

# Support Vector Machine

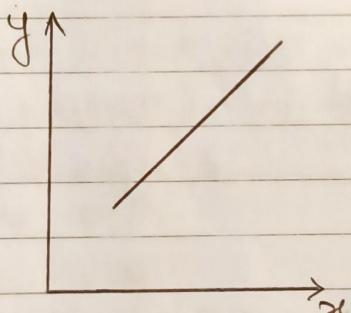
(SVM)

It can solve both classification and Regression problem.

1. classification → SVC (Support Vector classifier)

2. Regression → SVR (Support Vector Regressor)

some basics:



Equation of line:

$$y = mx + c \quad \text{OR}$$

$$y = \beta_0 + \beta_1 x \quad \text{OR}$$

$$ax + by + c = 0$$

$$\therefore y = \left( \frac{-a}{b} \right) x - \frac{c}{b}$$

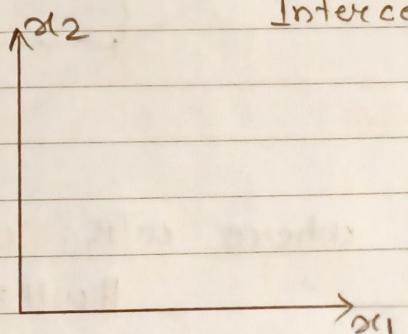
Coefficient      Intercept

$$ax_1 + bx_2 + c = 0$$

$$w_1 x_1 + w_2 x_2 + b = 0$$

$$\therefore w^T x + b = 0$$

If line passes through origin  $\therefore \underline{\underline{w^T x = 0}}$



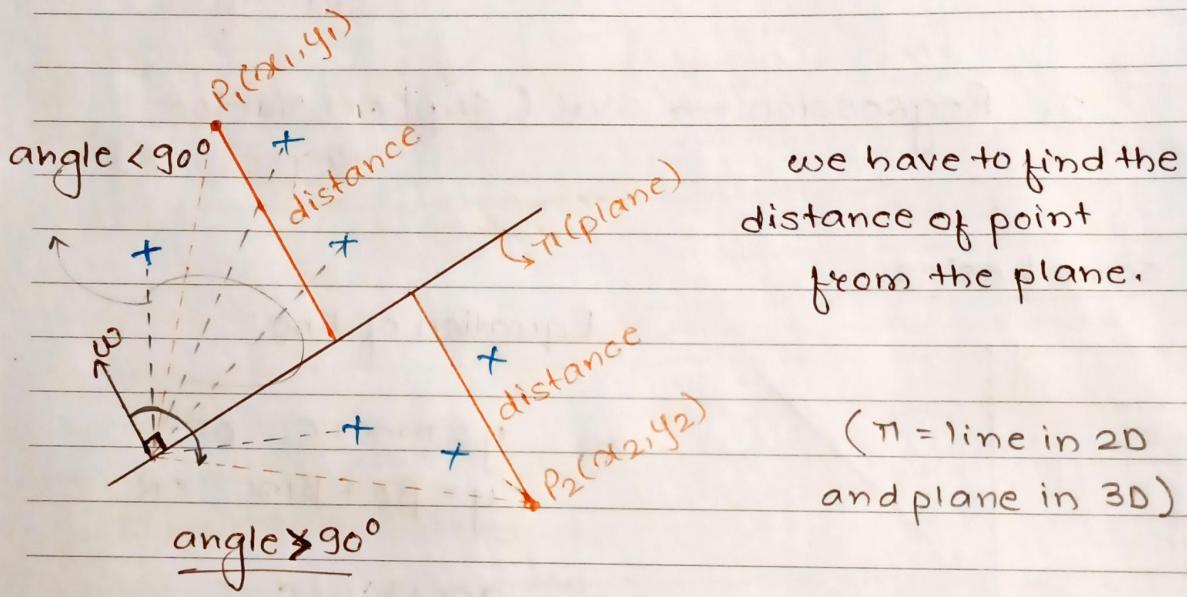
matrix multiplication:

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

$$w^T x + b = 0 \quad (w^T x: w \text{ Transpose } x)$$

Equation of line passing through origin is:

$$\underline{w^T x = 0}$$



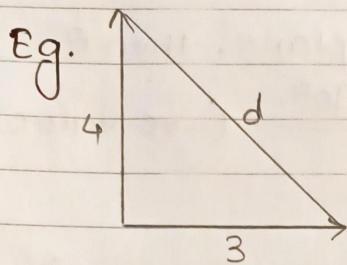
Distance of a point to the plane,

$$\text{distance}(d) = \frac{w^T p_i}{\|w\|}$$

$$= \|w\| \cdot \|p\| \cdot \cos \theta$$

where  $w^T p_i$ :  $w$  Transpose  $P_i$ ,  $w$ : vector,  
 $\|w\|$ : magnitude of  $w$ .

unit vectors: A vector which has a magnitude of 1 is basically called unit vector.



Now,

$$d = \sqrt{3^2 + 4^2} \\ = \sqrt{25}$$

$$\therefore d = 5$$

where vector,  $\hat{d} = \frac{d}{||d||}$

$||d||$  → magnitude

$$(3_{15}, 4_{15}) = d = \sqrt{(3_{15})^2 + (4_{15})^2} = \sqrt{25/25} \\ = 1.$$

∴ unit vector is a way to get focused on direction not on magnitude.

upward vector →

$$d = \frac{\omega \cdot P_1}{||\omega||}$$

$$d = ||\omega|| \cdot ||P_1|| \cdot \cos\theta$$

point above the plane

as  $\theta < 90^\circ$

$\cos\theta$  will always be +ve.

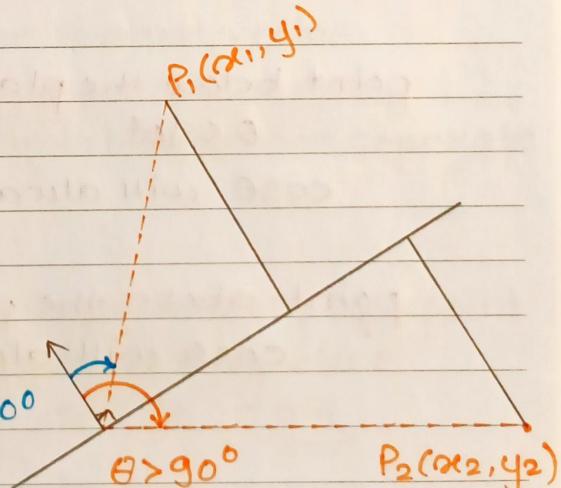
$\theta < 90^\circ$

$\theta > 90^\circ$

$P_2(x_2, y_2)$

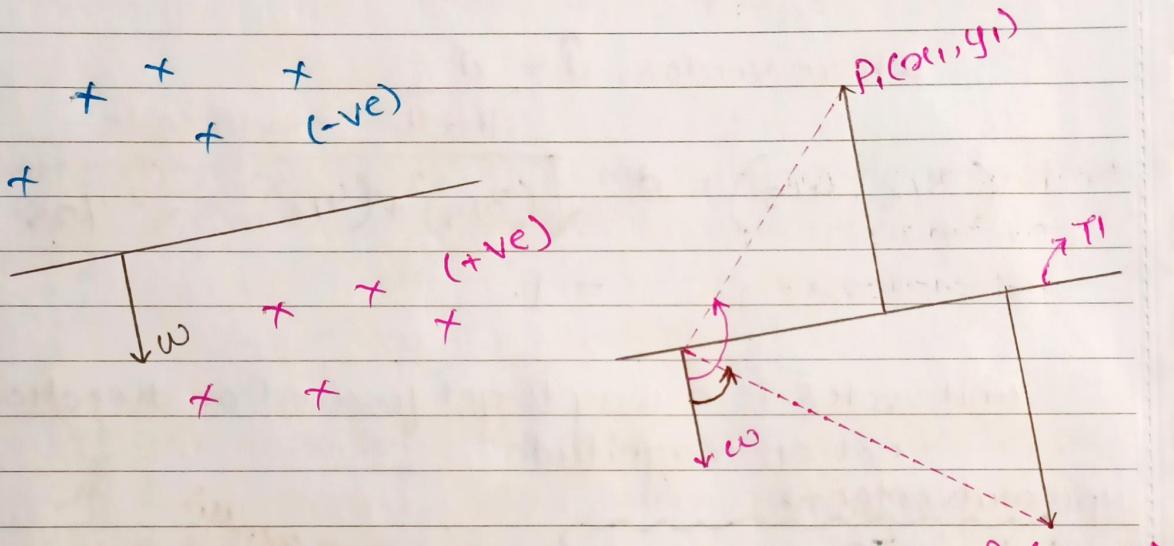
point below the plane, as  $\theta > 90^\circ$

$\cos\theta$  will always be -ve.



- If any point falling above the plane, then  $\theta$  must be less than  $90^\circ$  (+ve value)
- If any point falling below the plane, then  $\theta$  must be greater than  $90^\circ$ . (-ve value)

Downward vectors:



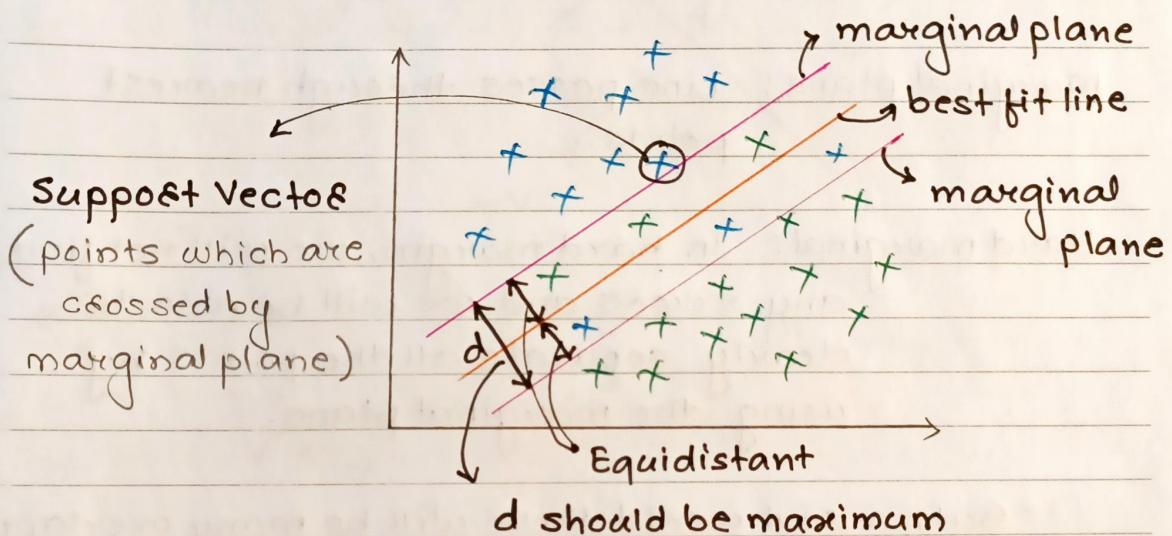
$\cos\theta$  will always be '+ve'.

point above the plane,  $\theta > 90^\circ$

$\cos\theta$  will always be '-ve'.

## Geometric Intuition Behind Support Vector Machine

### Support Vector classifier (SVC)



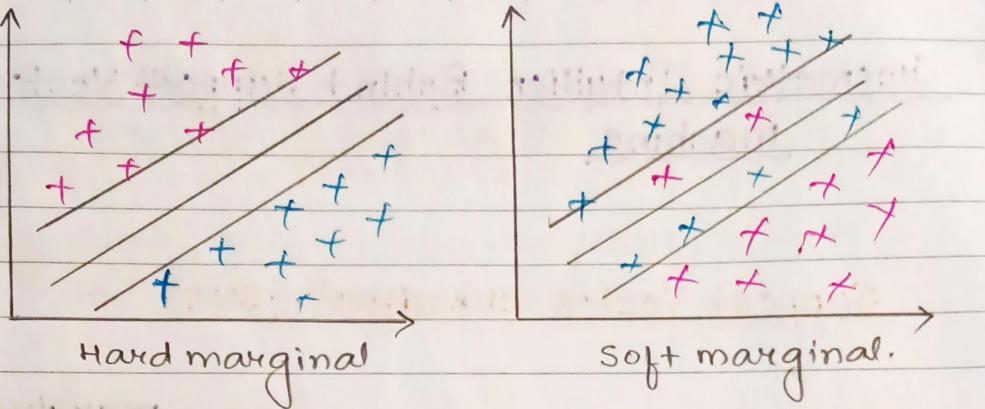
\* you can have more than one support vector.

There will be many possible hyperplanes that separate different classes.

We have learnt in LR that the probability of a point belonging to any class given at very close to the hyperplane will be close to 0.5.

So, we want a hyperplane that separates (+ve) pts and (-ve) pts as far away as possible.

Key idea of SVM: such hyperplane is called margin-maximizing plane.



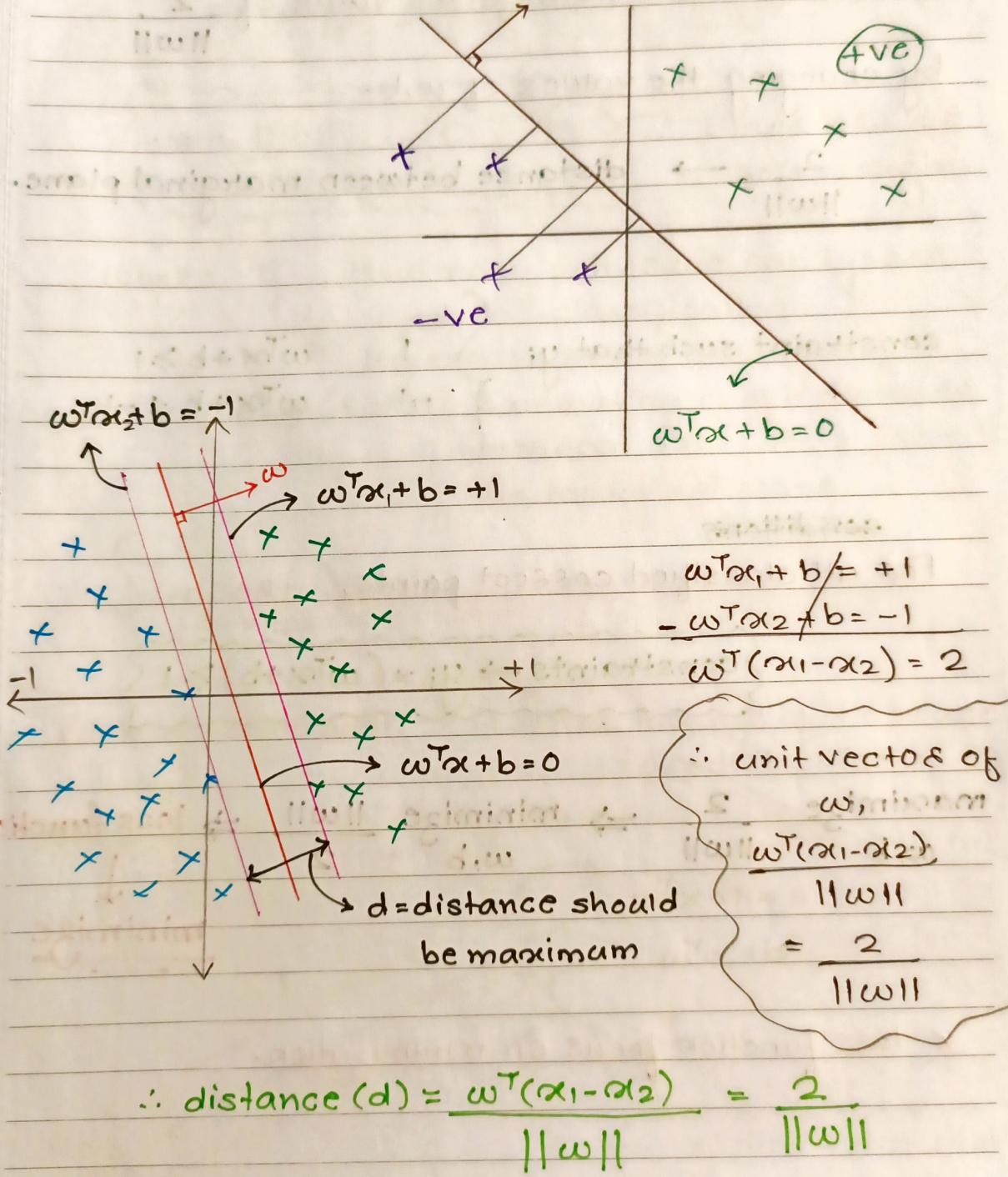
Marginal plane: Line passes through nearest points.

Hard marginal: In hard margin, we will not find any errors and we will be able to clearly separate all the points by using the marginal plane.

But, in real world there will be many overlapping with many errors, so marginal plane lines will be called soft margin.

→ marginal plane should be equidistance from best fit line.

## SVM Mathematical Intuition:



## Cost Function:

we have to maximize the value of  $\frac{2}{\|\omega\|}$

by changing the values of  $\omega, b$ .

$\frac{2}{\|\omega\|} \rightarrow$  distance between marginal plane.

constraint such that  $y_i$

$$\left\{ \begin{array}{l} 1 \quad \omega^T x + b \geq 1 \\ -1 \quad \omega^T x + b \leq -1 \end{array} \right.$$

## conditions

For all classified correct points,

$$\{ \text{constraints} \rightarrow y_i \times (\omega^T x + b) \geq 1 \}$$

maximize  $\omega, b \quad \frac{2}{\|\omega\|} \Rightarrow$  minimize  $\frac{\|\omega\|}{2} \Rightarrow$  loss function

$\downarrow$   
minimize

\* loss function focus on minimization.

## Cost function:

minimize  $\frac{\|w\|}{2}$  by changing  $w, b$ .

$$\min \frac{\|w\|}{2} + C_i \sum_{i=1}^n \xi_i$$

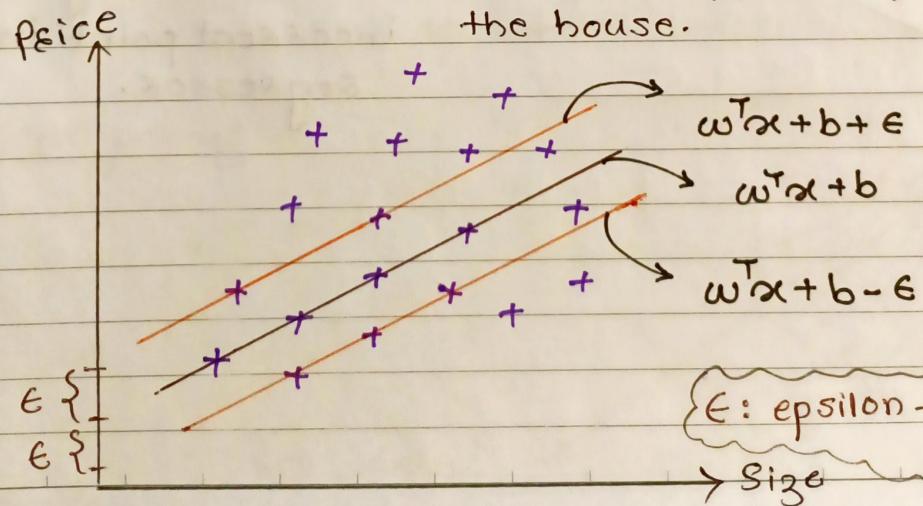
Hinge loss for soft margin.

where,  $C_i$ : How many points we can ignore  
for mis-classification.  
Hyperparameter

$\xi_i$  (Eta): Summation of the distance  
of incorrect data points from  
the marginal plane.

## Support Vector Regression:

Problem Statement: Based on the size of the house, we have to predict price of the house.



$\epsilon$ : epsilon  $\rightarrow$  marginal error

## cost Function:

$$\text{minimize}_{w,b} \frac{\|w\|}{2} + \left\{ C_i \sum_{i=1}^n \xi_i \right\} \rightarrow \text{Hinge loss}$$

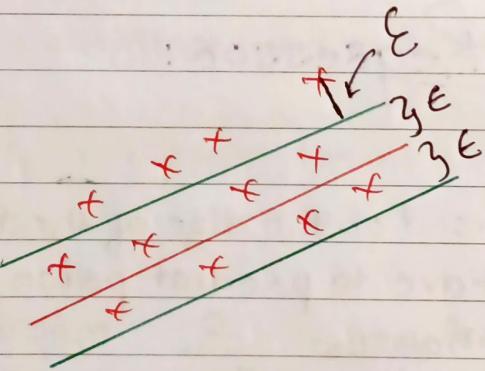
MAE

constraint:  $|y_i - w^T x_i| \leq \epsilon + \xi_i$

Truth point    predicted point                          epsilon

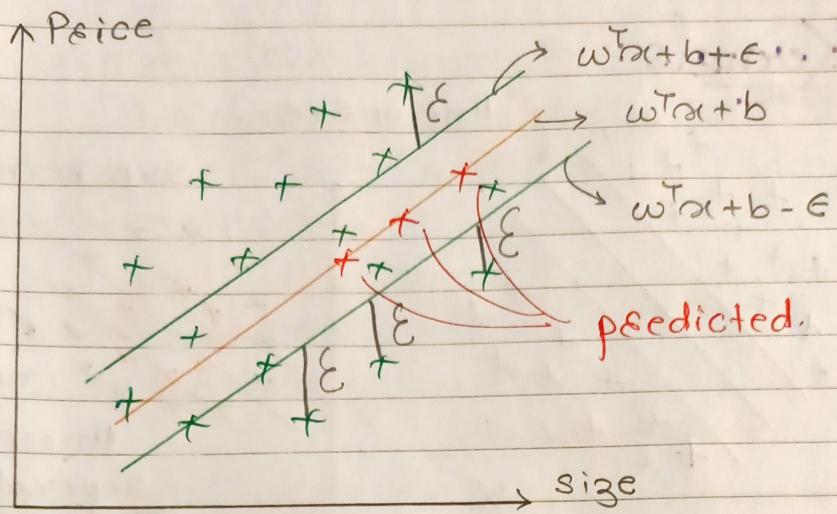
$\epsilon$ : margin of error (to decide original plane)

$\xi_i$ : errors above the margin.



## Hyperparameter:

- keep adjusting  $\epsilon$  to get best margin
- we can't say incorrect point in regressions.



In Regression, no complete incorrect value, because it will be continuous value.

Q.

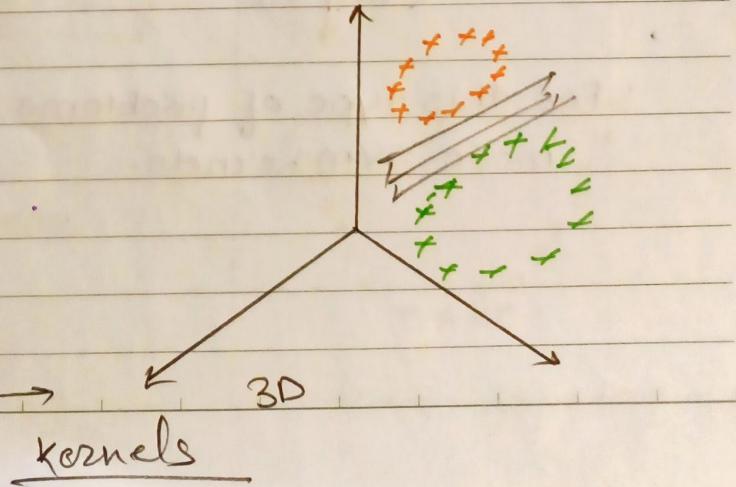
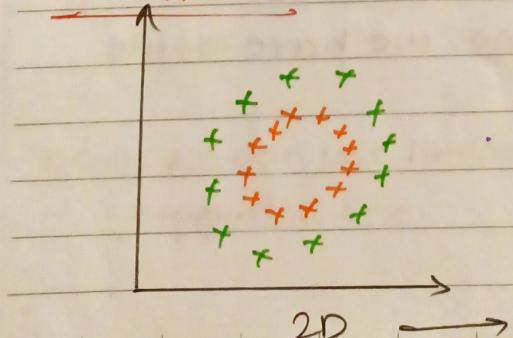
Is SVM impacted by the outliers?

Yes, SVM is impacted by the outliers.

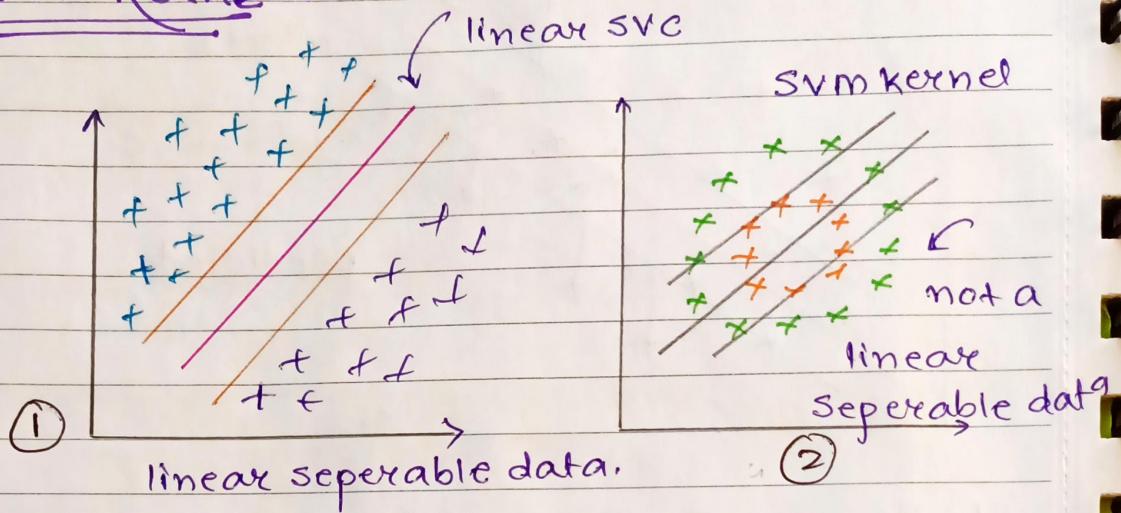
Does Standardization is need in SVM?

Yes, we need to perform Normalization and standardization.

SVM kernel:



## SVM Kernels:



- when we create this (1) type of best fit line and marginal plane, we are actually solving the linear separable data.
- → called as linear SVC. (Fig 1)
- If data is not a linear separable 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.

what does SVM kernels do?

→ The main aim is to apply some transformation technique. (some mathematical formula) on the dataset.

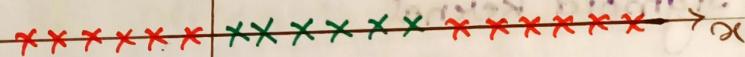
This transformation increases the dimension of the data.

SVM kernels → Transformation → Increasing the dimension of the data

(mathematical formula)

linearly separable line

→ will give error

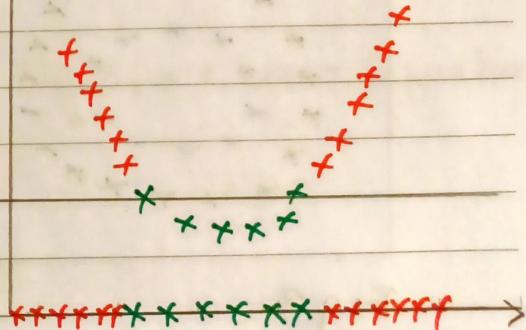


∴ we will transform the data from 1D  $\rightarrow$  2D.

$$y = x^2$$

After  $\longrightarrow$

Now, we can use linear separable line.



$$y = x^2$$

so if  $x = -7 \quad y = 49$   
 $x = -3 \quad y = 9$  and so on.

what is the advantage of doing this transformation?

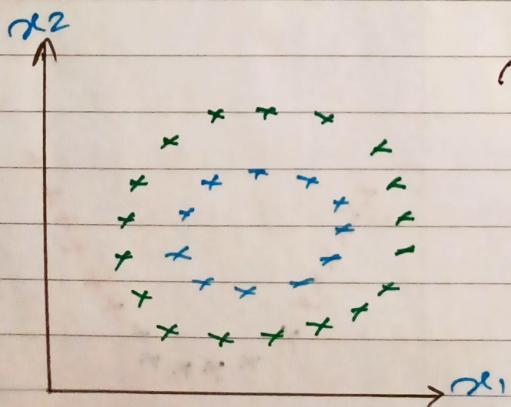
→ after transformation, we can apply linear SVM or SVC.

\* when we divide convert  $1D \rightarrow 2D$  then we can divide all the points using single line which is called Linear SVC.

### Types of SVM Kernel:

1. Polynomial kernel
2. RBF kernel
3. Sigmoid kernel

### 1. Polynomial kernel



not separable with best fit line.

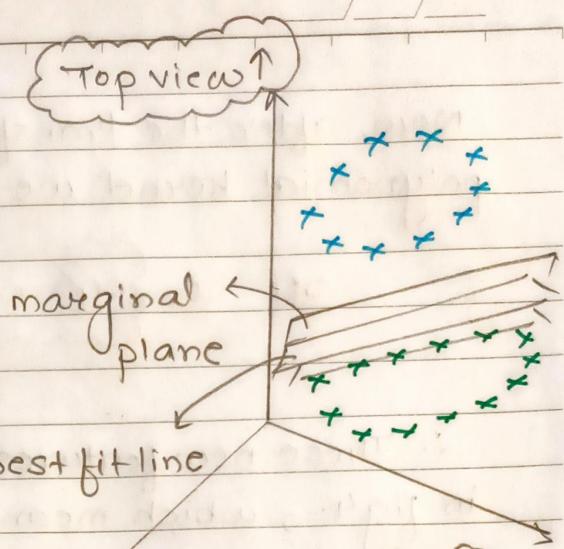
so, we need to convert  $2D \rightarrow 3D$ .

Fig 3.

Our main aim was to increase dimension,

$$2D \rightarrow 3D$$

so, hyperplane is created.



Fig(4)

Formula for Polynomial kernel:

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

$d = \text{dimension}$

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

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

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

3 unique values:  $x_1^2, x_1 \cdot x_2, x_2^2$

Now, initially at the time of Fig 3, we have 3 features

$$x_1 \ x_2 \ y$$

Now, after the transformation / formula of polynomial kernel we have 6 features

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

∴ 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  $x_2^2$

$z$  will be  $x_1 \cdot x_2$

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

→ use polynomial kernel, to get better accuracy.

## (2) Radial Basis Function Kernel (RBF Kernel)

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

hyperparameter

## (3) Sigmoid Kernel

It can be used as the proxy for neural networks.

$$k(x, x_i) = \tanh(6x^T x_i + \gamma)$$