Date
Page

Unit-2

Logistic Regression

→ Logistic Regression comes under Supervised Learning Technique.

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

→ Logistic regression predicts the output of a categorical dependent variable.

→ It gives the probabilistic values which lie b/w 0 & 1

→ Logistic Regression is used for solving the classification problems.

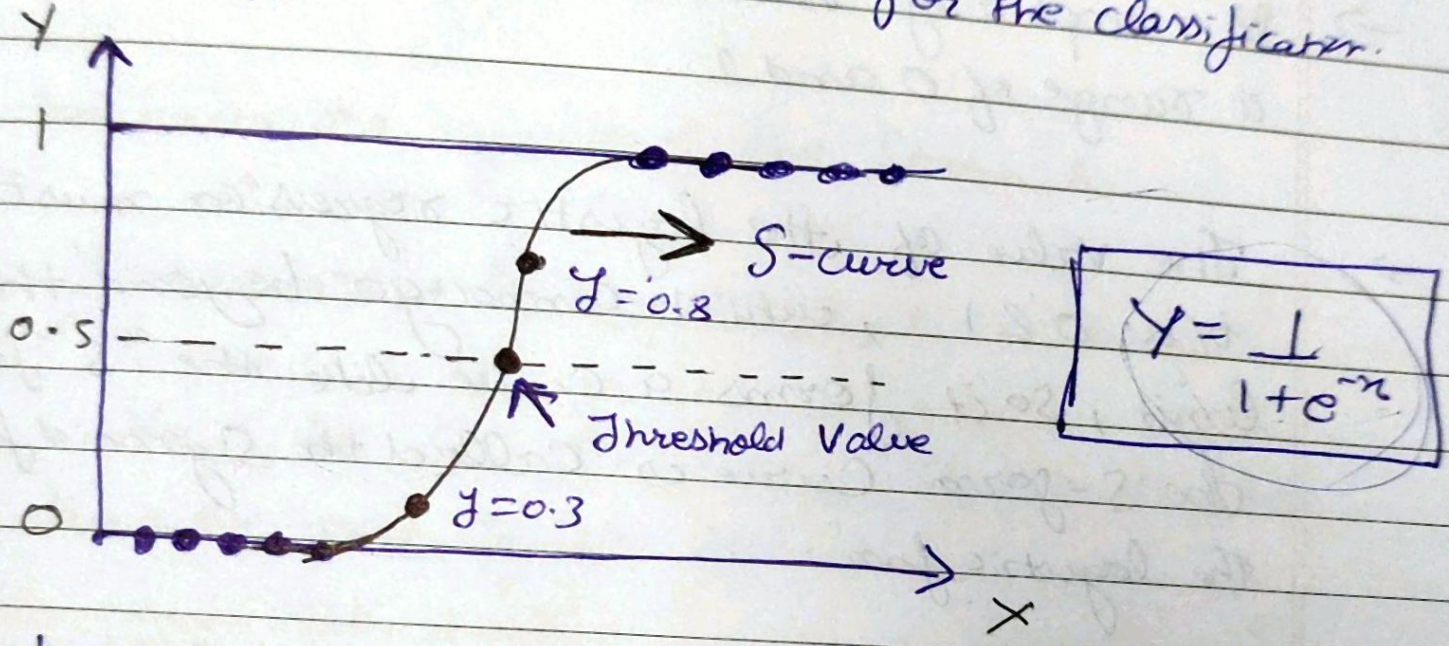
→ we fit an "S" Shaped logistic func^c which predicts two maximum values (0 or 1)

→ The curve from the logistic func^c indicates the likelihood of something.

→ It is a significant machine learning algorithm because it has the ability to provide

probabilites and classify new data using continuous and discrete datasets.

It can be used to classify the observations using different types of data and can easily determin the most effective variables used for the classificater.



Logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such

→ Logistic regresion uses the concept of predictive modeling as regression; therefore it is called logistic regression, but is used to classify samples. Therefore it falls under the classification algorithm.

Logistic Function (Sigmoid Function):

→ Sigmoid function is a mathematical fun^c 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 b/w 0 & 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid fun^c or the logistic fun^c.

→ We use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumption for Logistic Regression

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.
- We know the equation of the straight line

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

- Y can be b/w 0 & 1 only, so for this let's divide the above equation by $(1-y)$:

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

- But we need range b/w $-[\text{infinity}]$ to $+[\text{infinity}]$ then take logarithm of the equation it will become

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n$$

Types of Logistic Regression

- Binomial: There can be only two possible types of the dependent variables, such as 0 or 1, Pass or fail etc.
- Multinomial: ~~Unordered Logistic~~ there can be 3 or more possible unordered types of the dependent variable such as "cat", "dogs" or "Sheep"
- Ordinal: There can be 3 or more possible ordered type of dependent variables such as "Low", "Medium" or "High".

Steps in Logistic Regression:-

- Data Pre-processing step
- Fitting Logistic Regression to the Training Set
- Predicting the test result
- Test accuracy of the result (Creation of confusion matrix)
- Visualizing the test set result.

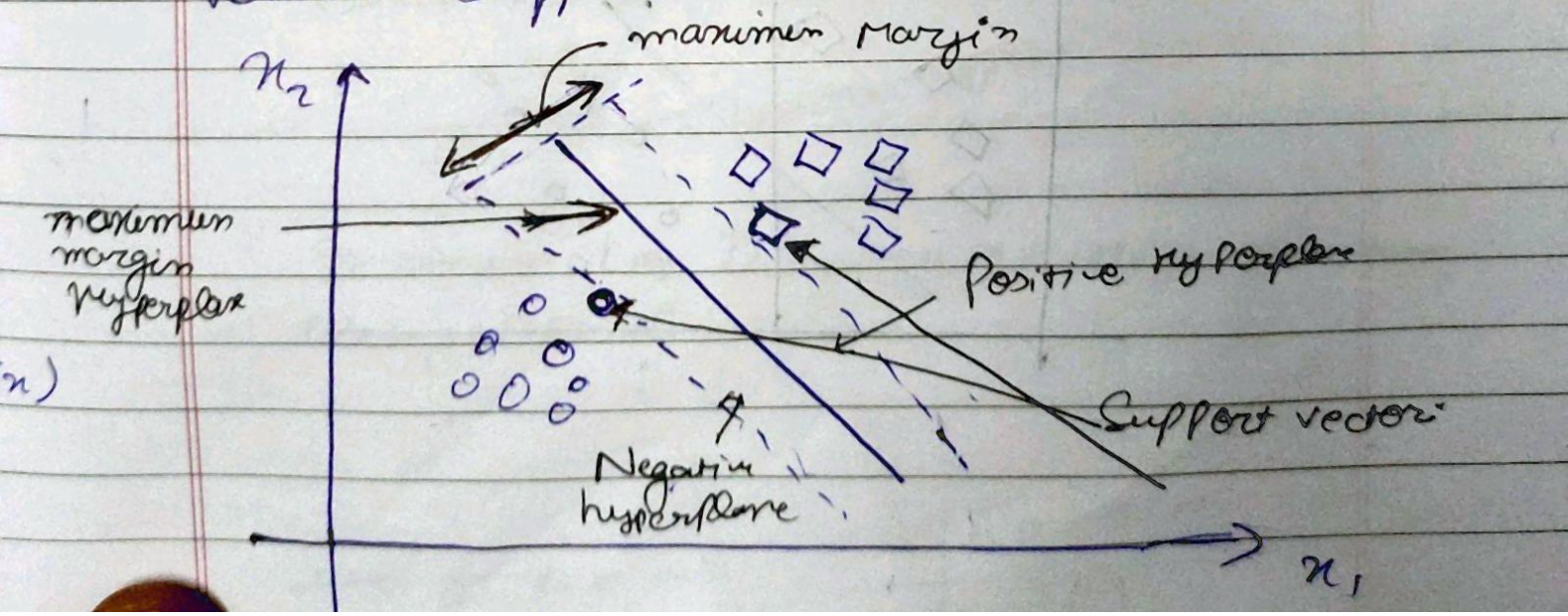
(-) | (+)

SVM

→ It is one of the most popular Supervised Learning algorithm, which is used for classification as well as Regression problems. However,

→ The Goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. The best decision boundary is called a hyperplane.

→ SVM choose the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as Support Vectors, and hence algorithm is termed as Support Vector Machine.



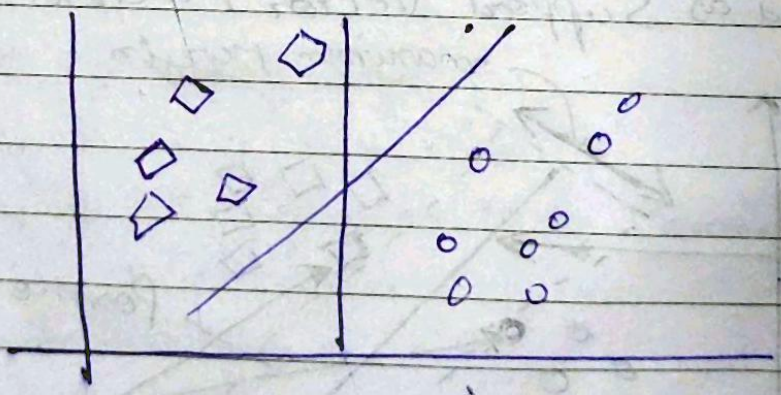
→ A SVM creates a decision boundary b/w two data and on basis of support vectors it will classify the data.

Note:- We always need choose the hyperplane in which we got max margin.

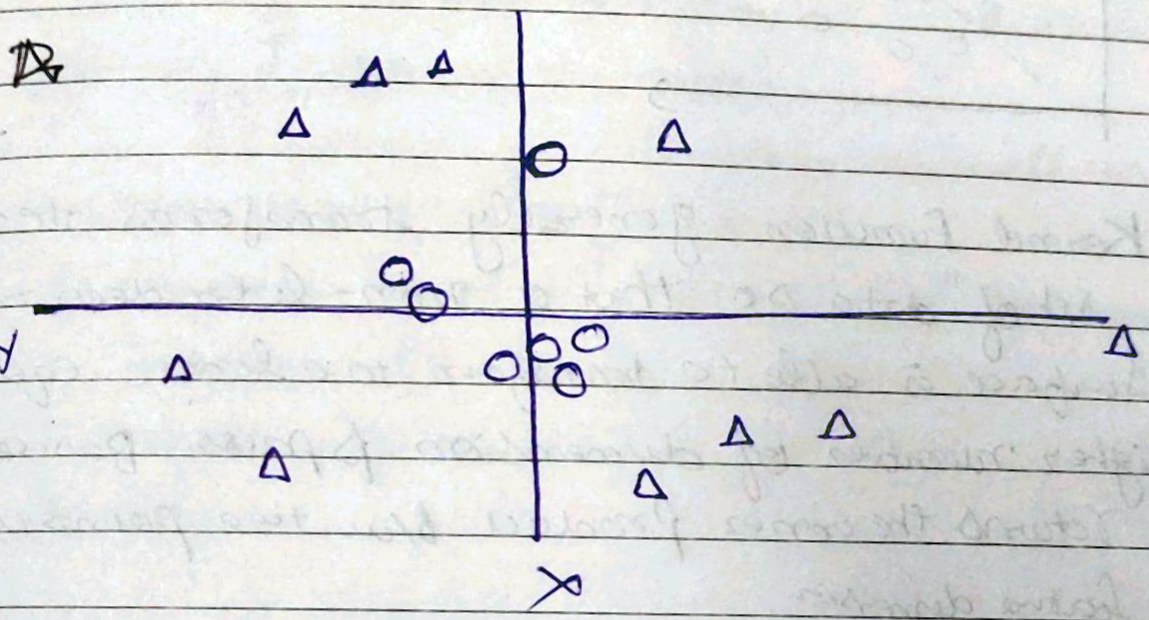
$$\text{margin} = d^- + d^+$$

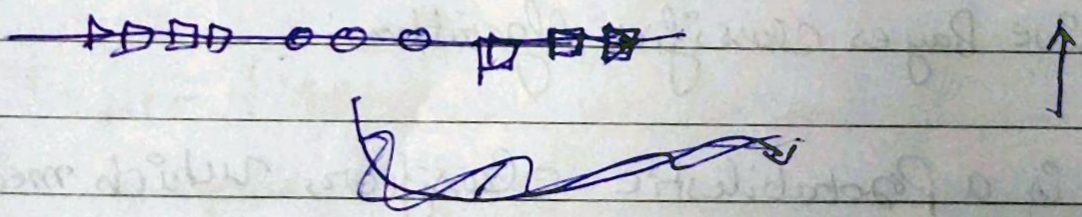
Types of SVM

1 Linear SVM: Linear SVM is used for linearly seperable data, which mean if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly seperable data

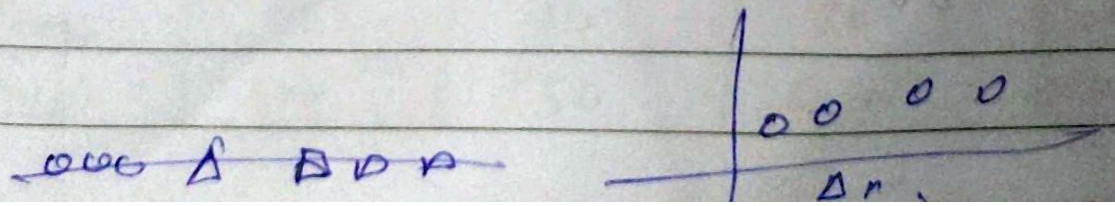


2. Non-linear SVM: It is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

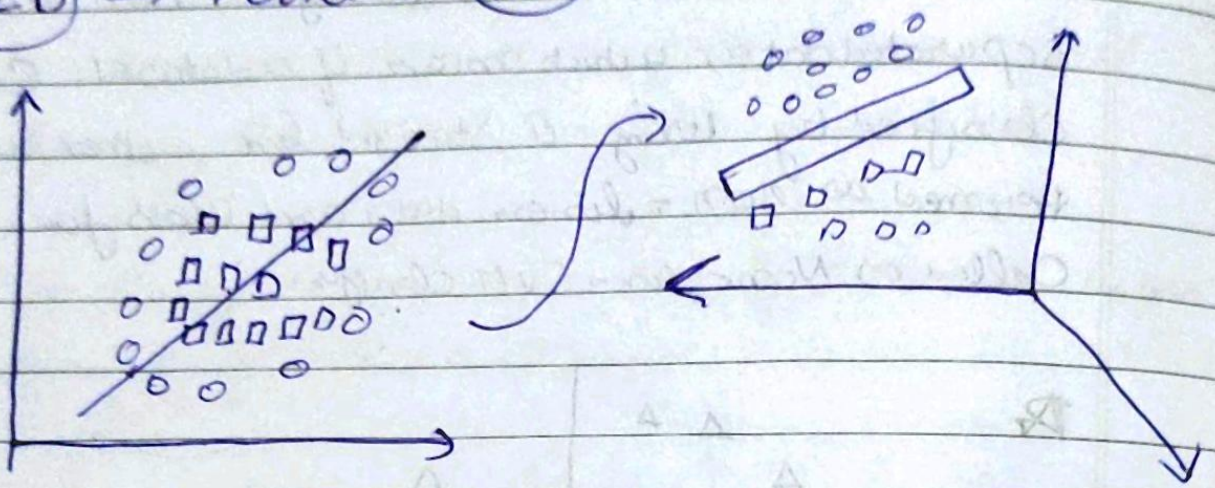




It convert Low Dimesion to High Dimension using kernel.



(LD) => Kernel => (HD)



Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Basically, it returns the inner product b/w two points in a standard feature dimension.

Naive Bayes classifier Algorithm

- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object
- Example of Naive Bayes
  Spam filtration, Sentimental analysis and classifying articles.

Naive:- it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

$$P\left(\frac{A}{B}\right) = \frac{P(B|A)\,P(A)}{P(B)}$$

Advantage

- It is one of the fast and easy ML algorithms to predict a class of datasets.
- Can be used for Binary as well as multi-class classification.
- It perform well in multi-class predictions as compared to the other Algorithms
- It is most popular choice for text classification problems.

Naive Bayes classifier Algo

Fruit = {Yellow, Sweet, long}

Fruit	Yellow	Sweet	Long	Total
Orange	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	150
Total	800	850	400	1200

65 | 358 — 29 — Date / Page

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(\text{Yellow} | \text{orange}) = \frac{P(\text{orange} \cap \text{yellow}) : P(\text{Yellow})}{P(\text{orange})}$$

$$= \frac{\frac{350}{800} \times \frac{800}{1200}}{\frac{650}{1200}} = 0.5$$

$$P(\text{Sweet} | \text{orange}) = \frac{P(\text{Orange} | \text{Sweet}) \cdot P(\text{Sweet})}{P(\text{orange})}$$

$$= \frac{\frac{350}{650} \times \frac{850}{1200}}{\frac{650}{1200}} = \frac{35}{120} \times \frac{120}{65}$$

$$= \frac{0.411 \times 0.70}{0.541} = 0.53$$

$$P(\text{Long} | \text{orang}) = \frac{P(\text{orange} | \text{Long}) \cdot P(\text{Long})}{P(\text{orange})}$$

$$= \frac{\frac{0}{400} \times \frac{400}{1200}}{\frac{650}{1200}} = 0$$

Do Same for Banana, and otherfruits also

P(Fruit | orange) = 0.53 × 0.69 × 0 = 0

P(Fruit | Banana) = 1 × 0.75 × 0.87 = 0.65

P(Fruit | Others) = 0.33 × 0.66 × 0.33 = 0.072

∴ The one whose probability is more. the answer is that Banana

Fruit = Banana.

KNN

query => X = (Maths = 6, CS = 8), K = 3

maths	CS	Result
4	3	Fail
6	7	Pass
7	8	Pass
5	5	Fail
8	8	Pass

Date
Page

(I) ∫

Euclidean distance:-

$$d = \sqrt{|x_0 - x_A|^2 + |x_{0_2} - x_{A_2}|^2}$$

(I) $\sqrt{(6-4)^2 + (8-3)^2} = \sqrt{29} = 5.38$

(II) $\sqrt{(6-4)^2 + (8-7)^2} = 1$

(III) $\sqrt{(6-7)^2 + (8-8)^2} = 1$

(IV) $\sqrt{(6-5)^2 + (8-5)^2} = \sqrt{10} = 3.16$

(V) $\sqrt{(6-8)^2 + (8-8)^2} = 2$

K=3 is given so we have to consider 3 nearest distance

Pass (P) P P P = 3-P

Fail O-F

3 > 0

So pass is more than Fail So we can declare x is also pass

→ K-NN algorithm assumes the similarity b/w the new case / data and available cases and put the new case into the category that is most similar to the available categories.

→ KNN is non-parametric algorithm which means it does not make any assumption on underlying data.

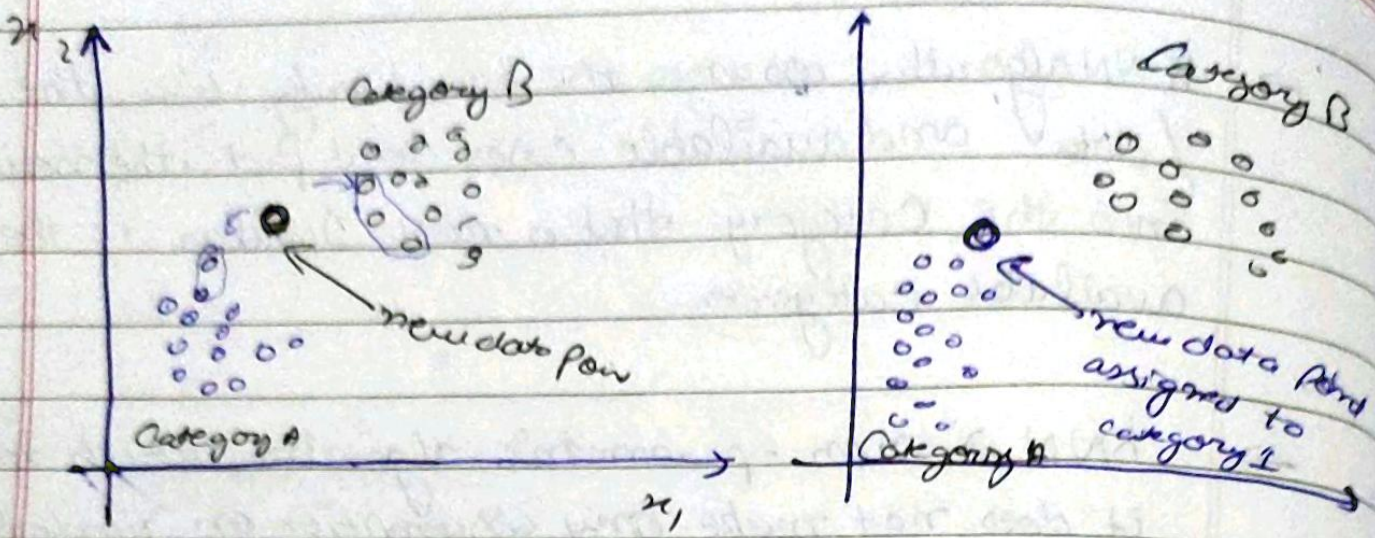
→ It is also called a lazy learner because it does not ~~make any assumption on underlying data.~~ learn from

→ the training set immediately instead it stores the dataset and ~~at~~ the time of classification it performs an action on the dataset.

→

Why we need KNN Algorithm

Category A & category B and we have a new data point x1, so this data will lie in which of these categories. To solve this problem we need K-NN



~~KNN works~~ Advantage & Disadvantage from Javapoint.

Cross Validation

1000 words

(70%) (30%)

Let suppose the data divide in 70% training data and 30% for accuracy.

Suppose we ~~are~~ ~~trying~~ take random-State 0 it will split & give some accuracy

When we do random-State 50 it will split & give some other accuracy

classmate
Date
Page

⊕ For different random state accuracy fluctuate

To get prevent from this we have a concept of cross validation

→ Cross validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

Basic Steps of Cross validation are

- Reserve a subset of the dataset as a validation set
- Provide the training to the model using the training dataset.
- Evaluate model performance using the validation set. If the model performs well with the validation set, perform the further step, else check for the issues

Math.

Methods Used for Cross-Validation

1. Validation Set Approach
2. Leave-p-out cross-validation
3. Leave one out cross-validation
4. K-fold cross-validation
5. Stratified K-fold cross validation

Leave-P-out-Cross Validation

→ The p datasets are left out of the training data.
→ If there are total n datapoints in the original input dataset, then n-p data points will be used as the training dataset and the p data points as the validation set.
→ This complete process is repeated for all the samples and the average error is calculated to know the effectiveness of the model.

Disadvantage.

It can be computationally difficult for the large p.

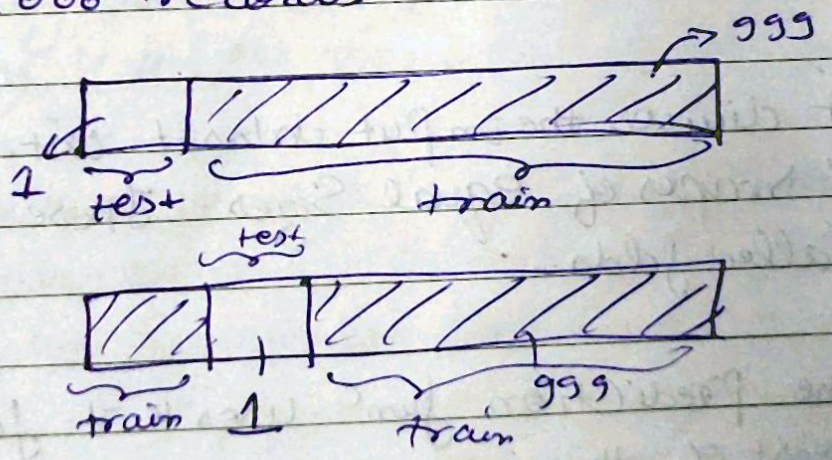
Leave one out Cross - Validation

Similar to leave-p-out CV but instead of p, we need to take 1 dataset out of training.

Disadvantage

- The process is executed for n times; hence execution time is high
- This approach leads to high variation in testing the effectiveness of the model as we iteratively check against one data point.

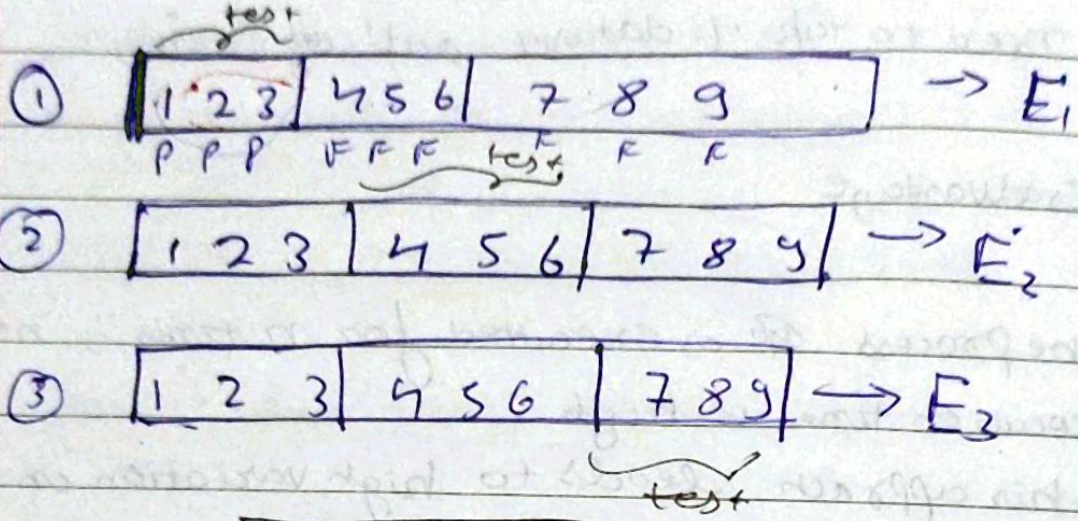
Eg 1000 records



- Bias is minimum as all the data points are used.

K-Fold Cross validation

Ex 3-Fold CV



$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

It divides the input dataset into K groups of samples of equal sizes. These samples are called folds.

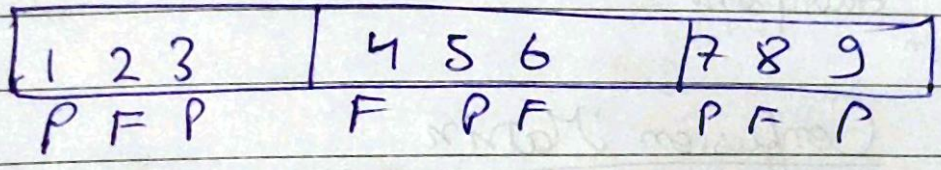
The prediction func uses k-1 folds, and the rest of the folds are used for the test set.

→ Easy to understand, and the output is less biased than other methods.

The steps for k-fold cross validation

- Stratified k-fold cross-validation

In this data is divided into same proportion.



Holdout Method

In this method we need to remove a subset of the training data and use it to get prediction results by training it on the rest part of the dataset.

Ek nikalte he test ke liye baki training ke liye rakh dete hen, fir Ek nikalte he bake k. training ke liye rakh dete hen.

limitation of cross validation

Java Point

Application

Javapoint

Confusion Matrix

Total Program = 165	NO	Yes	
		Predicted	
ACTUAL NO	50 [TN]	10 [FP]	60
ACTUAL Yes	5 [FN]	100 [TP]	105
	55	110	

- Accuracy = $\frac{TP + TN}{Total}$ $= \frac{100 + 50}{165} \approx 0.91$
- Error rate $= 1 - Accuracy$

  or

  $\frac{FP + FN}{Total} = 0.09$

- Precision: $\frac{TP}{\text{Predicted yes}} = \frac{100}{110} = 0.9$

- Recall = $\frac{TP}{\text{actual Yes}} = \frac{100}{105} = 0.95$

- F measure = $\frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$

- True Negative :- Model has given prediction NO, and the real or actual value was also NO.

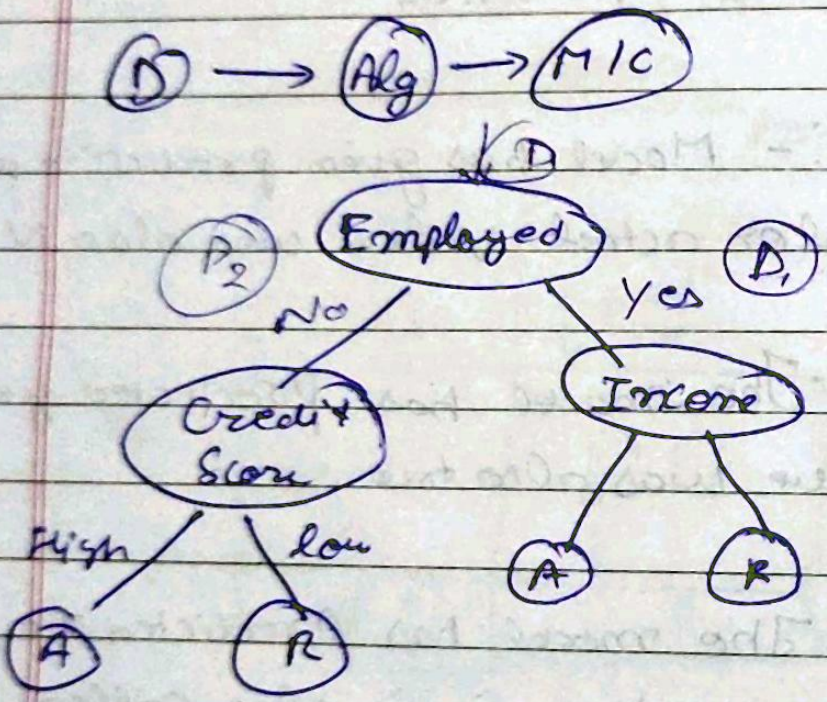
- True Positive - The model has predicted yes, and the actual value was also true.

- False Negative : The model has predicted no, but the actual value was Yes, it is also called as Type-II error.

- False Positive The model has Predicted Yes, but the actual value was NO. It is also called a Type-I error.

Decision Tree

A classifier (Tree structured)
Decision Node (Test)
Leaf Node (classification / value)
Also used for Regression
Test is performed on the feature / Attribute (value)



→ target attribute

Age	Competion	Type	Profit
old	Yes	S/w	Down
old	No	S/w	Down
old	No	H/w	Down
mid	Yes	S/w	Down
mid	Yes	H/w	Down
mid	No	H/w	up
mid	No	S/w	up
new	Yes	S/w	up
new	No	H/w	up
new	NO	S/w	up

$$IG = \frac{-P}{P+N} \log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N} \log_2\left(\frac{N}{P+N}\right)$$

$$E(A) = \sum_{i=1}^{v} \frac{P_i + N_i}{P+N} I(P_i, N_i)$$

$$\text{Gain} = IG - E(A)$$

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2}$$

$$IG = -\left[\frac{5}{10}\log_2\left(\frac{5}{10}\right) + \frac{5}{10}\log_2\left(\frac{5}{10}\right)\right]$$

$$= -\left[0.5 \times \log_2 2^{-1} + 0.5 \log_2 2^{-1}\right]$$

$$= -\left[0.5 \times (-1 \log_2 2) + 0.5 \times (-1 \log_2 2)\right]$$

$$= -[-0.5 - 0.5] = -[-1]$$

$$IG = 1$$

Age

	Down	up
old	3	0
mid	2	2
new	0	3

Probability

$$I(old) = -\left[\frac{3}{3}\log\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] = 0 \times \frac{3}{10} = 0$$

$$I(mid) = -\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] = 1 \times \frac{4}{10} = 0.4$$

$$I(new = -\left[\frac{0}{3}\log_2\left(\frac{0}{3}\right) + \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] = 0 \times \frac{3}{10} = 0$$

$$E(Age) = 0.4$$

Entropy:- Entropy is nothing but the measure of disorder. It is the measure of purity / impurity.

More the uncertainty more is entropy

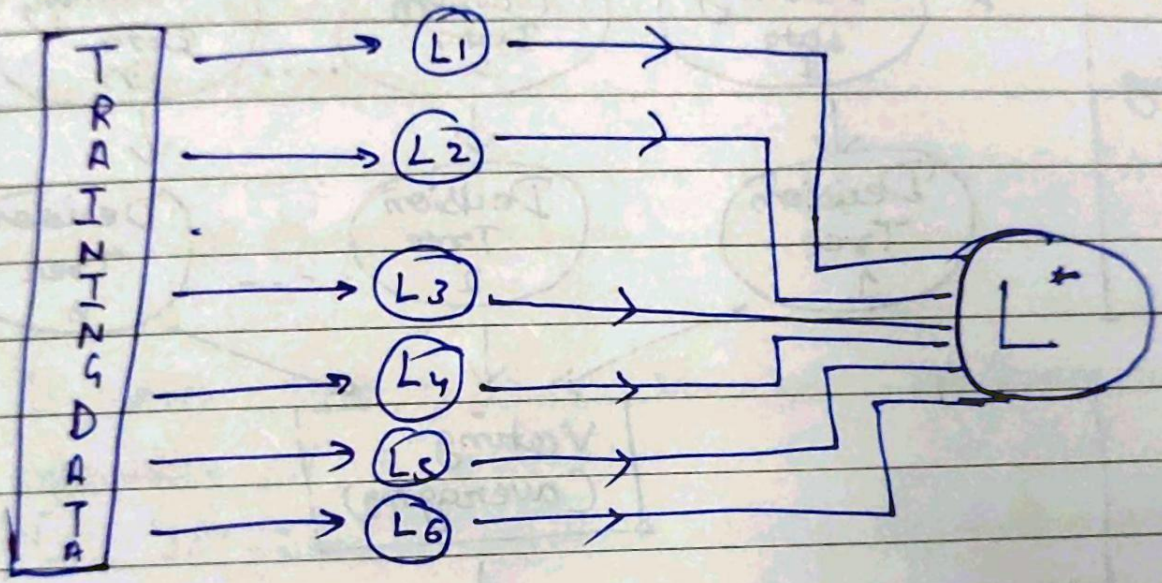
Information Gain :- It is a metric used to train Decision Trees. This metric measures the quality of a split.

The information gain is based on the decrease in entropy after a data-set is split on an attribute that returns the highest information gain

Gini impurity = $1-(P_x^2+P_m^2)$

Essemble method

Essemble methods are technique that aim at improving the accuracy of results in models by combining multiple model instead of using a single model.

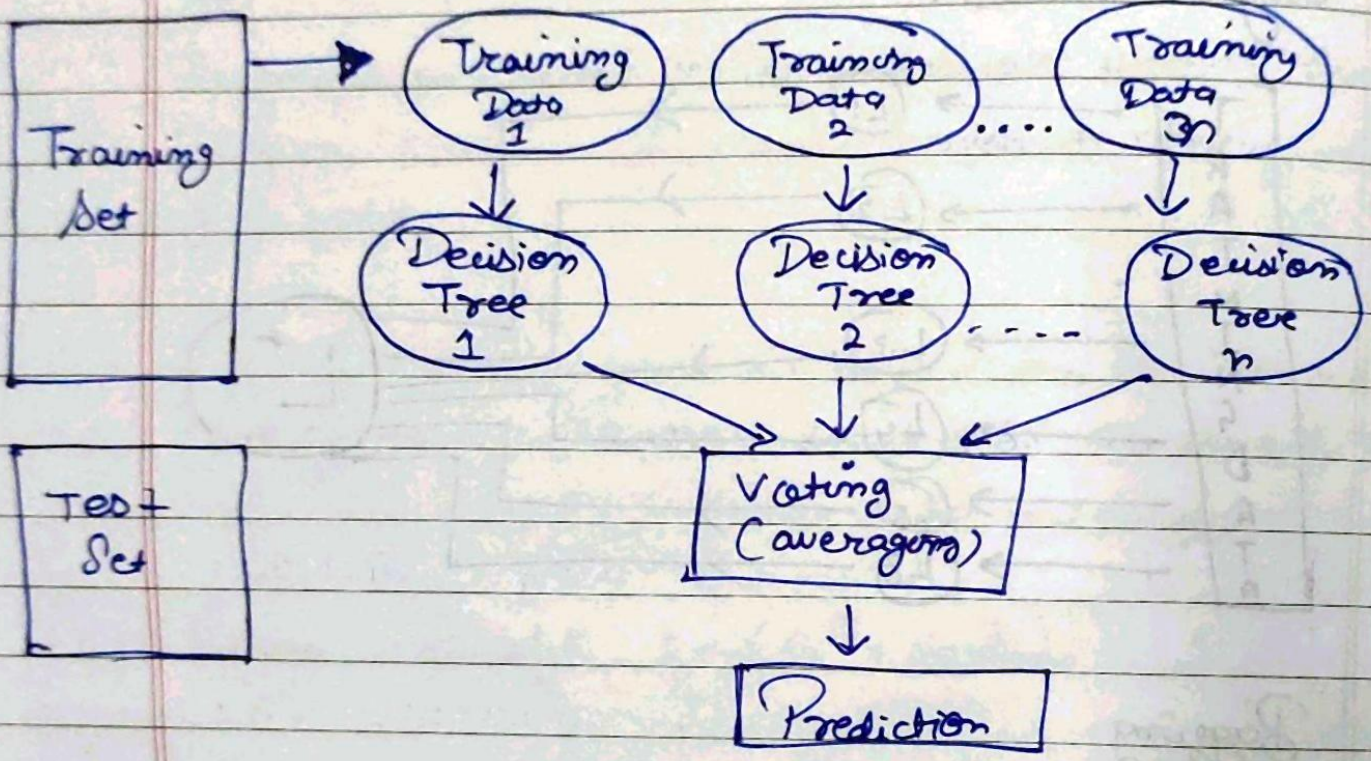


Bagging

It also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.

Random forest

The greater the number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.



Random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while other may not. But together, all the trees predict the correct output.

⇒ Two assumption

1) There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate result rather than guessed result.
2) The prediction from each tree must have very low correlations.

Why use Random forest

- It takes less training time as compared to other algorithm
- It predict output with high accuracy, even for the large dataset it runs efficiently
- It can also maintain accuracy when a large proportion of data is missing

Date
Page

Pruning

* Performance of a tree is increased by pruning

→ It values removing the branches that make use of features having low importance.

→ Remove each node with most popular class in that leaf → Reduced error pruning

→ A learning parameter is used to weigh whether nodes can be removed based on the size of the sub-tree → weakest link pruning.

Two types of pruning

i) Pre-pruning
we decide
we