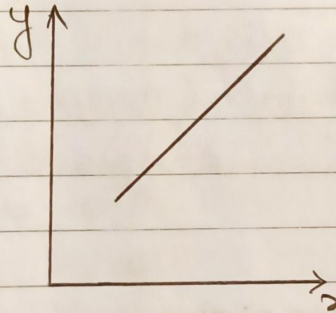
Support Vector Machine
(SVM)

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:

$y = mx + c$ OR

$y = \beta_0 + \beta_1 x$ OR

$ax + by + c = 0$

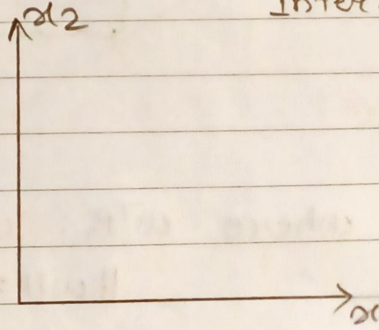
$\therefore \quad y = \overset{\text{coefficient}}{\left(\frac{-a}{b}\right)} x - \underset{\text{Intercept}}{\left(\frac{c}{b}\right)}$

$ax_1 + bx_2 + c = 0$

$w_1 x_1 + w_2 x_2 + b = 0$

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

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

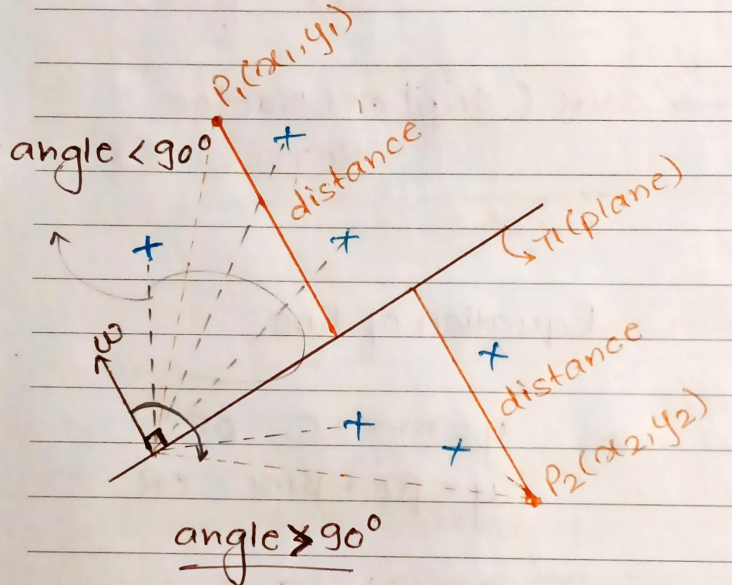


matrix multiplication: $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$

$w^Tx + b = 0$. (w^Tx: w Transpose x)

Equation of line passing through origin is:

$$\boxed{w^Tx = 0}$$



we have to find the distance of point from the plane.

(π = line in 2D and plane in 3D)

Distance of a point to the plane,

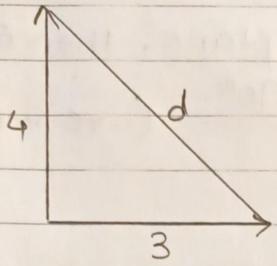
$$\boxed{\text{distance}(d) = \frac{w^TP_1}{\|w\|}}$$

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

where w^TP_1: w Transpose P_1, w: vector, $\|w\|$: magnitude of w.

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

Eg.



Now,

$$d = \sqrt{3^2+4^2}$$
$$= \sqrt{25}$$
$$\therefore \quad \underline{d = 5}$$

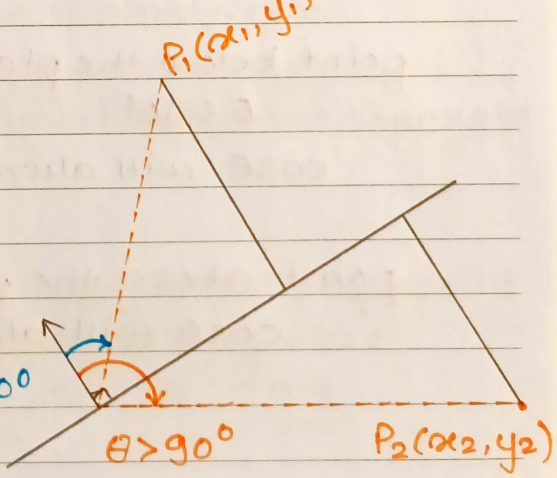
where vector, $\hat{d} = \dfrac{d}{\underbrace{\|d\|}_{\text{magnitude}}}$

$$\left(3/5, 4/5\right) = d = \sqrt{(3/5)^2 + (4/5)^2} = \sqrt{25/25}$$
$$= 1.$$

$\therefore$ unit vector is a way to get focused on direction not on magnitude.

upwardvector →

$$\boxed{\begin{aligned} d &= \frac{w^T P_1}{\|w\|} \\ d &= \|w\| \cdot \|P_1\| \cdot \cos\theta \end{aligned}}$$



point above the plane

as $\theta < 90^\circ$

cosθ will always be +ve.

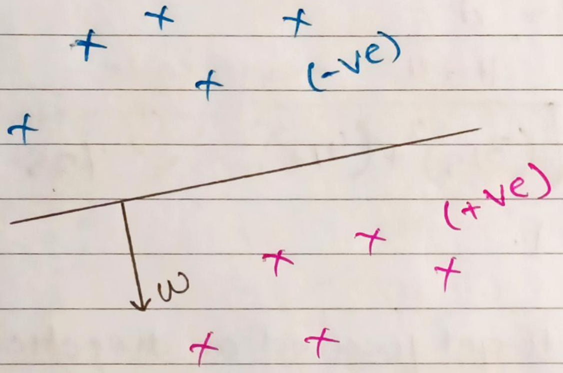
point below the plane, as $\theta > 90^\circ$

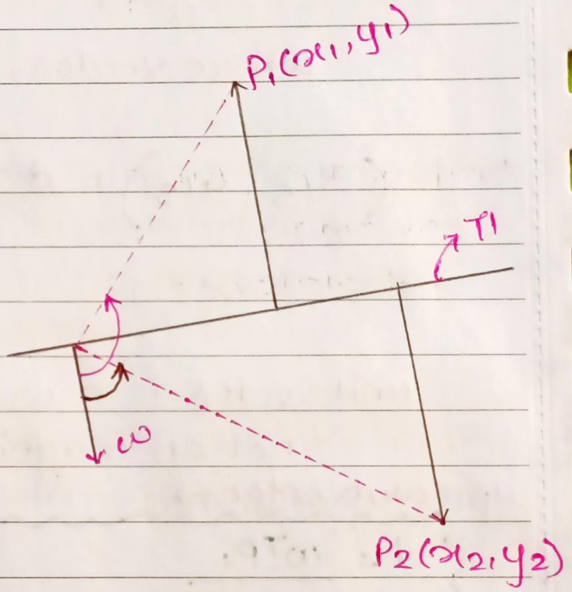
cosθ will always be -ve.

- If any point falling above the plane, then θ must be less than 90° (+value)

- If any point falling below the plane, then θ must be greater than 90°. (-ve value)

Downward vector:





point below the plane, $\theta < 90^0$

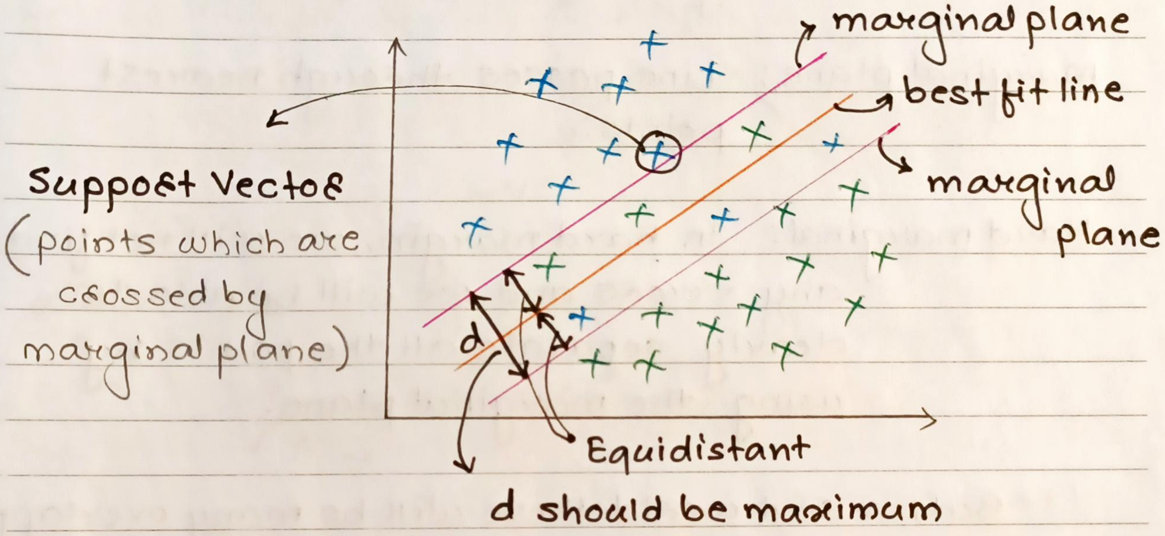
cosθ will always be '+ve'.

point above the plane, $\theta > 90^0$

cosθ will always be '-ve'.

Geometric Intuition Behind Support Vector Machine

Support Vector classifier (SVC)



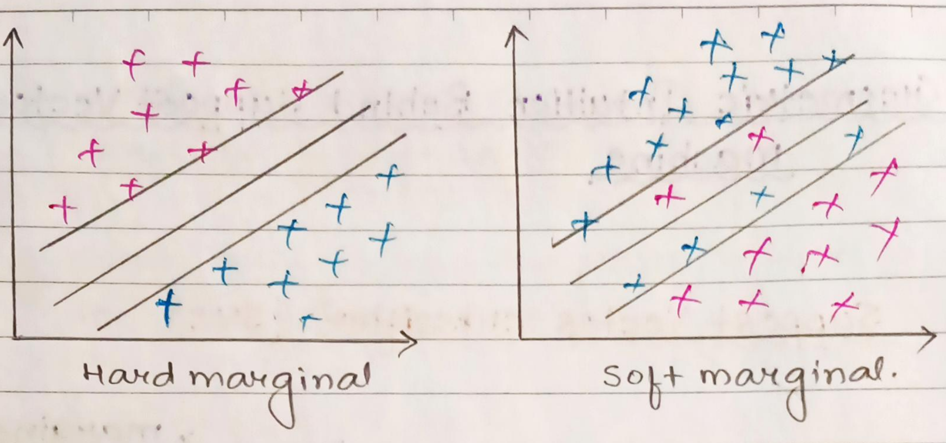
* you can have more than one support vector.

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 seperates (+ve) pts and (-ve) pts as far away as possible.

key idea of SVM, such hyperplane is called margin-maximizing plane.



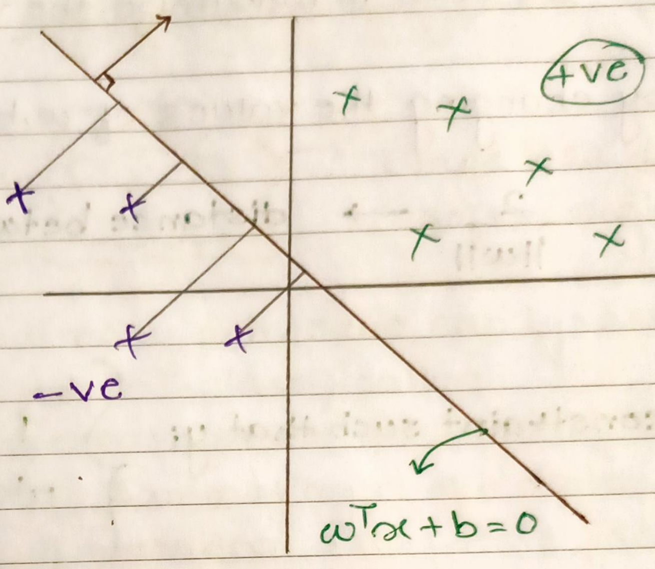
Marginal plane : Line passes through nearest points.

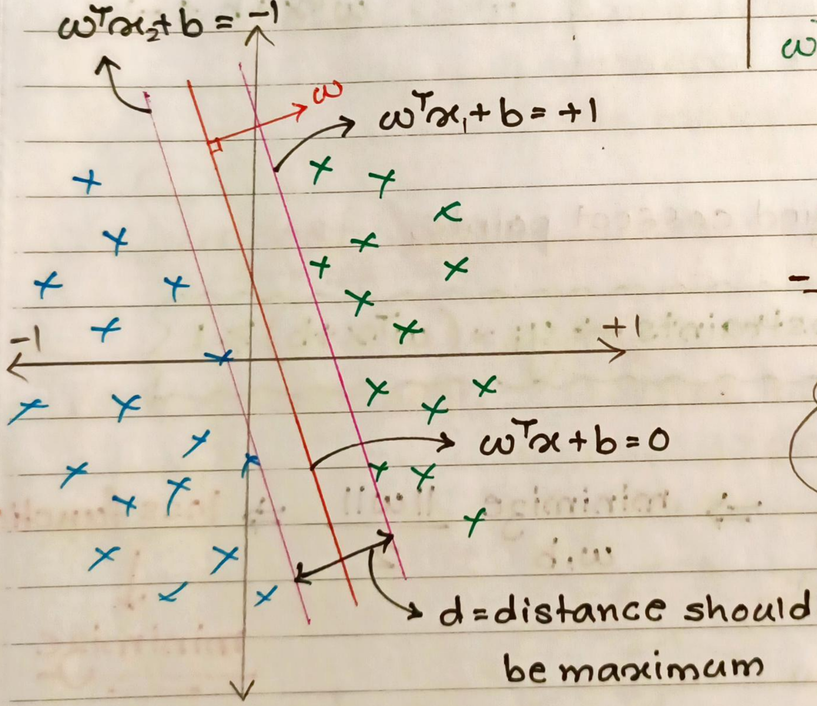
Hard marginal : In hard margin, we will not find any errors and we will be able to clearly separate all the points by using the marginal plane.

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:





$$\omega^T x_1 + \cancel{b} = +1$$
$$-\ \omega^T x_2 + \cancel{b} = -1$$
$$\overline{\omega^T (x_1 - x_2) = 2}$$

∴ unit vector of ω,

$$\frac{\omega^T (x_1 - x_2)}{||\omega||} = \frac{2}{||\omega||}$$

$$\therefore \text{distance } (d) = \frac{\omega^T (x_1 - x_2)}{||\omega||} = \frac{2}{||\omega||}$$

Cost Function:

we have to maximize the value of $\frac{2}{\|w\|}$ by changing the values of w, b.

$\frac{2}{\|w\|}$ → distance between marginal plane.

constraint such that $y_i \begin{cases} 1 & w^T x + b \geq 1 \\ -1 & w^T x + b \leq -1 \end{cases}$

~~conditions~~

For all classified correct points,

$$\boxed{\text{constraints} \rightarrow y_i * (w^T x + b) \geq 1}$$

$\underset{w,b}{\text{maximize}} \ \frac{2}{\|w\|} \Rightarrow \underset{w,b}{\text{minimize}} \ \frac{\|w\|}{2} \Rightarrow$ loss function

$\downarrow$

minimize

* loss function focus on minimization.

cost function:

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

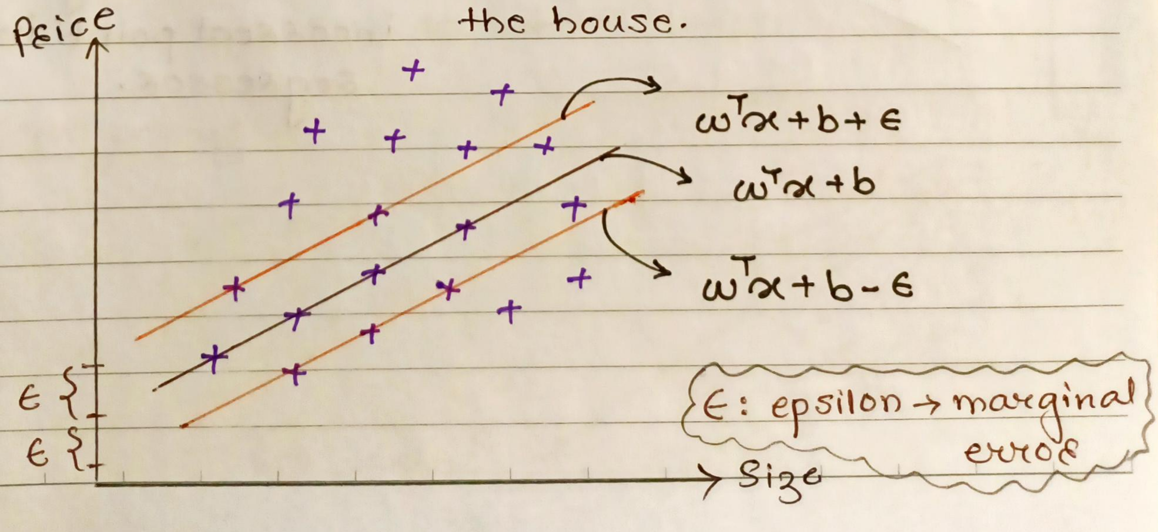
$$\boxed{\min \frac{\|w\|}{2} + \boxed{C_i \sum_{i=1}^{w} \xi_i}} \rightarrow \text{Hinge loss for soft margin.}$$

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

ξ_i (Eta): Summation of the distance of incorrect data points from the marginal plane.

Support Vector Regressor:

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



ϵ: epsilon → marginal error

cost Function:

$$\underset{\omega, b}{\text{minimize}} \; \frac{\|\omega\|}{2} + \boxed{C_i \sum_{i=1}^{\omega} \xi_i} \rightarrow \text{Hinge loss}$$

MAE

$$\boxed{\text{constraint: } |y_i - \omega_i x_i| \leq \epsilon + \xi_i} \rightarrow \text{Etaqi}$$

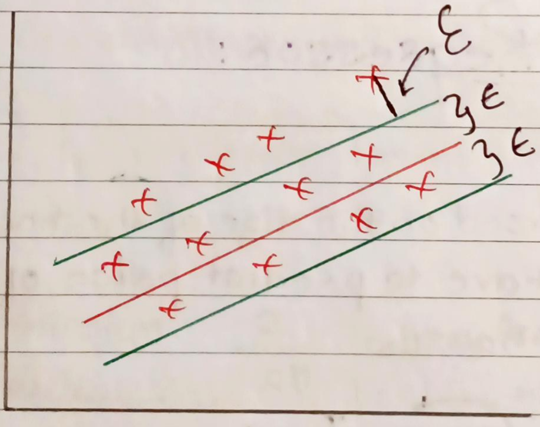
Truth point

predicted point

epsilon

ϵ : margin of error (to decide original plane)

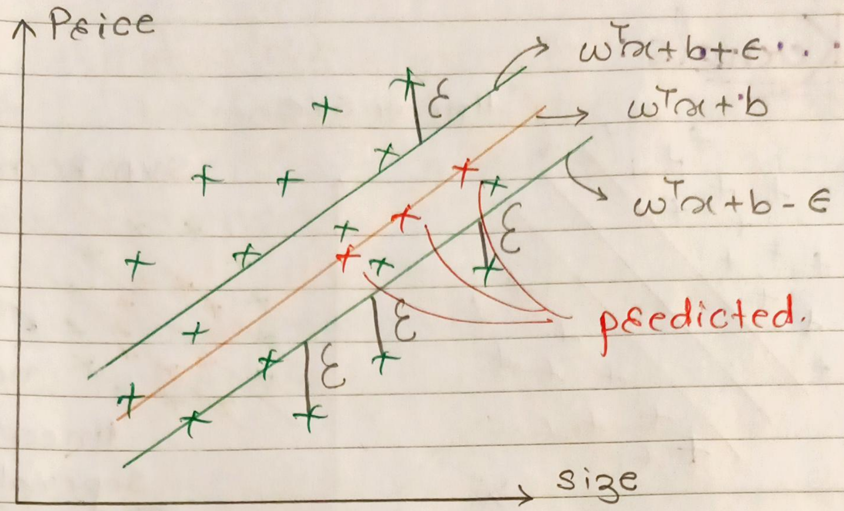
ξ_i : error above the margin.



Hyperparameter:

→ keep adjusting ε to get best margin

→ we can't say incorrect point in regressor.



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

Q.

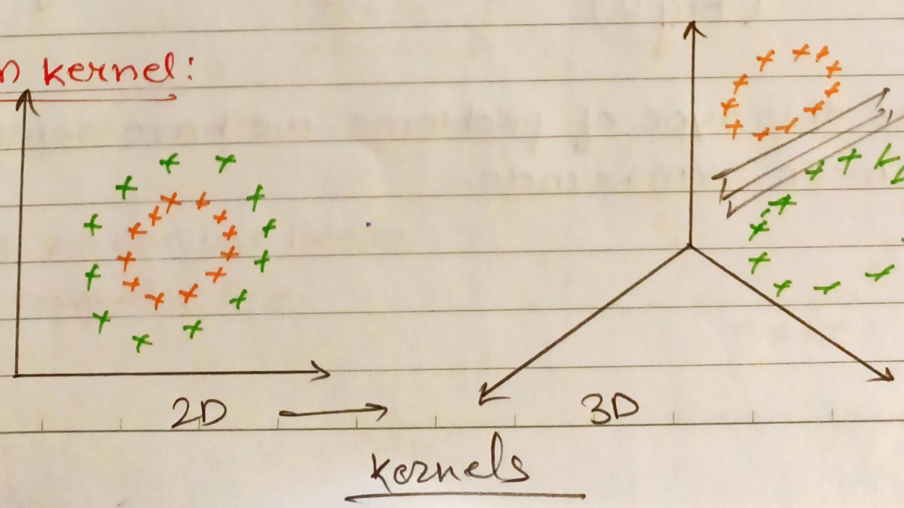
Is svm impacted by the outliers?

Yes, svm is impacted by the outliers.

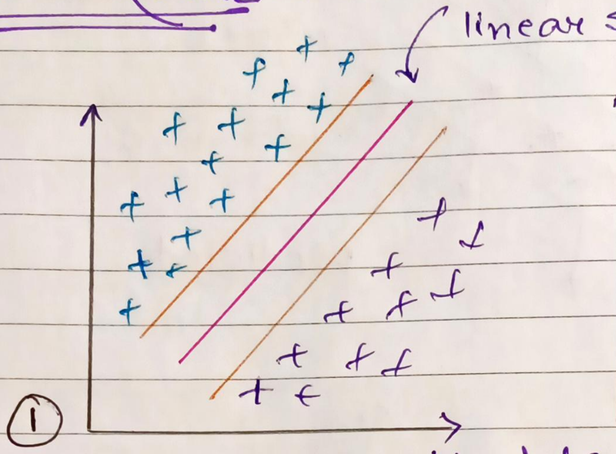
Does Standardization is need in svm?

Yes, we need to perform Normalization and standardization.

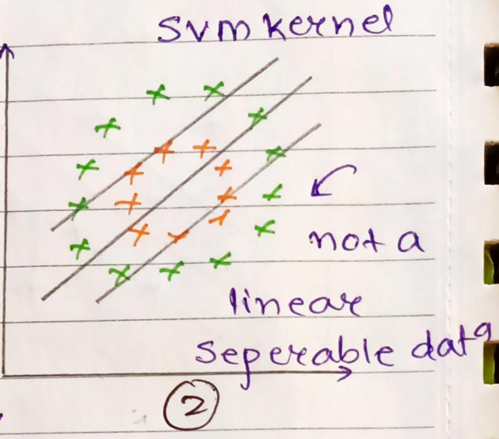
SVM kernel:



SVM Kernals :



linear seperable data.



- 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. (Fig①)

- 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.

what does SVM kernels do?

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

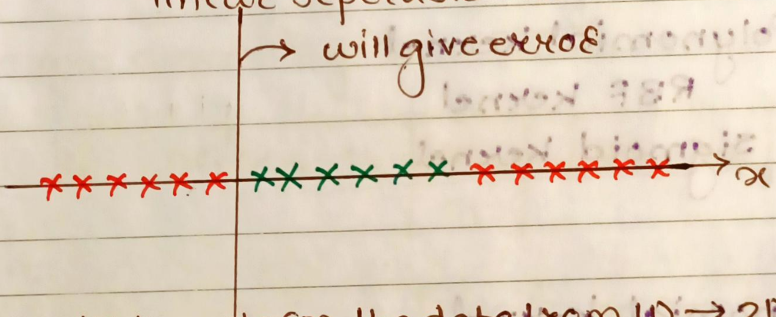
This transformation increases the dimension of the data.

(mathematical formula)

SVM kernels → Transformation → Increasing the dimension of the data

linear seperable line

→ will give error

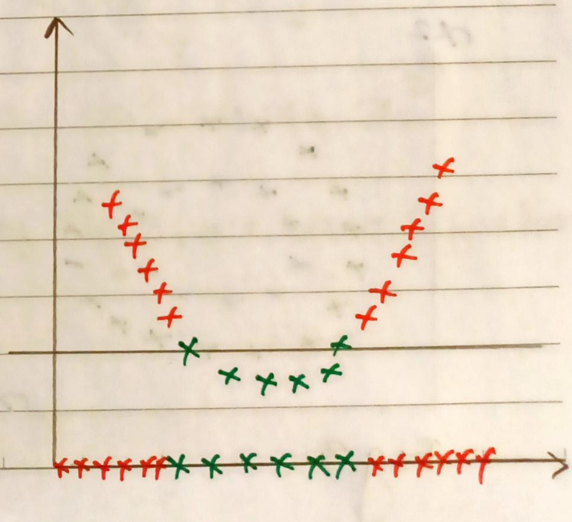


∴ we will transform the data from 1D → 2D.

$\boxed{y = x^2}$

After →

Now, we can use linear seperable line.



$y = x^2$

so if $x = -7$ $y = 49$

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

what is the advantage of doing this transformation?

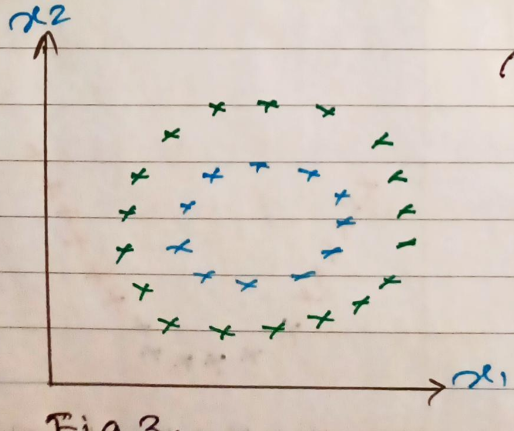
→ after transformation, we can apply linear SVM or SVC.

* when we ~~divide~~ convert 1D → 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 seperable with best fit line.

so, we need to convert 2D → 3D.

Fig 3.

Our main aim was to increase dimension,

$2D \rightarrow 3D$

so, hyperplane is created.

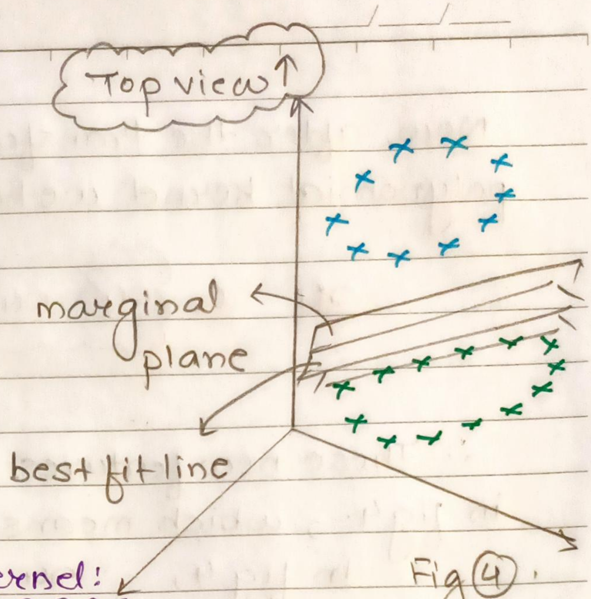


Fig ④.

Formula for Polynomial kernel:

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

d = dimension

If we are converting 2D→3D, the value of $\underline{\underline{d=3}}$.

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

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

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.

$x_1 \quad x_2 \quad y$

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

$$x_1 \quad x_2 \left\{ x_1^2 \quad x_1 \cdot x_2 \quad x_2^2 \right\} y$$

∴ These new features can be plotted as the 3D in fig4., 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 seperate the points.

→ use polynomial kernel, to get better accuracy.

② Radial Basis Function Kernel (RBF kernel)

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

hyperparameter

③ Sigmoid kernel

It can be used as the proxy for neural networks.

$$K(x, x_i) = \tanh\left(\sigma X^T X_i + r\right)$$