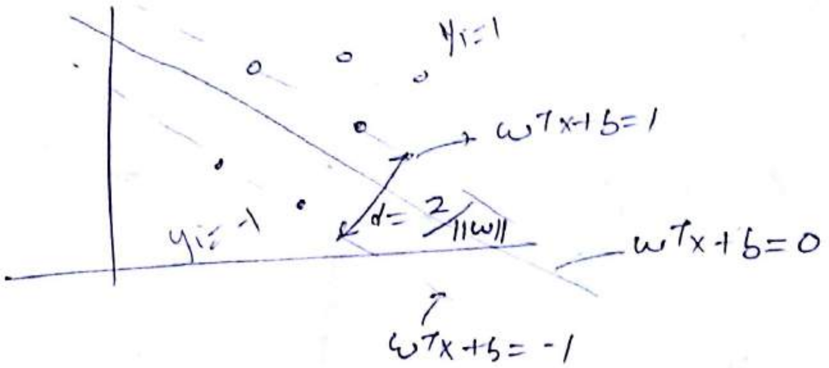
Q1

Let $\{x_1, x_2, \ldots x_n\}$ be out data points

& $y_i \in \{1,-1\}$ be the class label of x_i



$y_i (w^T x_i + b) \geq 1$
$\forall (x_i, y_i)$

For $y_i=1$. Now decision boundary can be found as

$$\text{minimize } \frac{1}{2} \|w\|^2$$
$$\text{subject to } y_i (w^T x_i + b) \geq 1$$

So Lagrangian is

$$L = \frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i \left(1 - y_i (w^T x_i + b)\right)$$

$\alpha_i \geq 0$

Differentiating w.r.t. w & b

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w - \sum \alpha_i y_i x_i = 0$$
$$\Rightarrow \bar{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \underline{\sum \alpha_i y_i = 0}$$

So If we substitute $\bar{w} = \sum \alpha_i y_i x_i$ to L

we get

$$L = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i^T \sum_{j=1}^{n} \alpha_j y_j x_j + \sum \alpha_i \left(1 - y_i \left(\sum_{j=1}^{n} \alpha_j y_j x_j^T x_i + b\right)\right)$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{n} \alpha_j y_j x_j^T x_i - b \sum \alpha_i y_i$$

$$= -\frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i$$

since $\sum_{i=1}^{n} \alpha_i y_i = 0$

So the dual problem is:-

$$\boxed{\begin{array}{l} \max \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{subject to } \alpha_i \geq 0, \ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{array}}$$

SVM solver for XOR data set,

Class 1: $x_1 = (0,0)$, $x_4 = (1,1)$ $] \to y = -1$

class 2: $x_2 = (1,0)$, $x_3 = (0,1)$ $] \to y = 1$

Kernel function :- $x = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ $x' = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$

polynomial of order 2

$$K(x, x') = (x^T x' + 1)^2$$
$$= (u_1 v_1 + u_2 v_2 + 1)^2$$
$$= 1 + 2u_2 v_2 + 2u_1 v_1 + 2u_1 v_1 u_2 v_2 + (u_1 v_1)^2 + (u_2 v_2)^2$$

$\phi(X)$ i.e. $= \phi([x_1, x_2])$

can be

$$= [1, \sqrt{2} x_1, \sqrt{2} x_2, \sqrt{2} x_1 x_2, x_1^2, x_2^2]$$

$\therefore \phi(x_1) = \phi([0,0])$

$= [1, 0, 0, 0, 0, 0]^T$

$\phi(x_2) = \phi([1,0]) = [1 \ \sqrt{2} \ 0 \ 0 \ 1 \ 0]^T$

$\phi(x_3) = \phi([0,1]) = [1 \ 0 \ \sqrt{2} \ 0 \ 0 \ 1]^T$

$\phi(x_4) = \phi([1,1]) = [1 \ \sqrt{2} \ \sqrt{2} \ \sqrt{2} \ 1, 1]^T$

objective function is,

$$\mathcal{L}_D(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2} \sum_{i,j=1}^{4} \sum \alpha_i \alpha_j y_i y_j k_{ij}$$

s.t. $\sum \alpha_i y_i = 0 \Rightarrow \alpha_1 + \alpha_4 = \alpha_2 + \alpha_3$

$\alpha_i \geq 0$

inner product is represented as

$$K = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 1 & 4 \\ 1 & 1 & 4 & 4 \\ 1 & 4 & 4 & 9 \end{bmatrix}$$

Now optimizing w.r.t. Lagrange multiplier

leads to,

$\cancel{\frac{\partial L}{\partial \alpha_1} = 0 \Rightarrow}$

$$L(\alpha) = \sum \alpha_i - \frac{1}{2}\left(\alpha_1^2 + 4\alpha_2^2 + 4\alpha_3^2 + 9\alpha_4^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 + 2\alpha_2\alpha_3 - 8\alpha_2\alpha_4 - 8\alpha_3\alpha_4\right)$$

$$\frac{\partial L}{\partial \alpha_1} = 0 \Rightarrow \quad \alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1$$

$$\frac{\partial L}{\partial \alpha_2} = 0 \Rightarrow \quad 4\alpha_2 - \alpha_1 + \alpha_3 - \cancel{8}\, 4\alpha_4 = 1$$

$$\frac{\partial L}{\partial \alpha_3} = 0 \Rightarrow \quad 4\alpha_3 - \alpha_1 + \alpha_2 - 4\alpha_4 = 1$$

$$\frac{\partial L}{\partial \alpha_4} = 0 \Rightarrow \quad 9\alpha_4 + \alpha_1 - 4\alpha_2 - 4\alpha_3 = \cancel{0}\, 1$$

Solving these equations,

gives

$\alpha_1 = 13/3$; $\alpha_2 = 8/3$, $\alpha_3 = 8/3$

$\alpha_4 = 2$.

Thus all the data points are support vectors in this case.

$$w = \sum_{i=1}^{4} \alpha_i y_i \, \phi(x_i)$$

$$= \alpha_1 (-1) \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T + \alpha_2 (1) \begin{bmatrix} 1 & \sqrt{2} & 0 & 0 & 1 & 0 \end{bmatrix}^T$$

$$+ \alpha_3 (1) \begin{bmatrix} 1 & 0 & \sqrt{2} & 0 & 0 & 1 \end{bmatrix}^T + \alpha_4 (-1) \begin{bmatrix} 1 & \sqrt{2} & \sqrt{2} & \sqrt{2} & 1 & 1 \end{bmatrix}^T$$

$$= \begin{bmatrix} -13/3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 8/3 \\ \frac{8\sqrt{2}}{3} \\ 0 \\ 0 \\ 8/3 \\ 0 \end{bmatrix} + \begin{bmatrix} 8/3 \\ 0 \\ 8/3\sqrt{2} \\ 0 \\ 0 \\ 8/3 \end{bmatrix} + \begin{bmatrix} -2 \\ -2\sqrt{2} \\ -2\sqrt{2} \\ -2\sqrt{2} \\ -2 \\ -2 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \\ \frac{2\sqrt{2}}{3} \\ \frac{2\sqrt{2}}{3} \\ -2\sqrt{2} \\ 2/3 \\ 2/3 \end{bmatrix}$$

So $g([x_1, x_2])$

$$= w^T \phi(x)$$

$$= w^T \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1x_2 & x_1^2 & x_2^2 \end{bmatrix}^T$$

$$= \begin{pmatrix} -1 & \frac{2\sqrt{2}}{3} & \frac{2\sqrt{2}}{3} & -2\sqrt{2} & 2/3 & 2/3 \end{pmatrix} \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

$$= -1 + 4/3\, x_1 + 4/3\, x_2 - 4x_1x_2 + 2/3\, x_1^2 + 2/3\, x_2^2$$

is the decision boundary,

(Q2) <u>Bayesian Classifier</u>: is a simple probabilistic classifier based on applying ~~bay~~ bayes' theorem. It is based on the idea that the role of class is to predict the values of features for the members of that class. If an agent know the class it can predict the values of other features. if it doesn't know Bayes rule can be applied to predict the class given some features.

so if (X, Y) takes values in $R^d \times \{1, 2, \ldots k\}$
Y is class label of X.

$X \mid Y = r \sim P_r$ for $r = 1, \ldots k$

$\sim$ means "is distributed as" P_r is prob. distribution

A Classifier assigns to an observation $X = x$ a estimate of what unobserved label $Y = r$ actually was.

$$C^{Bayes}(x) = \underset{r \in \{1, 2, \ldots k\}}{\operatorname{argmax}} P(Y = r \mid X = x)$$

$$P(C \mid F_1, F_2, \ldots F_n) = \frac{P(C)\, P(F_1, \ldots F_n \mid C)}{P(F_1 \ldots F_n)}$$

$$\propto P(C)\, P(F_1 \ldots F_n \mid C)$$
$$= P(C)\, P(F_1 \mid C)\, P(F_2 \mid C, F_1)\, P(F_3 \mid C, F_1, F_2) \ldots P(F_n \mid C, F_1, F_2, \ldots F_{n-1})$$

(Q3) According to Fisher's Linear Discriminant methods–

If we have,

$$\bar{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} \bar{x}_n \quad \text{and} \quad \bar{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} \bar{x}_n$$

Simplest measure of seperation of classes comes out to be when project onto ω, i's the seperation of projected class means.

so $$m_2 - m_1 = \omega^T (\bar{m}_2 - \bar{m}_1) \quad , m_k = \omega^T \bar{m}$$

and $$\omega \propto \bar{m}_2 - \bar{m}_1 \quad \text{(using Langrange multiplier)}$$

and within class variance i's given by

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

so fisher criteria,

$$J(\omega) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

$$J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega}$$

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$

$$S_W = \sum_{n \in C_1} (\bar{x}_n - \bar{m}_1)(\bar{x}_n - \bar{m}_1)^T + \sum_{n \in C_2} (\bar{x}_n - \bar{m}_2)(\bar{x}_n - \bar{m}_2)^T$$

$$\therefore \boxed{\omega \propto S_W^{-1} (\bar{m}_2 - \bar{m}_1)} \quad \text{—} (1)$$

Now we show that if we use least square method using $t = N/N_1$ for class C_1 & $t = -\frac{N}{N_2}$ for

So sum of squares error is given by

$$E = \frac{1}{2} \sum_{n=1}^{N} (w^T x_n + w_0 - t_n)^2$$

$$\frac{\partial E}{\partial w_0} = 0$$

$$\Rightarrow \sum_{n=1}^{N} (w^T x_n + w_0 - t_n) = 0$$

$$\Rightarrow N w_0 = - w^T \sum_{n=1}^{N} x_n$$

$$\left(\because \sum t_n = 0 \right)$$

$$= \frac{N_1}{N_1} \times N_1 - \frac{N_2}{N_2} N_2 = 0$$

$$\Rightarrow \boxed{w_0 = - w^T \bar{m}}$$

$$\bar{m} = \frac{1}{N} \sum x_n = \frac{1}{N} (N_1 \bar{m}_1 + N_2 \bar{m}_2)$$

Now $\frac{\partial E}{\partial w} = 0 \Rightarrow \sum_{n=1}^{N} (w^T \bar{x}_n + w_0 - t_n) \bar{x}_n = 0$

Now by substituting $t_n = \frac{N}{N_1}$ for $n \in C_1$ & $t_n = \frac{-N}{N_2}$ for $n \in C_2$

we get

$$\left(S_W + \frac{N_1 N_2}{N} S_B \right) w = N (m_1 - m_2)$$

Now $S_B w \propto m_1 - m_2$

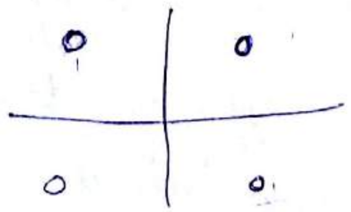
$\therefore$ $\cancel{w \propto (S_W^{-1} \bar{m}_2 - \bar{m}_1)}$

$$S_W w = K(\bar{m}_1 - \bar{m}_2) \quad K \text{ is some constant.}$$

$$w \propto S_W^{-1} (\bar{m}_1 - \bar{m}_2)$$

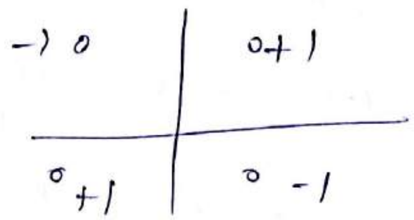
which is same as $\rightarrow$ (1).

Q5



(a) Linear Kernel

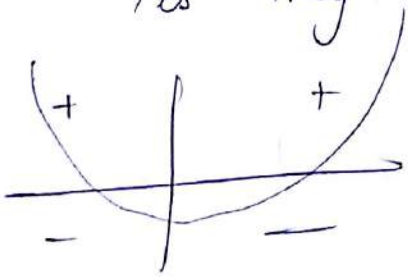
No it can't be shattered, since VC dim of the data given is 3. Clearly we can't find any linear boundary for say such classes

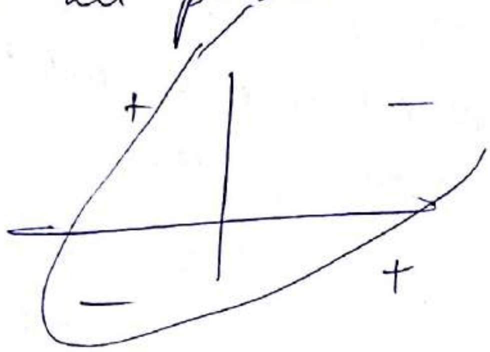


Can't be shattered.

(b) Polynomial kernel of degree 2.

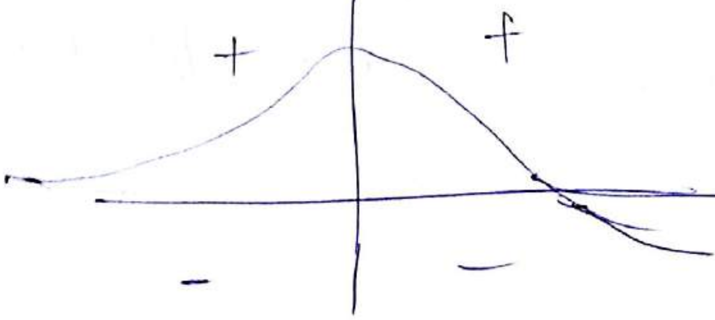
Yes they can shatter all possible labels

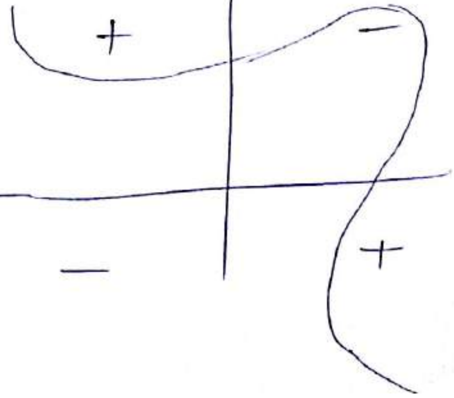




(c) Gaussian kernel

Now again a gaussian kernel can also shatter all possible labels.





Scanned by CamScanner