} " " f_2 " d_2

2. How many similar elements should be considered for deciding the class label of each test data element?

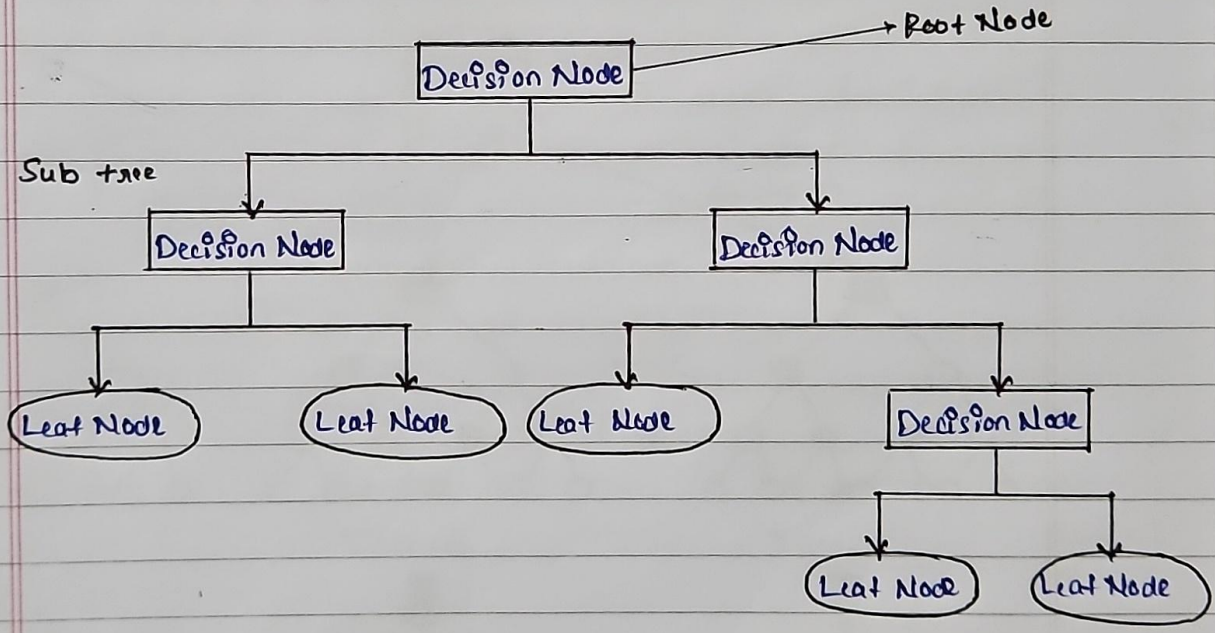
6. Explain the Concept of Decision Trees with an example?

Decision Tree:

Decision Tree is a type of supervised Learning algorithm that is used for classification & regression Tasks. Decision Trees are type of model that uses a tree-like structure to make decision based on input features.

The basic idea of decision tree is to split the data set into smaller & smaller subsets based on the features of the data until a decision can be made. This process is repeated recursively for each subset until the algorithm has generated a tree that can accurately classify the new data.

A Decision tree is usually represented in the format below,



For Example:

Consider an example for Decision Tree of Car Driving. The Decision to taken is whether to 'Keep Going' or to 'Stop' which depends on various situations as dipicted in the figure.

If the signal is RED in color, then car should be stopped. If there is not enough gas in the car, the car should be stopped at the next available gas station.

7. Explain the Concept of Random Forest with a neat diagram?

Random Forest:

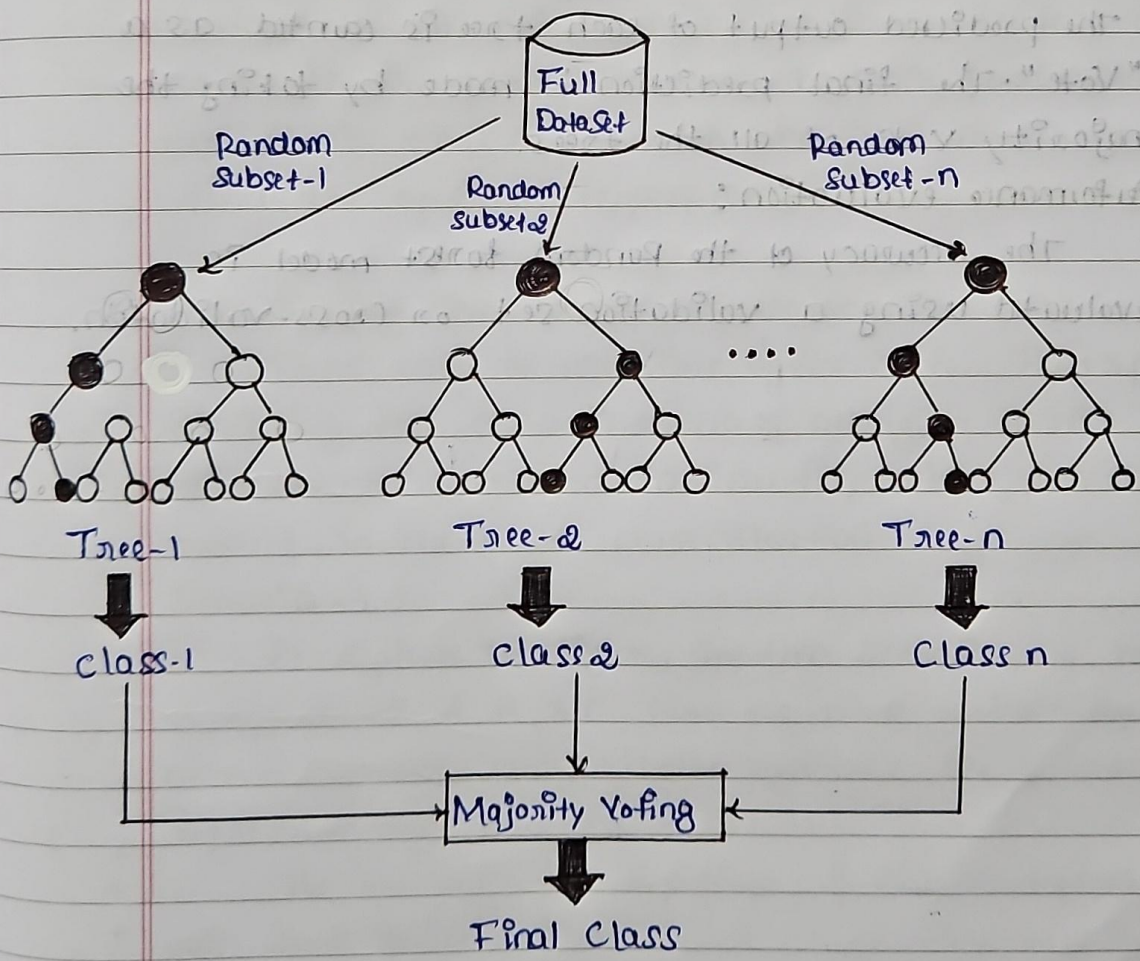
Random Forest is an ensemble classifier i.e: a combining classifier that combines & uses many decision tree classifiers.

Random Forest is a popular group learning method that builds multiple decision trees & aggregates their outputs to make a final prediction.

It works by creating a set of decision trees where each tree is trained on a randomly selected subset of training data & a randomly selected subset of features.

on the basis of the majority votes from the 'n' trees.

The final prediction is then made by aggregating the prediction of all the trees. The randomness in selecting the data & feature helps to increase the accuracy of the model.



4 Discuss the SVM model in detail with diff scenarios.

SVM Model:

Support Vector Machine (SVM) is a type of supervised learning algorithm used for both classification & regression analysis.

SVM is based on the concept of a surface, called a Hyperplane. A Hyperplane is a multidimentional plane that is used to separate diff classes of data in a classification problem. In SVM, a Hyperplane is used to find the best possible boundary that separates data points into diff classes.

The Goal of SVM analysis is to find a plane or Hyperplane which separates the instances on the basis of their classes.

In the overall training process the SVM algorithm analyses input data & identifies a surface in the multi-dimentional feature space called the Hyperplane. There may be many possibilities, & one of the challenges with the SVM model is to find the optimal Hyperplane.

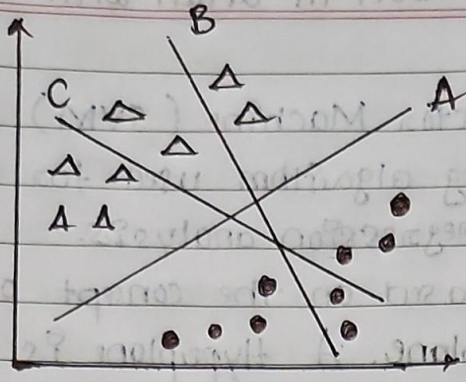
Identifying the correct Hyperplane in SVM:

There may be multiple options for Hyperplanes dividing the data instances belonging to the diff classes. We need to identify which one will result in the best classification.

Senario-1:

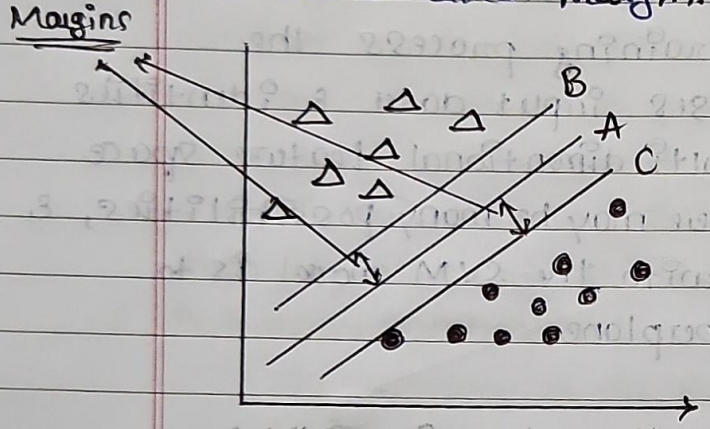
As depicted in figure, in this scenario, we have 3 Hyperplanes A, B & C. Now we need to identify the correct hyperplane which better segregates the 2 classes represented by Triangles & Circles.

As we can see Hyperplane A has performed this task quite well.



Senario 2:

As shown in below fig, we have 3 hyperplanes, A, B & C. Here maximizing the distance b/w the nearest data points of the both the classes & hyperplanes will help us decide the correct hyperplane. This distance is called margin.

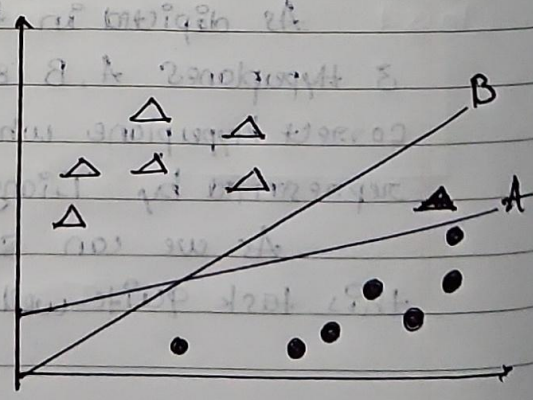


Hence, Hyperplane A is the correct Hyperplane. because the margine for Hyperplane A is high as compared to those B & C.

Scenario-3:

When we use the rules as discussed previously to identify the correct Hyperplane in the scenario shown in the figure there may be a chance of selecting the Hyperplane B as it has higher margin than A.

But here is the catch; SVM selects the hyperplane which classifies the classes accurately before maximizing the margin. Here Hyperplane B has classification error & A has classified all the data instances correctly.



Therefore A is the correct Hyperplane.

Scenario-4:

In this scenario as shown in the fig, it is not possible to distinctly segregate the 2 classes by using a straight line, as one data instance belonging to one of the classes (Δ) lies in the territory of the class (●) as an outlier.