Date :- 30th October, 2022

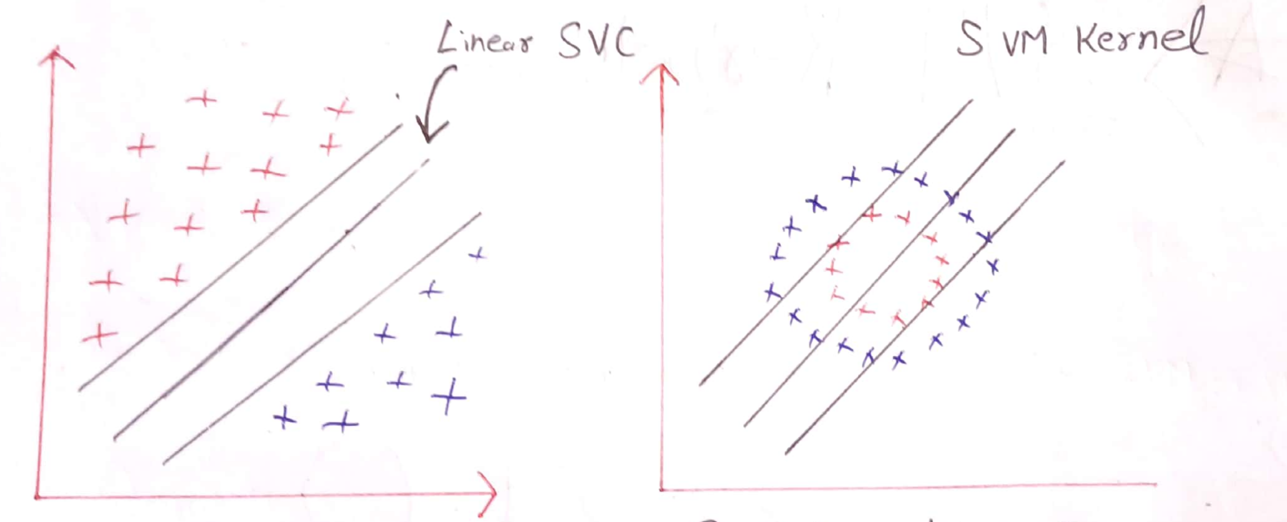
(61)

Machine Learning (Day-6)

Agenda

1) SVM Kernel
2) ROC - AUC Curve

*) SVM Kernels :-



① Linear Separable data ② Not a Linear Separble Data

Note :- When we create this ① Type of Best Fit line and Marginal Plane, we are actually Solving the Linear Seperable Data.

→ Called as Linear SVC (Fig ①)

P.T.O

*) If data is not a Linear Seperable data, you will not be able to create Best Fit Line and Not able to create a marginal Plane Even though we create it, the accuracy will be very low. (Fig ②)

*) For this Type of Problems, We have Some more SVM Kernels.

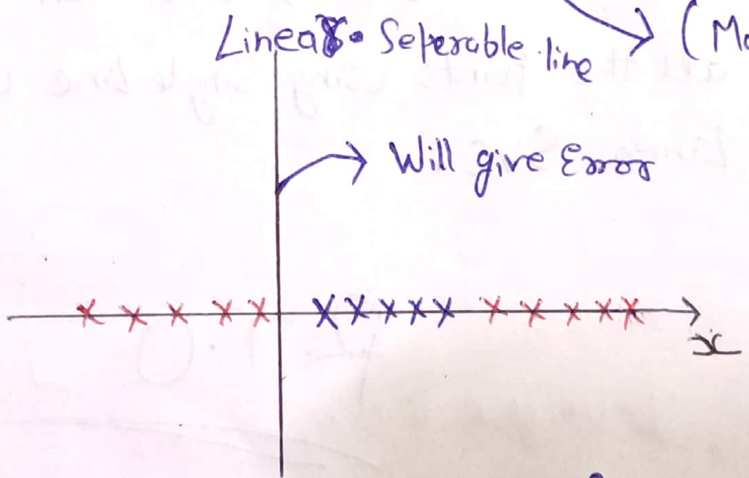
Q) What does SVM Kernels do?

⟹ The Main aim is to apply some Transformations technique. (Some Mathematical Formula) on the dataset.

This Transformation increase the dimension of the data.

SVM Kernels ⟶ Transformation ⟶ Increasing the dimension of the data.

(Mathematical formula)



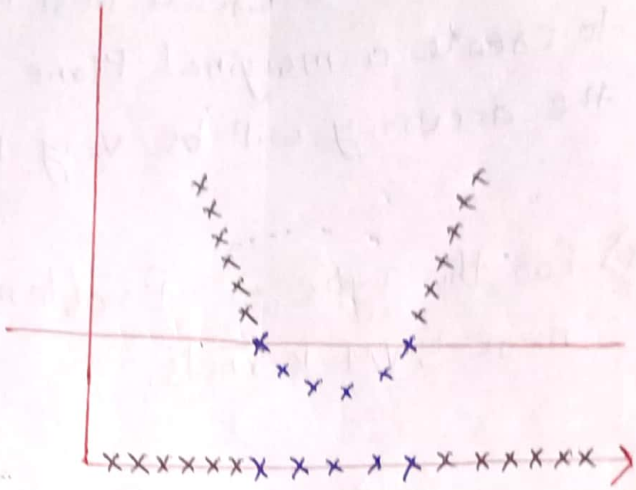
∴ We will transform the data from 1D → 2D

$y = x^2$

P.T.O

After →

Now, we can use Linear Seperable line.



$$\boxed{y = x^2}$$

$y = x^2$

So, if $x = -7$ $y = 49$

$x = -3$ $y = 9$ and so on...

*) What is the advantage of doing this Transformation ?

→ After Transformation, we can apply Linear SVM or SV

*) When we convert 1D → 2D then we can Divide all the points using Single Line which is called Linear SVC.

P.T.O

A) Types of SVM Kernel :—

1) Polynomial Kernel
2) RBF Kernel
3) Sigmoid Kernel.

1) Polynomial Kernel :—

⇒ Polynomial Kernel represent the Similarity of Vectors in the Training set of Data in a Feature Space over Polynominal of the original Variables used in the Kernels.

Fig. 3

→ Not Seperable with Best Fit Line.

∴ So, we need to Converts 2D → 3D.

P.T.O

Our main aim was to increase dimension, $2D \rightarrow 3D$

So, hyperPlane is created.

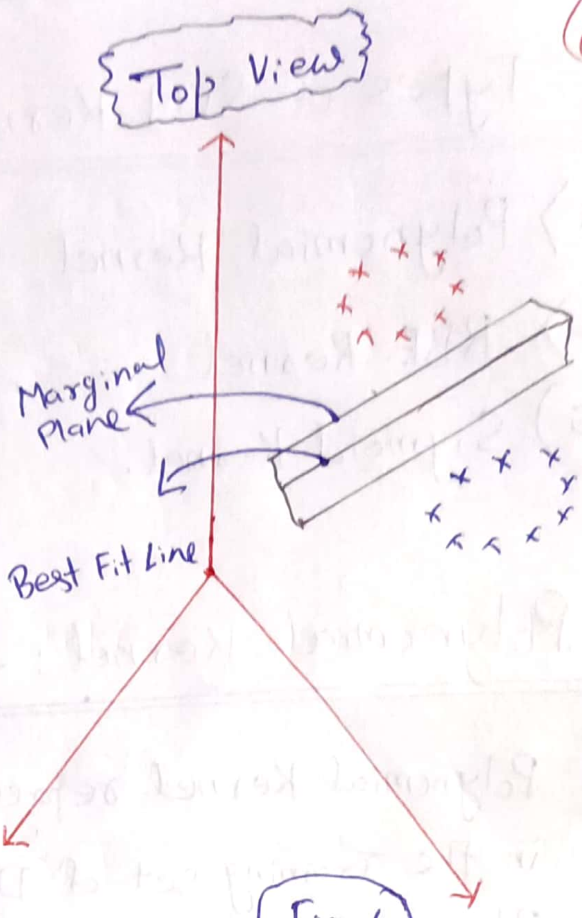


Fig. 4

Formula for Polynomial Kernel :—

$$\boxed{f(x_1, x_2) = (x_1^T x_2 + 1)^d}$$

$\therefore d$ = dimension,,

If we are converting $2D \rightarrow 3D$, the value of $d = 3$,,

$$\therefore x_1^T x_2 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 & x_1 \cdot x_2 \\ x_1 \cdot x_2 & x_2^2 \end{bmatrix}$$

Here, 3 Unique values $\Rightarrow x_1^2, x_1 \cdot x_2, x_2^2$

Now, initially at the time of Fig. 3, we ~~have~~ have 3 Features x_1 x_2 y.

Now, after the Transformation / Formula of Polynomial Kernel we have 6 Features.

$$x_1 \quad x_2 \quad \boxed{x_1^2 \quad x_1 \cdot x_2 \quad x_2^2} \quad y$$

∴ These New Features can be plotted as the 3D in Fig. 4, which means that

In Fig. 4 :- x_1 will be x_1^2
x_2 will be x_2^2
y will be $x_1 \cdot x_2$

and once we have all these points, we will be able to clearly seperate the points.

→ Use Polynomial Kernel, to get better Accuracy.

2 Radial Basis Function Kernel (RBF Kernel)

$$K(\vec{x}_i, \vec{l}_1) = e^{-\frac{\|\vec{x} - \vec{l}_i\|}{2\sigma^2}}$$

→ Hyper Parameter.

3 Sigmoid Kernel :-

⇒ IT can be used as the Proxy for Neural Networks

$$K(x, x_i) = \tanh(\sigma x^T x_i + \gamma)$$

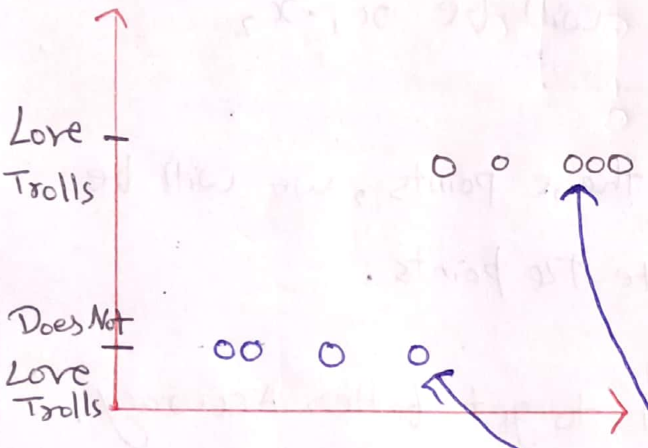
2) ROC And AUC :-

ROC ⇒ Receiver Operating Characteristic

AUC ⇒ Area Under Curve.

*) Threshold ⇒ Threshold is Super Important

Problem :- Whether or Not Someone Love the Movie Trolls.

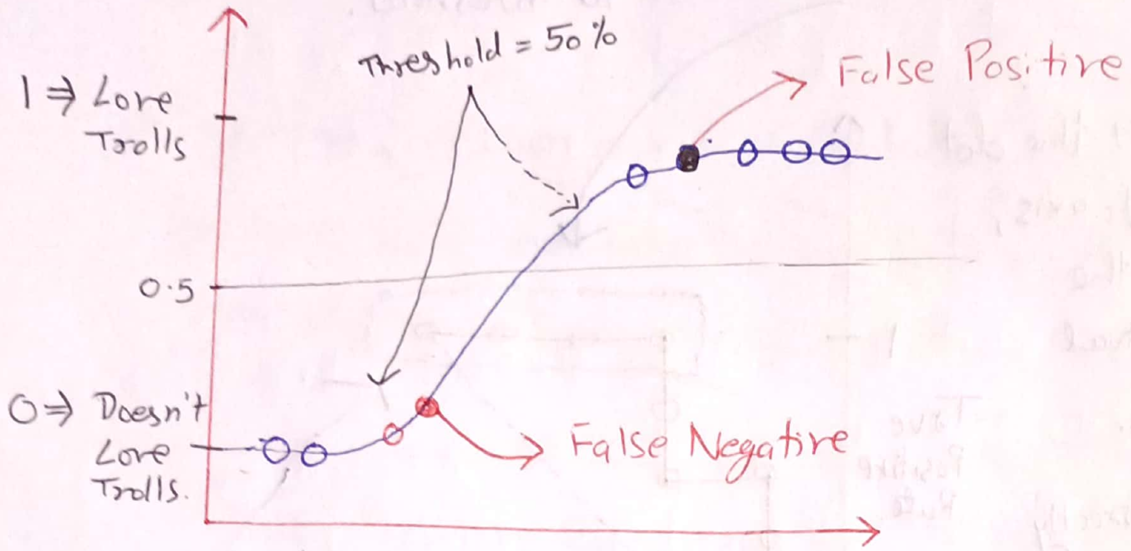


In this Example,
We measured the amount of Popcorn a bunch of People ate (grams).
Which is continous and Whether they
Love Trolls or Does Not Love Trolls, discrete.

The goal is to make a classifier that used the amount of Popcorn someone eats to classify whether or not they Love Trolls.

P.T.O

*) After using Logistic Regression to Predict →



*) To find correct identity, we ~~can~~ keep on changing threshold value and get different confusion Matrix.

Ultimately, we can try any classification threshold from 0 to 1.

and when we do, we end up with 10 Different Confusion Matrix (Matrices) that we can choose from.

*) The Threshold under each Confusion Matrix is just one of Many that will result in the exact ~~to~~ same Confusion Matrix.

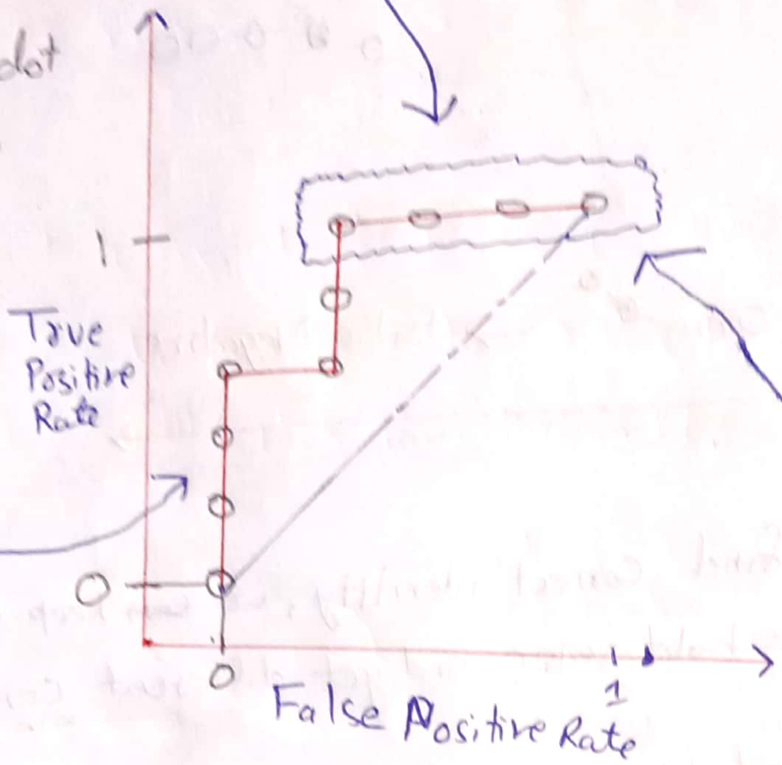
→ Try to find the ideal - Classification threshold among all ~~of~~ of these Confusion Matrices is tedious and annoying.

Wouldn't it be awesome, if we could consolidate them into one easy to interpret Graph? → that's when ROC graphs come.

*) Each black dot on the ROC Graph tells us the True Positive Rate and the False Positive Rate for a Specific Classification Threshold.

The higher the dot is along y-axis, the higher the % of actual Positives.

Where correctly Classified.



at a glance, we can look at the top row of points and tells that the Classification threshold that Resulted in the point on the left side Performed better than the others because they all have the same TRUE Positive Rate, but the Point on the Left has Low False Positive Rate.

Diagonal Line Shows, TRUE Positive Rate = False Positive Rate

*) To Construct ROC Graph, we will Start by using a Classification threshold 1, and Construct / Calculate Confusion Matrix.

Using that Confusion matrix, Calculate True Positive Rate and False Positive Rate. and then Plot that Point on ROC Graph.

$$\boxed{\text{TRUE Positive Rate (TPR)} = \frac{TP}{TP + FN}}$$

$$\boxed{\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}}$$

Now, lower the threshold (Such as 0.975, 0.965, ...) and Calculate Confusion matrix for that Particular Threshold.

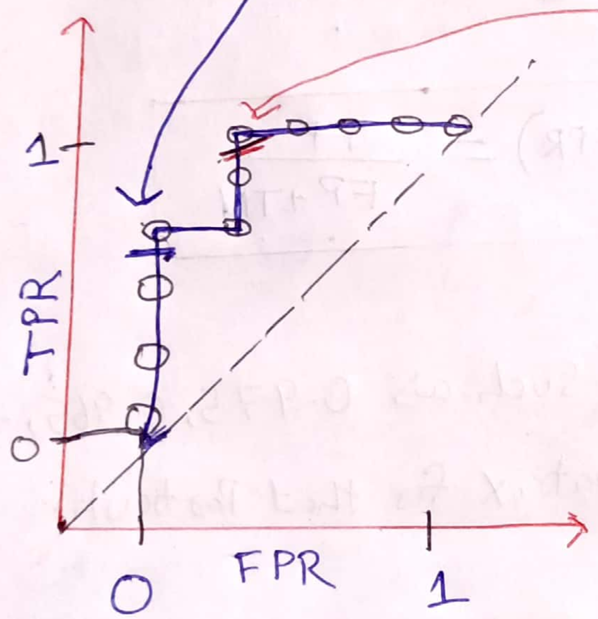
Thus, Calculate TPR, FPR and Plot the points on Graphs.

Likewise, For each threshold that increases the Number of Positive Classification, we calculate TPR and FPR until everyone is ~~as~~ classified as Positive after we finish plotting the points from each Possible Confusion Matrix, and connect the dots.

Now, without having to sort through a huge Pile of Confusion Matrices, we can uses the ROC graph to Pick a Classification threshold.

If we want to avoid all False Positives, but want to maximize the Number of actual Positives Correctly Classified, we would Pick this threshold.

--- but if we can tolerate a few False Positives, we would pick this threshold because it correctly Classifies all of the actual Positives.



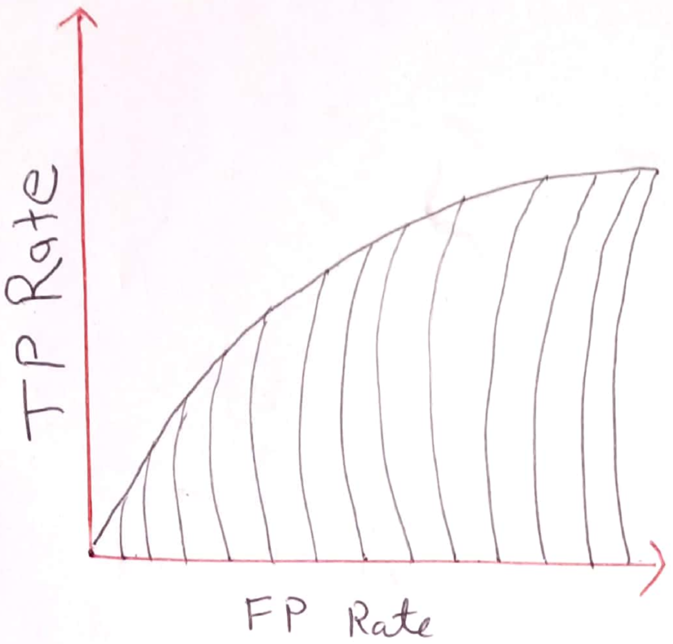
ROC Graphs are great for Selecting an Optimal Classification threshold for a Model.

But, what if we want to Compare how one perform Vs another??

~~this~~ This is where AUC (Area Under the Curve) Comes into Picture.

A) Area Under the Curve (AUC) :-

⇒ AUC measures the entire two Dimensional Area underneath the entire ROC curve from (0,0) to (1,1)



Note :- The More the area when compared to different model, better the Model.