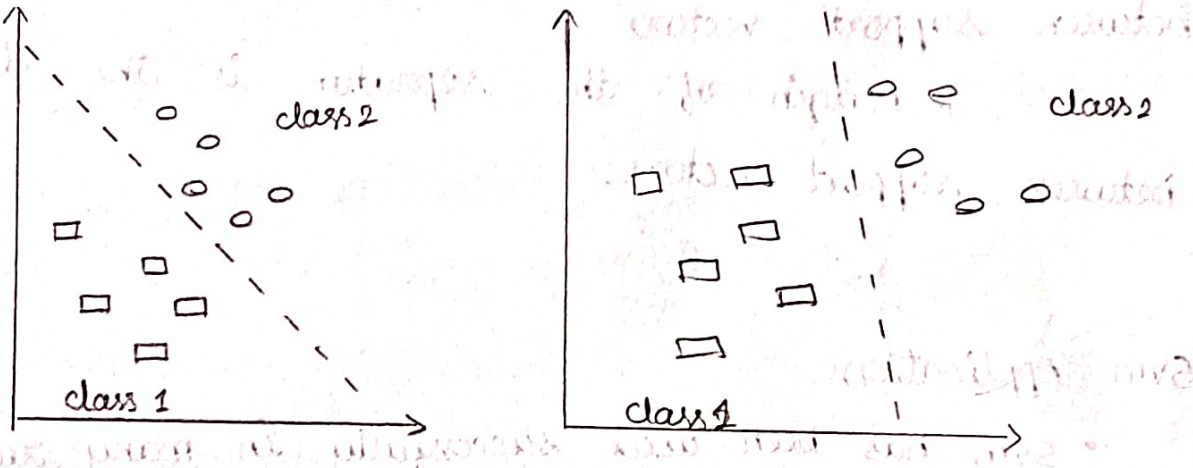
Support vector machine (svm)

definition:

svm are a set of supervised learning methods. Learn from dataset, used for classification.

Large Margin classifier

* It is a vector space based machine learning method where the goal is find decission boundary between two classes that is maximally from the data.

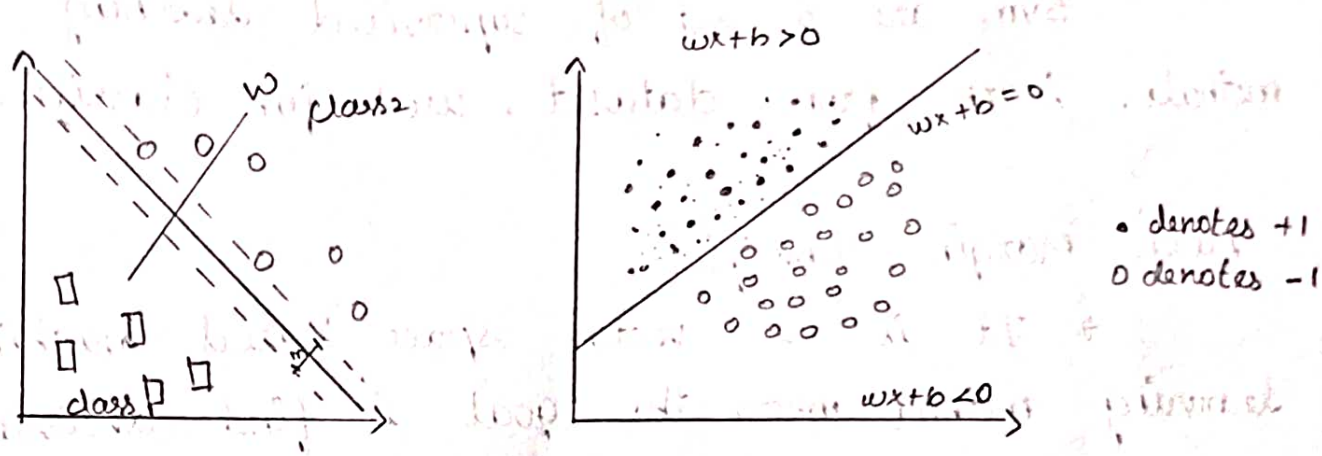


Bad decision boundary of SVM

* 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.

Good decision boundary

The decision boundary should be faraway from the data of both class as possible.



Good decision boundary.

m - margin

* The gap between data point and the classifier boundary.

between support vectors.

* Margin of the seprator is the distance between support vectors.

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

SVM Applications

• SVM has been used successfully in many real-world problems,

* Text (and hypertext) categorization
* Image classification
* Hand-written character recognition
* Bioinformatics (protein classification, cancer classification)
* Determination of SPAM email.

Types of SVM

* Simple or linear SVM
* Kernel or non-linear SVM

Simple or linear SVM

* A linear SVM refers to the SVM type used for classifying linearly separable data.
* A simple SVM is typically used to address classification and regression analysis problems.

Kernel or non-linear SVM

* Non-linear data that cannot be segregated into distinct categories with the help of a straight line is classified using a kernel or non-linear SVM.
* Kernel SVMs are typically used to handle optimization problems that have multiple variables.

Advantages of SVM:

* SVM works relatively well when there is a clear margin of separation between classes.

* SVM is more effective in high dimensional spaces.

* SVM is effective in cases where the number of dimensions is greater than the number of samples.

* SVM is relatively memory efficient

Disadvantages of SVM:

* SVM algorithm is not suitable for large data set.

* SVM does not perform very well when the data set has more noise.

* It requires more time to process

* It is less accurate

Linear regression in machine learning

* Linear regression is also a type of machine learning algorithm more specifically a supervised machine learning algorithm.

* which can be used for prediction on new datasets.

Types of linear regression,

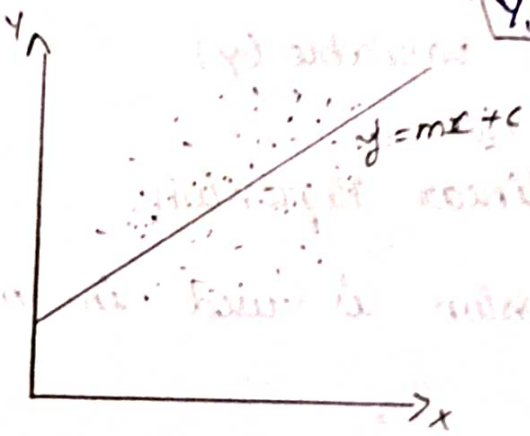
* When there is only one independent feature, it is known as Simple linear Regression.

* when there are more than one feature, it is known as Multiple linear regression

* Similarly, when there is only one dependent variable, it is considered as univariate linear regression.

* When there are more than one dependent variable, it is known as Multivariate Regression.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{i_2} + \beta_x X_{i_x} + \epsilon_i$$



* Here y is called dependent or target variable

* x is called an Independent variable also known as the predictor of y.

Linear Regression line

* The linear regression line provides valuable insights into the relationship between the two variables.

* Positive linear regression line
* Negative linear regression line

Positive linear regression line :

* A positive linear regression line indicates a direct relationship between the independent variable (x) and dependent variable (y)

Negative linear regression line :

* A negative linear regression line indicates an inverse relationship between the independent variable (x) and dependent variable (y)

Applications of linear Regression.

* Linear regression is used in many different fields, including.

* Finance
* economics
* psychology

Simple linear regression

$$y = \beta_0 + \beta_1 x$$

where

* y is the dependent variable
* x is the independent variable
* β_0 is the intercept
* β_1 is the slope

Multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n$$

where:

* y is the dependent variable
* $x_1, x_2 \dots, x_n$ are the independent variables
* β_0 is the intercept
* $\beta_1, \beta_2, \dots \beta_n$ are the slopes.

Advantages of linear Regression

* Linear regression is a simple algorithm and easy to implement.

* Linear regression is computationally efficient and can handle large datasets effectively.

* Linear regression is relatively robust to outliers compared to other machine learning algorithms.

* It is a well established algorithm with a rich history and widely available in various machine learning libraries and software packages.

Disadvantages of linear regression.

* If the relationship is not linear with dependent and independent variable may not perform well.

* It is sensitive to multicollinearibility.

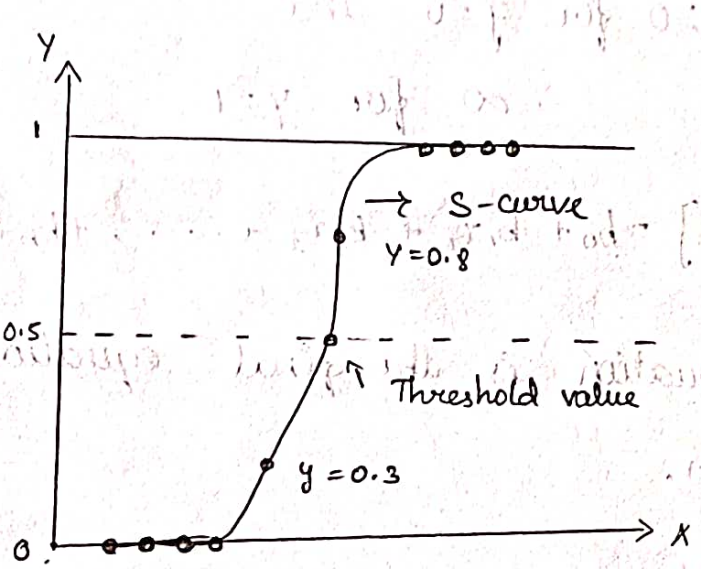
* Over fitting occurs.

* More advanced machine learning techniques may be necessary for deeper insights.

Logistic regression

* Logistic regression is one of the most Popular Machine learning algorithms, which comes under the supervised learning techiique.

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



Logistic function (sigmoid function):

* The sigmoid function is a mathematical function used to map the predicted values to probabilities.

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

Logistic regression Equation:

* The logistic regression equation can be obtained from the linear regression equation.

* we know the equation of the straight line can be written as:

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

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

$$\therefore \infty \text{ for } Y = 1$$

$$\log\left[\frac{Y}{1-Y}\right] = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

* The above equation is the final equation for logistic regression.

Types of logistic Regression:

* On the basis of the categories, logistic regress can be classified into three types,

* Binomial
* Multinomial
* Ordinal

Binomial:

In binomial logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, pass or fail ecte,

Multinomial:

In multinomial logistic regression, there can be 3 or more possible unordered types of the dependent variable. such as "cat", "dogs", or "sheep".

Ordinal:

In ordinal logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "medium", or "High".

Advantages

* Logistic regression is easier to implendent interpret, and very efficient to train.

* It makes no assumptions about distributions of classes in feature space.

* It can easily eoctend to multiple classes.

* It is very fast at classifying unknown records.

Disadvantages

* If the number of observation is lesser than the number of features, logistic regression should not be used.

* It constructs linear boundaries.

* It can only be used to predict discrete functions.

* In linear regression independent and dependent variables are related linearly.

DECISION TREE:

* A decision tree is a flowchat-like structure used to make decisions or predictions.

* Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test.

* Each leaf node corresponds to a class label or a continuous value.

STRUCTURE OF A DECISION TREE:

1. ROOT NODE:
Represents the entires dataset and the initial decision to be made.

2. INTERNAL NODE:
Represent decisions or tests on attributes. Each internal node has one or more branches.

3. BRANCHES:
Represent the outcome of a decision or test, leading to another node.

4. LEAF NODES:

Represent the final decision or prediction. No further splits occur at these nodes.

The process of creating a decision tress involves.

1. SELECTING THE BEST ATTRIBUTE:

Using a metric like Gini impurity, entropy, or information gain, the best attribute to split the data is selected.

2. SPLITTING THE DATA-SET:

The dataset is split into subsets based on the selected attributed.

3. REPEATING THE PROCESS:

The process is repeated recursively for each subset, creating a new internal node or leaf node until a stopping criteria in met.

ADVANTAGES:

* Simplicity and interpretability: Decision trees are easy to understand and interpret.

* versatility: can be used for both classification and regression tasks.

* Decision trees do not required normalization or scaling of the data.

* capable of capturing non-linear relationships between features and target variables.

DISADVATAGES:

* Decision trees can easily overfit the training data, especially if they are deep with many nodes.

* Small variations in the data can result in a completely different tree being generated.

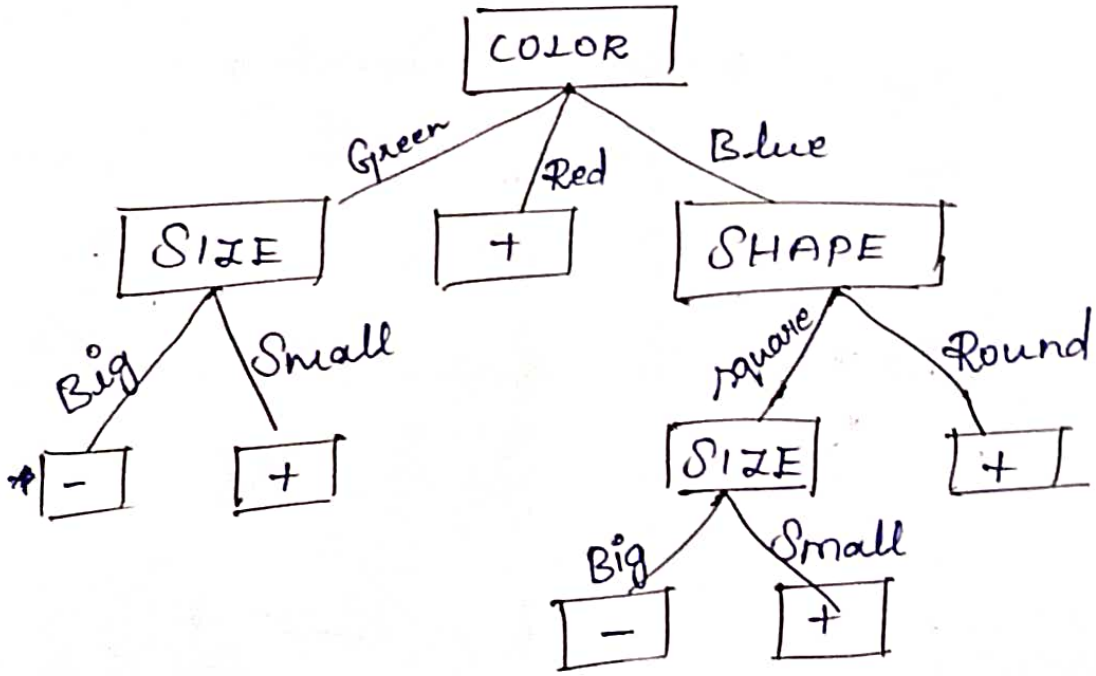
APPLICATIONS OF DECISION TREES:

* Used in stratigic planning and resource allocation.

* Assists in diagnosing diseases and suggesting treatment plans.

* Helps in credit scoring and risk assessment.

* Used to segment customer and predict customer behaviour.



② RANDOM FOREST:

* Random forest is a powerful tree learning technique in machine learning.

* It works by creating a number of decision trees during the training phase.

* This randomness introduces variability among individual trees, reducing the risk of overfitting.

RANDOM FOREST ALGORITHM WORK

The working technique may be explained within the below steps and diagram.

STEP 1: Select random k statistics points from the schooling set.

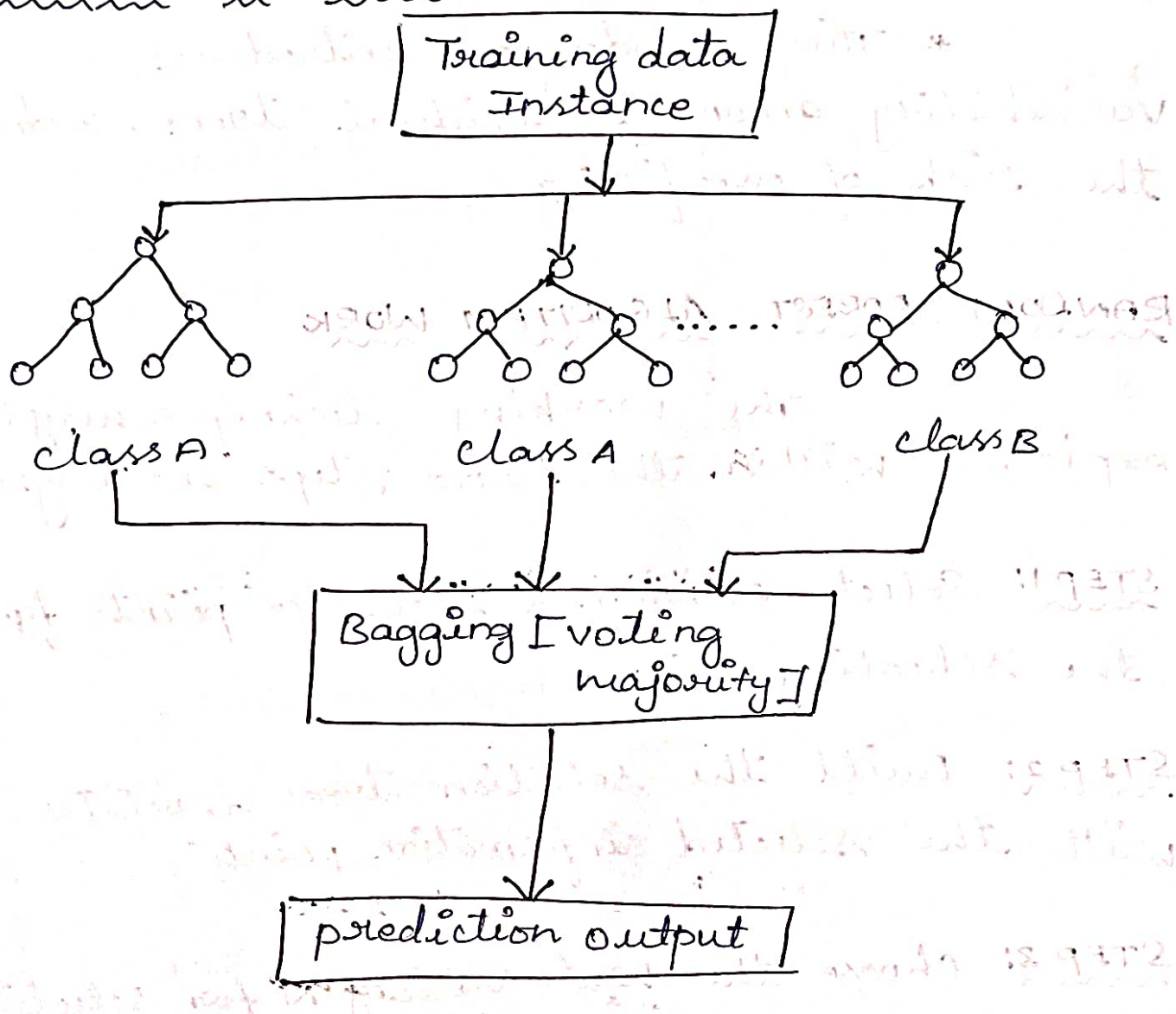
STEP 2: Build the selection trees associatis with the selected information points.

STEP 3: Choose the wide variety N for selection trees which we want to build.

STEP 4: Repeat step 1 and 2

STEP 5: For new factors, locate the predictions of each choice tree and assign the new records factors to the category that wins most people's votes.

EXAMPLE OF RANDOM FOREST



APPLICATIONS OF RANDOM FOREST:

i) BANKING: Banking zone in general uses this algorithm for the identification of loan danger.

ii) MEDICINE: With the assistance of this set of rules, disorder traits and risks of the disorder may be recognized.

iii) LAND USE: We can perceive the areas of comparable land use with the aid of this algorithm.

iv) MARKETING: Marketing tendencies can be recognized by the usage of this algorithm.

ADVANTAGES:

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

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

DISADVANTAGES:

* The random forest can be used for both class.

* Regression responsibilities, it isn't extra appropriate for regression obligations.