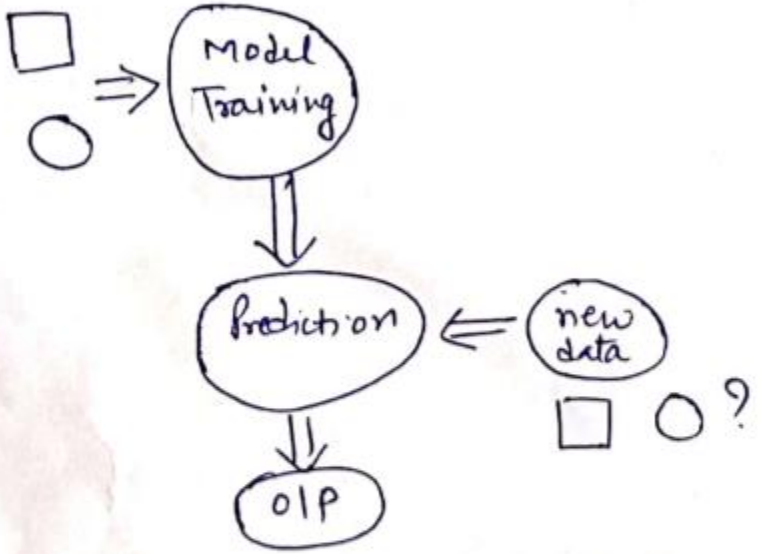
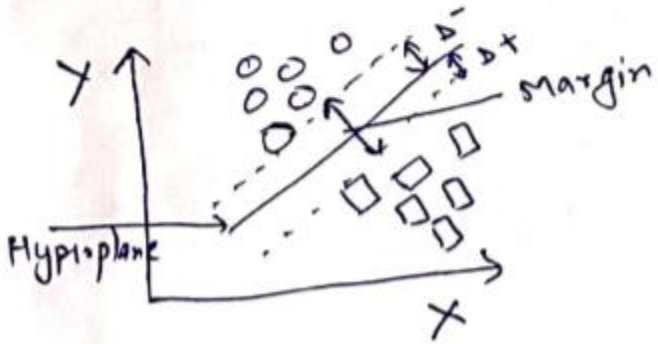
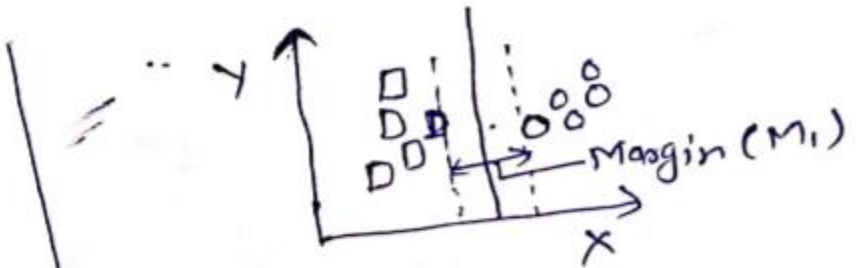
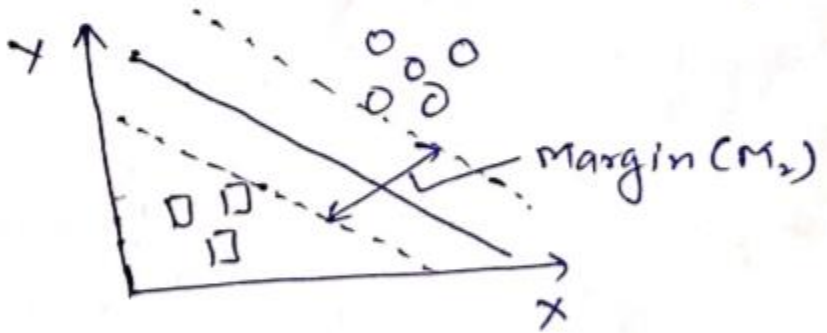
Support Vector Machine



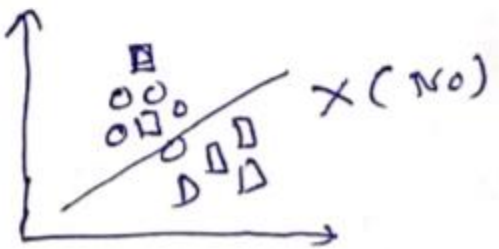




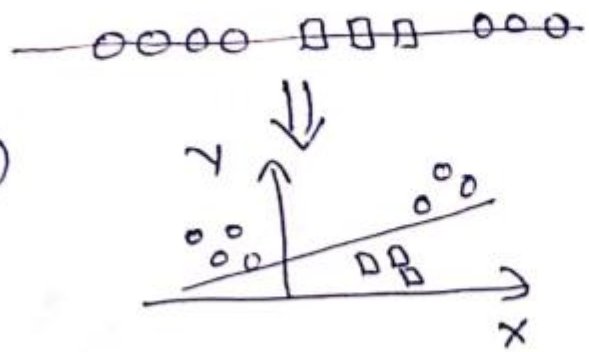
$M_2 >> M_1$



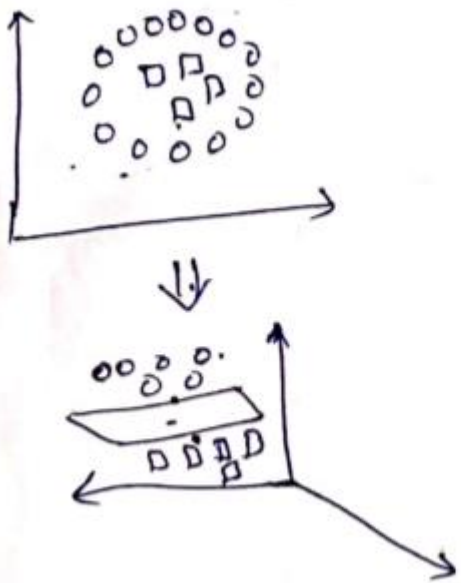
Maximum Margin Hyperplane



Non Linear SVM & Kernel function



(1)



(2)

L.D ⇒ Kernel ⇒ H.D

⇒ Kernel is used due to set of mathematical function used in SVM provides the window to manipulate data. So Kernel function generally transforms the training set of data so that a non-linear decision surface is able to ~~transformed~~ transformed to a linear function/ equation in a higher number of decision spaces.

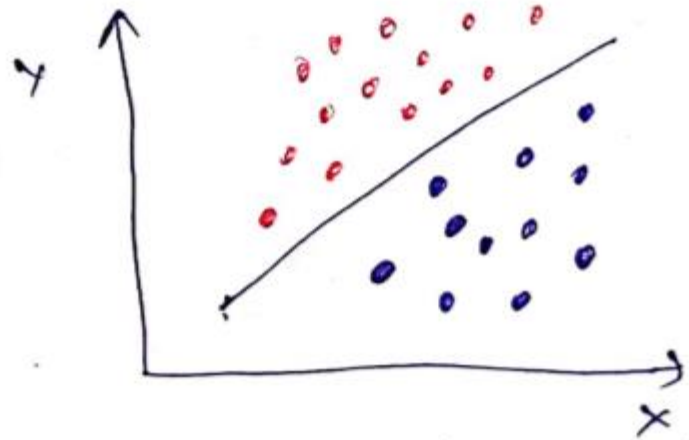
Types of kernel function

1) Linear kernel function

Linear kernel svm is used when the data is linearly separable i.e it can be seperated using a single line. It is one the most ~~most~~ common kernel to be used. It is most used when there are a large number of feature in particular data set.

Example

Where there are lot of features is Text classification, at each alphabet is a new feature. So we most use Linear kernel in Text classification.



In the above figure, there are two features "Blue" and the other "Red" features. Since there can be easily separated or in other words, they are linearly separable. So the Linear kernel can be used.

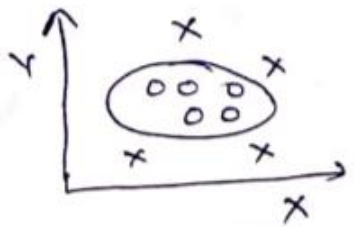
Note! It is most faster than other kernel functions.

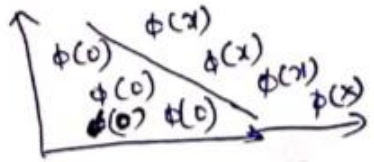
2) Polynomial kernel function

Polynomial kernel is a function commonly used in SVM & other kernelized models, that represents the similarity of vectors in a feature space over polynomial of the original variables, allowing learning of non-linear models.

$$K(x, y) = (x^T y + c)^d$$

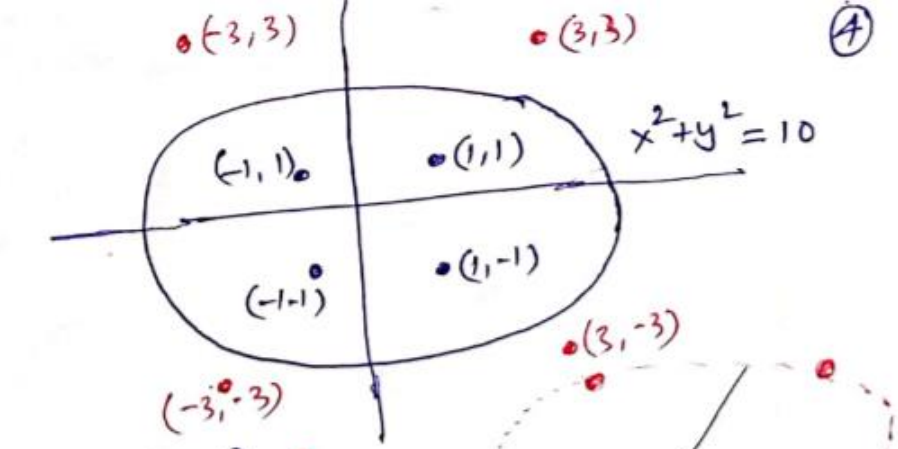
$x, y \rightarrow$ vectors in input space

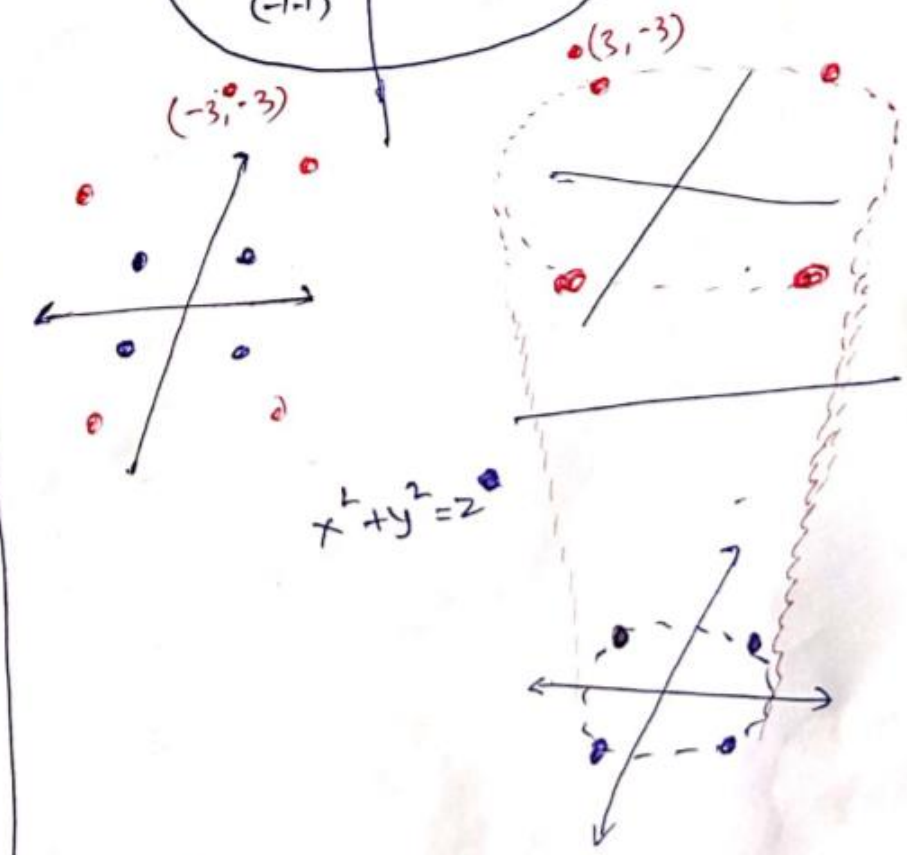


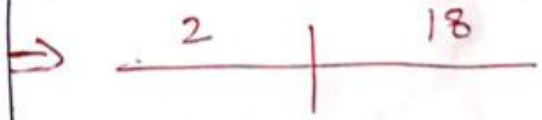


Ex

	-3,-3	3,-3	-3,3	3,3	-1,1	1,-1	-1,1	1,1
x+y	-6	0	0	6	-2	0	0	2
xy	9	-9	-9	9	1	-1	-1	1
x^2+y^2	18	18	18	18	2	2	2	2



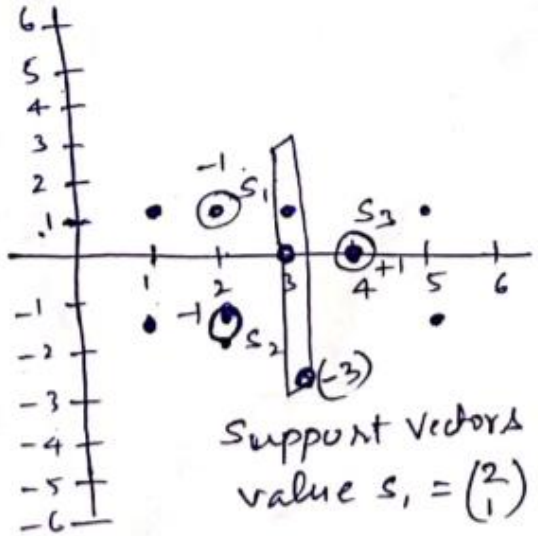




Hyperplane

I/P value (x_1, x_2), where x_1 & x_2 I/P value

(1,1), (2,-1), (1,-1), (2,-1), (4,0) (5,1) (5,-1) (6,0)



Support vectors $S_1, S_2 S_3$

value $S_1 = \begin{pmatrix}2\\1\end{pmatrix}$, $S_2 = \begin{pmatrix}2\\-1\end{pmatrix}$, $S_3 \begin{pmatrix}4\\0\end{pmatrix}$

Then Bias (1) into support vector

$$\bar{S}_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}, \bar{S}_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} \bar{S}_3 \begin{pmatrix}4\\0\\1\end{pmatrix}$$

Linear Eq $= \alpha_1 \bar{S}_1 \cdot \bar{S}_1 + \alpha_2 \cdot \bar{S}_1 \bar{S}_2 + \alpha_3 S_1 S_3 = -1$

$= \alpha_1 \bar{S}_1 \cdot \bar{S}_2 + \alpha_2 \cdot \bar{S}_2 \bar{S}_2 + \alpha_3 S_2 S_3 = -1$

$= \alpha_1 \bar{S}_1 \cdot \bar{S}_3 + \alpha_2 \bar{S}_2 \bar{S}_3 + \alpha_3 \bar{S}_3 \cdot \bar{S}_3 = +1$

$\Rightarrow$

$$\left.\begin{aligned} 6\alpha_1 + 4\alpha_2 + 9\alpha_3 &= -1 \\ 4\alpha_1 + 6\alpha_2 + 9\alpha_3 &= -1 \\ 9\alpha_1 + 9\alpha_2 + 17\alpha_3 &= +1 \end{aligned}\right\} \begin{aligned} &= -3.25 \\ &= -3.25 \\ &= 3.5 \end{aligned}$$

(weigh_vector)

Then weight $\bar{w} = \sum \alpha_i \cdot \bar{S}_i$

$$= -3.25 \begin{pmatrix}2\\1\\1\end{pmatrix} + (-3.25) \begin{pmatrix}2\\-1\\1\end{pmatrix} + 3.5 \begin{pmatrix}4\\0\\1\end{pmatrix} = \begin{pmatrix}1\\0\\-3\end{pmatrix}$$

(optimal) Hyperplane .

weigh value = $\begin{pmatrix}1\\0\end{pmatrix}$

Bias = (-3)

$b + 3 = 0$

$y = w^T x + b \le -1$ — (1)

$y = \begin{pmatrix}1\\0\end{pmatrix} \bar{x}_1 + (-3) \le -1$

$y \equiv \begin{pmatrix}1\\0\end{pmatrix} \bar{x}_1 + (-3) > 1$

$y = w^T x + b > 1$ — (2)

$y = w^T x + b = 0$ — (3)

Distance

Distance b/w S_3 & hyperplane

$a = S_3 = (4,0)$

$w = (1,0)$

length of weight $= \sqrt{1^2 + 0^2} = \sqrt{1^2} = 1$

weigh value $\left[\frac{w}{\|w\|}\right] = \left(\frac{1}{1}, \frac{0}{1}\right) = (1,0)$

distance b/w support vector

$u = (-1, 0)$

$a = (4, 0)$

$P = (u.a)\, u$

$= (-1 \times 4 + 0 \times 0)\, u$

$= (-4 + 0)\, u$

$= \underline{\underline{-4u}}$

$P = -4u$

$= -4(-1 \times 1, -4 \times 0)$

$= (4, 0)$

Distance of point $= \sqrt{4^2 + 0^2} = \sqrt{4^2} = 4$

Maximum margin

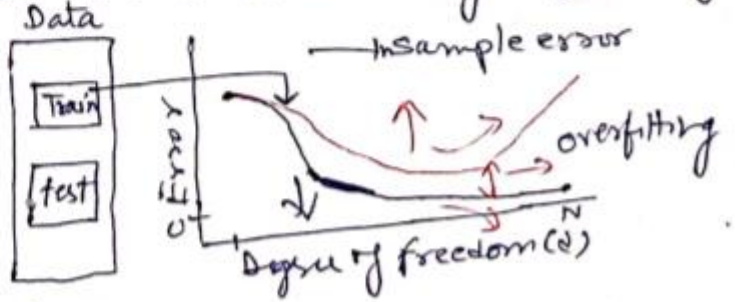
$= 2P$

$= 2 \times 4$

$= \underline{\underline{8}}$

Properties of SVM

1) Flexibility in choosing a similarity function
2) Sparseness of soln when dealing with large data sets
   + only support vectors are used to specify the seperating hyperplane
3) Ability to handle large feature space
   + complexity does not depend on the dimensionality of the feature space.
4) Overfitting can be controlled by soft-margin approach.



5) Feature selection

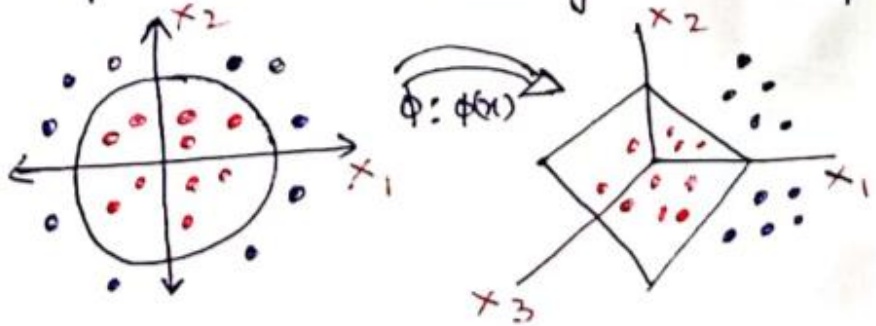
Issues in SVM

1) SVM algorithm is not suitable for large data sets
2) SVM does not perform very well, when the data set has more noise i.e target classes are overlapping
3) As the support vector classifier works by putting data points, above & below the classification hyperplane there is no probabilistic explanation for the classification.

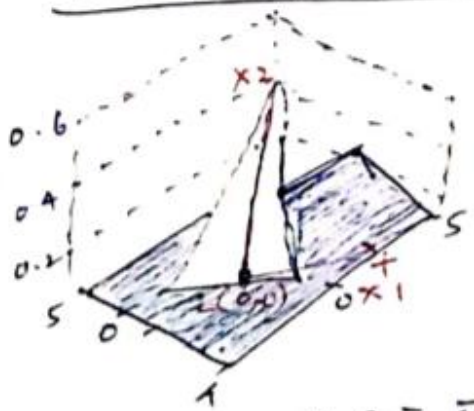
SVM for Non-linear classification

→ General idea:

The original I/p space can always be mapped to some-dimensional feature space where training set is separable

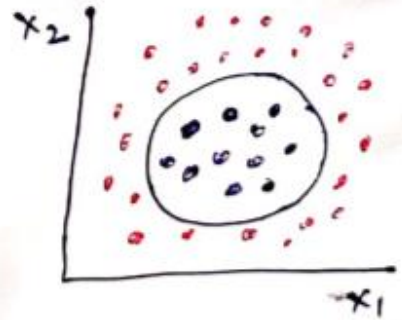


→ Gaussian Kernel Function [K(x,v)]



As the distance from the land mark i.e point (0,0) decreases, the kernel function tends to '1', As we go further from landmark, the kernel function tends to '0'.

$$K(\bar{x}, \bar{l}^i) = e^{-\frac{|\bar{x} - \bar{l}^i|^2}{2\sigma^2}}$$



$$K(\bar{x}, \bar{l}^i) = e^{-\frac{|\bar{x} - \bar{l}^i|^2}{2\sigma^2}}$$

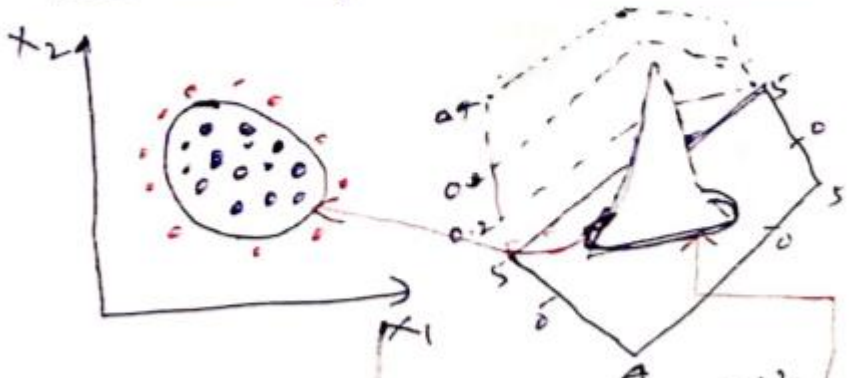
In this formula, if the kernel fn = 0, then assign class 0, if it is >0 then assign class 1. The points inside the curve are classified as blue color (1) and all the points in the purple zone are classified as red (0)

→ ~~Relevance of standard deviation~~

(σ)

if (σ) ~~indicate~~ ↑ means

Relevance of Standard deviation (σ)

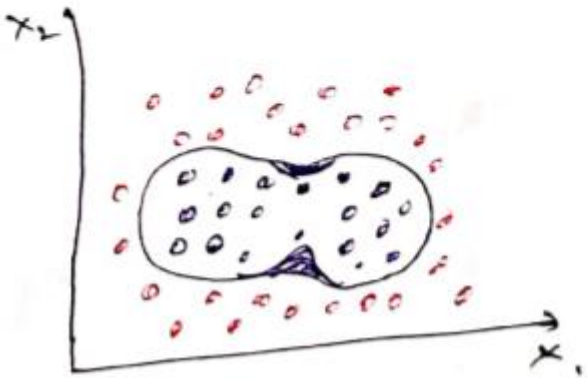


$$K(\bar{x}, \bar{l}^1) = e^{-\frac{|\bar{x} - \bar{l}^1|^2}{2\sigma^2}}$$

A high stand deviation will mean high circumference. This will imply more points being classified as (1)

$\Rightarrow$ σ is high

Gaussian kernel function to create non-linear, complex boundary decision



$K(\bar{x}, \bar{l}^1) + K(\bar{x}, \bar{l}^2)$ → Simplified formula

Blue color

$K(\bar{x}, \bar{l}^1) + K(\bar{x}, \bar{l}^2) > 0$

Red color

$K(\bar{x}, \bar{l}^1) + K(\bar{x}, \bar{l}^2) = 0$