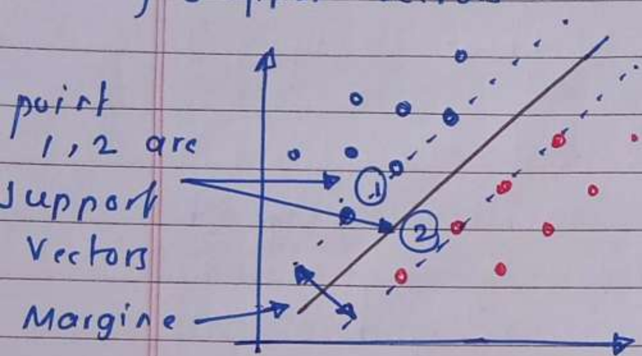
Support Vector Machine (SVM)

Basic about SVM.

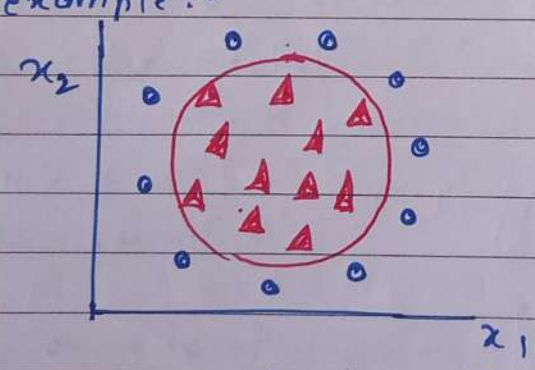
1) it is supervised ML model.
2) it can be used for both classification as well as regression but it is predominantly used for binary classification.
3) Hyperplane.
4) Support vectors



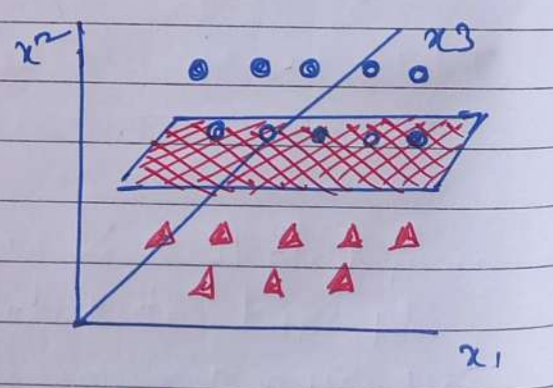
position of hyperplane depends upon support vectors.

- for 2D data it is easy to draw hyperplane but where data point can't seperated by line need to convert into 3D where we can seperate the datapoint by hyperplane

Example :-



{Here Not easy to seperate by line}



2D → 3D
and seperate by hyperplane

<u>Hyperplane</u> :- Hyperplane is line (in 2D) or plane that seperate the data point into two classes

Support vectors :- these are the datapoints which are nearest to hyperplane if these datapoints changes position of hyperplane changes.

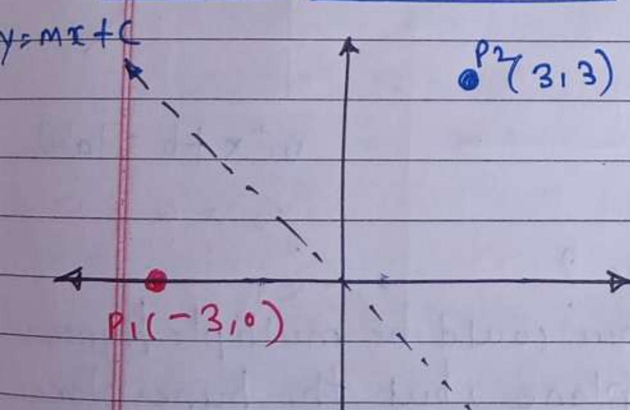
Advantages of SVM.

1) works fine with smaller dataset
2) works fine or efficiently where there is clear margin of seperation
3) works well with high dimensional data

Disadvantages

1) Not suitable for large dataset as training time would take very large.
2) Not suitable for noiser (outlier) dataset with overlapping classes.

Math Behind SVM.



let slope and intercept of hyperplane is,

$m = -1$

$c = 0$ {since passing through origine}

- let parameters of hyperplane save in W which is nothing but weight

$w \rightarrow (m, c) = (-1, 0)$

- let multiply x or P_1 by transpose of w

$$w^T x = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} -3 & 0 \end{bmatrix} = 3 \quad \text{(positive)}$$

[Note : why transpose? → for matrix multiplication no of column of 1st Matrix must be equal ~~row~~ no. of rows of second matrix]

- positive value indicates all the points ... of hyperplane will be positive class.

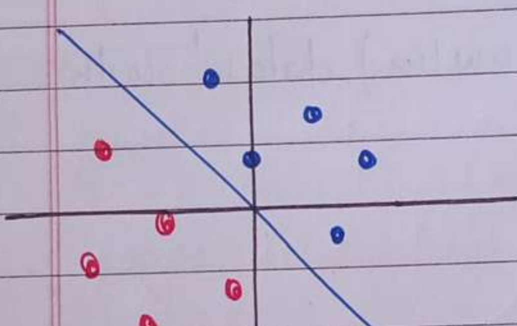
for $P_2(3,3)$

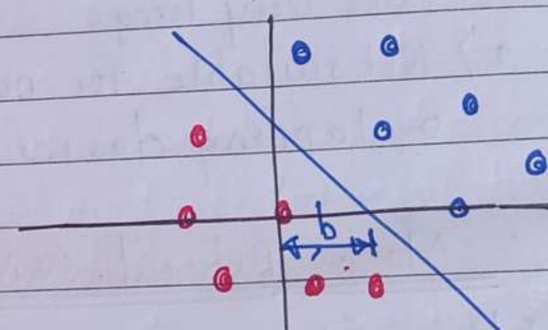
$$w^T x = \begin{bmatrix} -1 \\ 0 \end{bmatrix} [3, 3]$$

$$= -3 \quad (\text{Negative})$$

Here for all the points which lie on the right side of hyperplane will be belong to negative class.

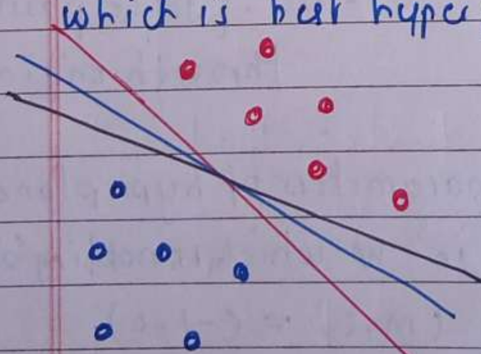
But Not all the time hyperplane will pass through Orgine.





$w^T x = \text{label}$ $\qquad$ $w^T x + b = \text{label}.$

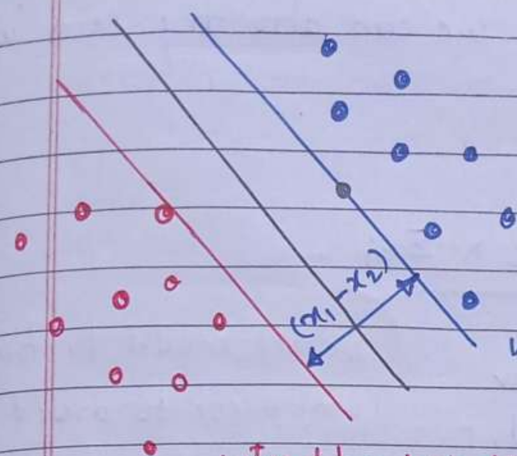
which is best hyperplane ?



there could be multiple hyper-plane, but the hyperplane with maximum margine size will be the best hyperplane.

→ optimization for maximum margine

$w^T x + b = \text{label}$



$w^T x + b = -1$

Equation of point or blue support vector & its output value any negative value.

$w^T x + b = 1 \Rightarrow$ this is equation of point or red support vector and its output value could be any positive value

to get margine let substract one from another.

$$\begin{aligned} w^T x_1 + b &= 1 \\ (-)\ w^T x_2 + b &= -1 \\ \hline w^T (x_1 - x_2) &= 2 \end{aligned}$$

$$w^T (x_1 - x_2) = 2$$

divide both side by $||w||$

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

$$(x_1 - x_2) = \frac{2}{||w||} \longleftarrow$$ this is nothing but magnitude of vector.

and

$$y_i = \begin{cases} -1 & w^T x_1 + b \leq -1 \\ 1 & w^T x_1 + b \geq 1 \end{cases} \quad \text{(label)}$$

so $\max\left(\frac{2}{||w||}\right)$ such that.

$$y_i = \begin{cases} -1 & w^T x_1 + b \leq -1 \\ 1 & w^T x_1 + b \geq 1 \end{cases}$$

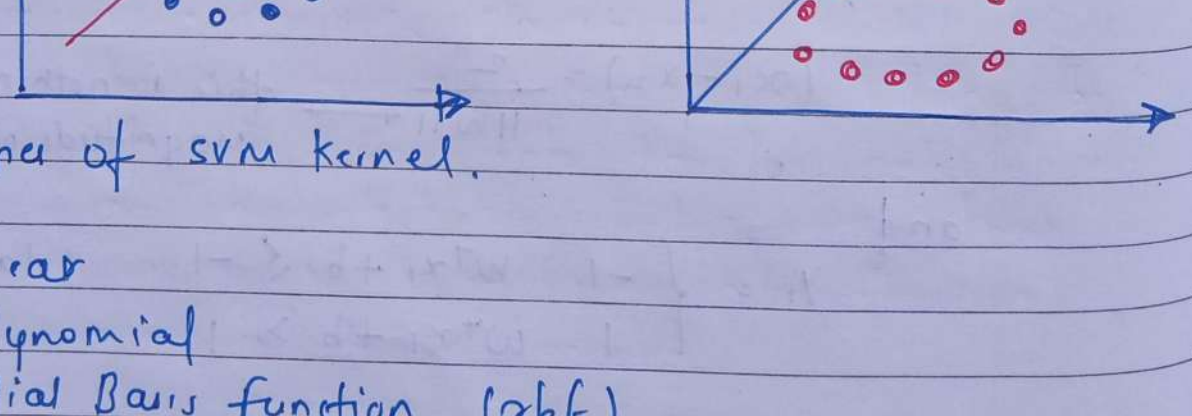
instead of using $\max_{\lambda}\left(\frac{2}{||w||}\right)$ we can also try Min which make better sense

$$\min\left(\frac{||w||}{2}\right) + \underbrace{c * \Sigma \varepsilon_i}$$

(we all model to train with some error to avoid overfitting (i.e it will be good and train and will bad for test data)

c = Number of error
ε_i = error magnitude

Kernels' in SVM : Generally function of the kernel is to transform the training set of data so that non-linear decision surface can be transformed to a linear equation in higher number of dimension space it return the inner product between two points in standard feature dimension



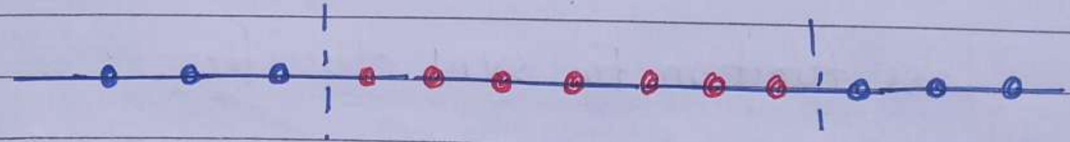
Types of SVM kernel.

1) Linear
2) polynomial
3) Radial Basis function. (rbf)
4) sigmoid.

Suppose feature (x)

x	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
x^2	36	25	16	9	4	1	0	1	4	9	16	25	36

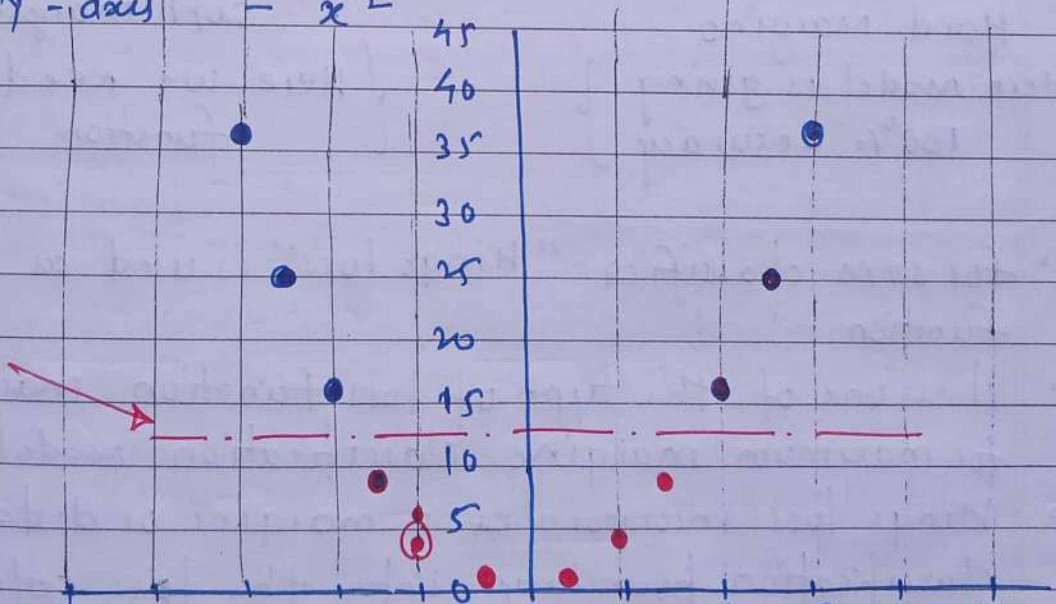
if you try to plot x on this 1D line.



- we can see None of the line could seperate the two class perfectly.
- that's why we add another feature which is function of x i.e x^2

x-axis — x

y-axis — x^2



Now this is separable data.

1) Linear kernel :- $K(x_1, x_2) = x_1^T, x_2$

{ best suitable for having too many features }

2) polynomial kernel.

$K(x_1, x_2) = (x_1^T x_2 + r)^d$

degree

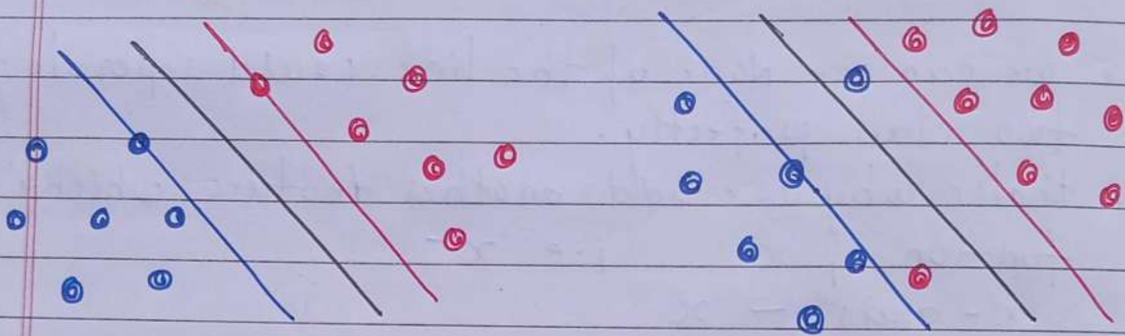
3.) radial basis function. (rbf kernel)

$$k(x_1, x_2) = \exp(-\gamma \cdot \|x_1 - x_2\|^2)$$

4. Sigmoid function

$$k(x_1, x_2) = \tanh(\gamma \cdot x_1^T x_2 + r)$$

Loss function for SVM classifier.



Hard Margine {Here model is giving 100% accuracy}

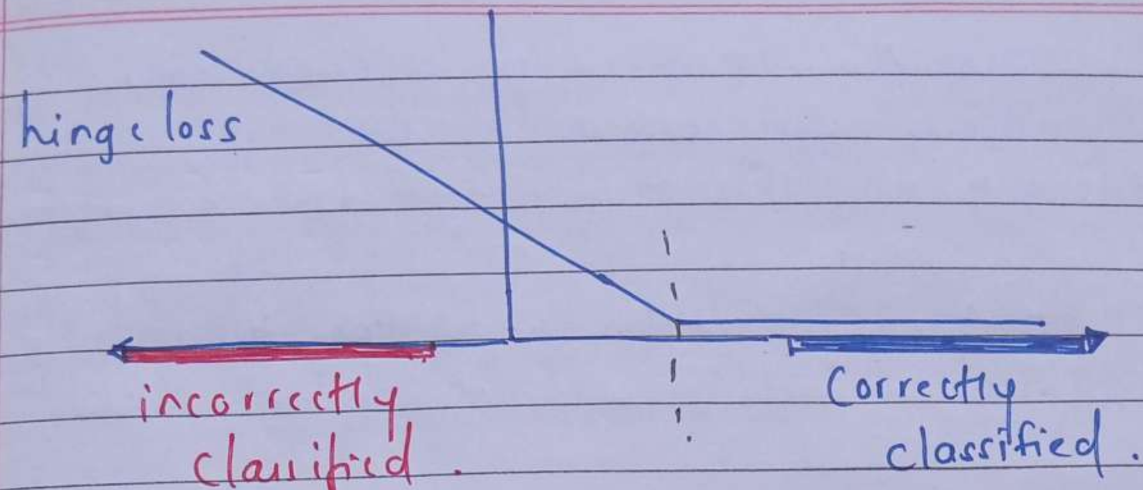
soft margine {Here we need loss function}

- for SVM classifier "Hinge loss" is used as loss function.
- it is one of the type of loss function mainly used for maximum margine classification model.
- Hinge loss incorporates a margine or distance from classification boundary into the loss calculation Even if new observation classified correctly they can incure penalty if the margine from decision Boundary is not large enough.

$$L = \max(0, 1 - y_i(w^T x + b))$$

0 - for correct classification

1 - for wrong classification.



let test for wrong misclassification.

$y_i = 1$, $\hat{y}_1 = -1$ $\qquad$ $y_i = -1$ $\quad \hat{y} = 1$

$L = (1 - 1(-1)$ $\qquad$ $L = (1 - (-1)(1)$

$= 1 + 1$ $\qquad$ $= 1 + 1$

$= 2$ $\qquad$ $= 2$

{ both are high loss value }

Now test for correct classification.

$y_i = 1$ $\quad \hat{y}_1 = 1$ $\qquad$ $y_1 = -1$ $\quad \hat{y}_1 = -1$

$(0 - (1)(1)$ $\qquad$ $(0 - (-1)(-1)$

$0 - 1$ $\qquad$ $0 - 1$

-1 $\qquad$ -1

{ both are low loss value }