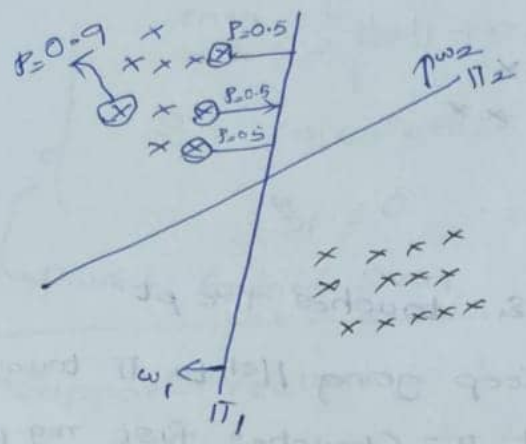
23/12/19:-

Support Vector Machines.

Geometric Intution :-



We can draw so many hyperplanes to seperate these two (+ve, -ve) pts

Actually many hyperplanes that seperate my +ve & -ve pts.

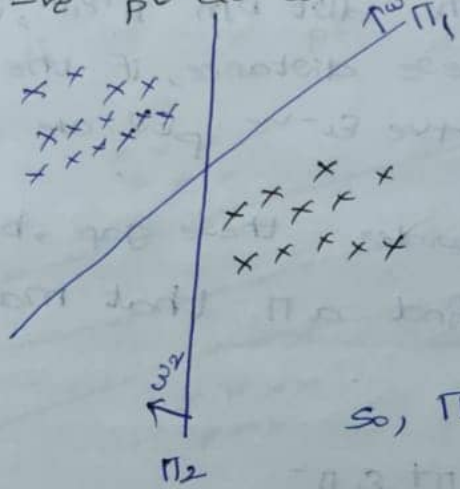
#We need to choose the best hyperplane.

Suppose if i take Π_1, so many pts are close to hyperplane

→ In Log. Reg we said, $P(y_i = 1) = \sigma(w^T x_i)$

So, pts close to hyperplane have low prob. (about 0.5) & pts that are far away have high probability.

#Key Idea:- Find a hyperplane that seperates my +ve & -ve pt as wide as possible.



Blue & Black pts are as far away as possible to Π_1 as compared to Π_2

So, Π_1 better than Π_2

→ So such a hyperplane is called Margin Maximising hyperplane.

→ Π^+ is //el to Π & touches +ve pt.

Similarly, if i keep going //el to Π towards my -ve pts i get Π^- (touches first neg. pt)

Π^- is //el to Π

Since Π^+ & Π^- are //el to each other, so dist. b/n Π^+ & Π^- are having const dist

Now, we want to find a hyperplane Π S.T if i go //el to Π & touch +ve pt with Π^+ & //el to Π, touch -ve pt with Π^-, the margin is the dist b/n Π^+ & Π^-, i need to maximise these distance, if the dist. is maximized +ve & -ve pts are quite faraway from Π,

→ The wider these gap, better for us.

→ SVM:- Try to find a Π that maximises the margin.

↓

dist b/n Π^+ & Π^-

So, if margin is high chances of misclassification will decrease & my Gener. Error will improve.

As Margin Incr my generaliz. acc improve

↓ Future/unseen

→ Dual Form of SVM:-

Soft-margin"

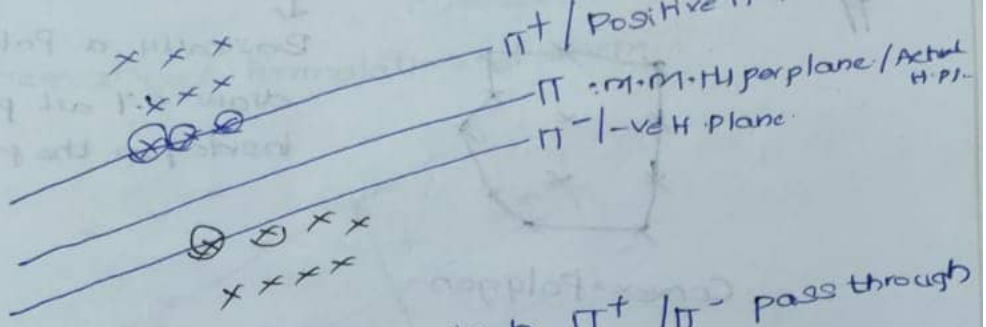
$$\min_{w,b} \frac{1}{2}\|w\| + C \cdot \sum_{i=1}^{n} \xi_i$$

S.T $y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i$

$\xi_i \geq 0$.

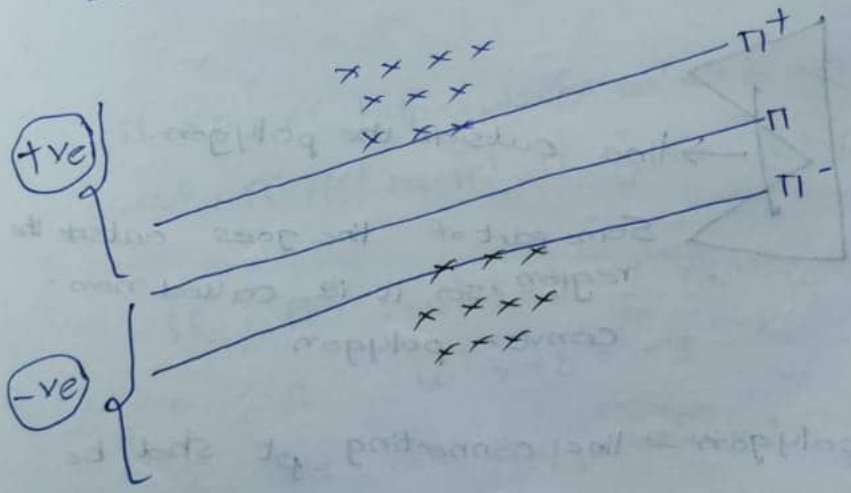
→ Primal Form of SVM

→ Support Vector:-



The pts through which Π^+ / Π^- pass through are called support Vectors.

I not only use $\Pi^+ \& \Pi^-$, Any pt above Π is +ve & Below Π is -ve pt.



Alternative Geometric Interpretation of SVM:-

First draw Convex-hull

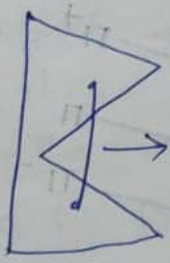
↓

Basically a Polygon, you can draw s.t all pts are either inside / on the polygon.

#

Convex-Polygon

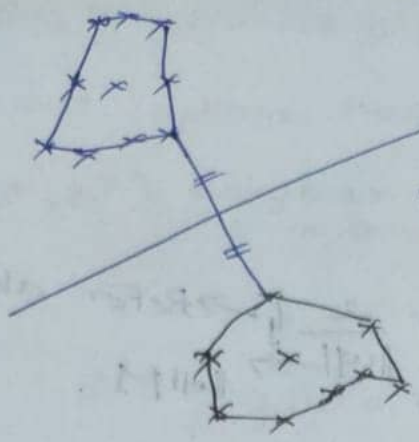
If i need to go from one pt to another pt the line is going inside the triangle only.
So △le is Convex-Polygon.



→ line outside the polygon.

Some part of line goes outside the region, so it is called non-convex polygon.

→ So, convex polygon – line connecting pt shd be inside the polygon.



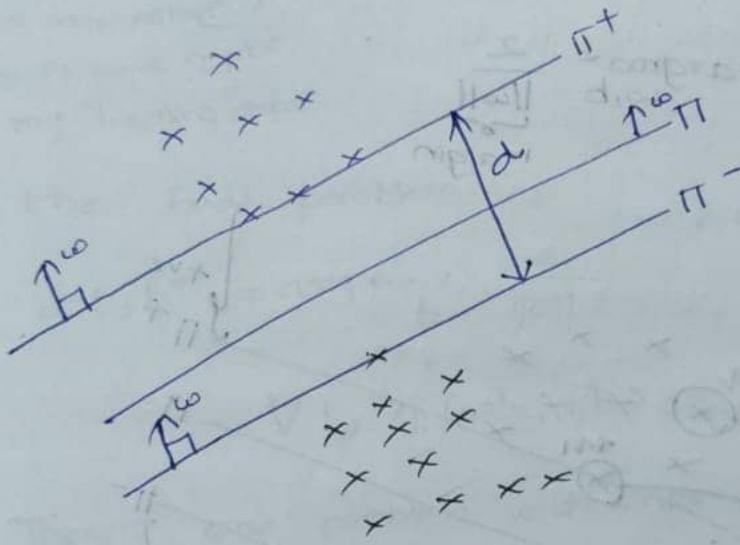
① Construct a convex hull for +ve pts & -ve pts.

② Find the shortest line connecting these hulls

③ Bisect the line

The plane that bisects it is called margin maximising hyperplane.

→ Mathematical Formulation of SVMs:-



Π → Eq: of Hyperplane: $\omega^T x + b = 0$ ($\omega \perp \Pi$)

as, $\Pi \parallel$el to Π^+, Π^-, so $\omega \parallel$el to Π^+, Π^-

If $\Pi^+ : \omega^T x + b = 1$ (as ω is $\parallel$el to Π^+)

$\Pi^- : \omega^T x + b = -1$

I did not say, $\omega^T \omega = 1$.

$\omega^T \omega \neq 1$; ω is not a unit vector.

Let, ω be some vector, not necessarily unit vector

$\Pi : w^T x + b = 0$

$\Pi^+ : w^T x + b = 1$

$\Pi^- : w^T x + b = -1$

Margin :- dist $d = \frac{2}{\|w\|}$ → Refer alexander pdf

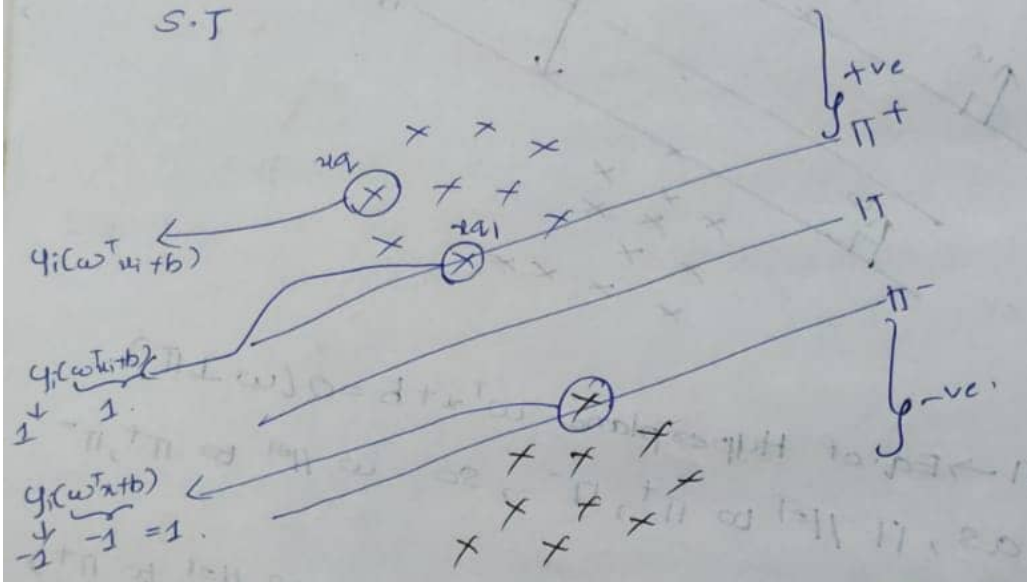
$\|w\| \rightarrow \|w\| \neq 1$.

→ We want to find w^* & b^* s.t which maximise the margin

s.t (constraint)

$$(w^*, b^*) = \underset{w,b}{\text{argmax}} \underbrace{\frac{2}{\|w\|}}_{\text{margin}}$$

s.t



→ For x_q :-

$y_i(w^T x_i + b) > 1$.

+1 dist > 1.

For x_{q_1} :-

$y_i(w^T x_i + b) = 1$

+1 dist = 1.

$w^T x + b = 1$

So, For S.V it's 1.

→ So, our optimiz. Problem is

$$(w^*, b^*) = \underset{w,b}{\arg\max} \underbrace{\frac{2}{\|w\|}}_{\text{margin}}$$

S.T $\underbrace{y_i(w^T x_i + b) \geq 1}_{\text{If all +vepts are above } \Pi^+ \text{ \& all -ve pts are above } \Pi^-}$ for all x_i

It's n-constraint because we've n-pts in my Training data.

So, the Final problem is

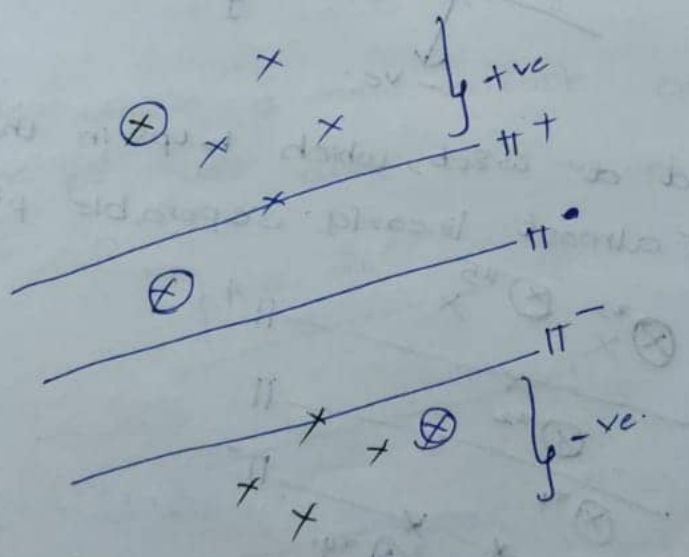
$$w^*, b^* = \underset{w,b}{\arg\max} \frac{2}{\|w\|}$$

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

Constr. optimiz. Problem of S.V.M.

→ This only works if data is linearly Seperable

There's one problem with this.



So, here there are 3 pts, which is violating our rule, +pt below Π^-

-ve pt above Π^+

& another pt(+ve) in no man's land (i.e, b/w Π^+, Π^-)

So, these 3 pts never satisfy our constraint

So, here it's almost Linearly seperable, not linearly seperable, in such a case its difficult to find w, b.

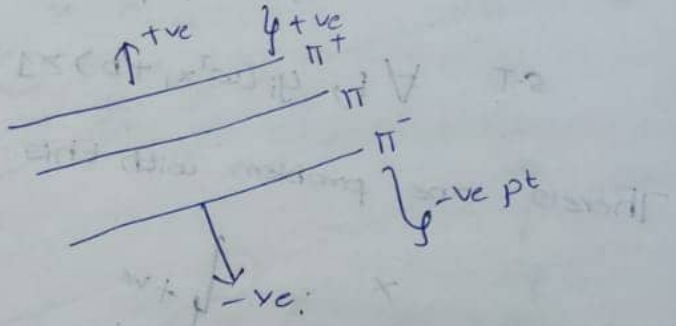
So,

$$(w^*, b^*) = \underset{w,b}{\text{argmax}} \frac{2}{\|w\|}$$

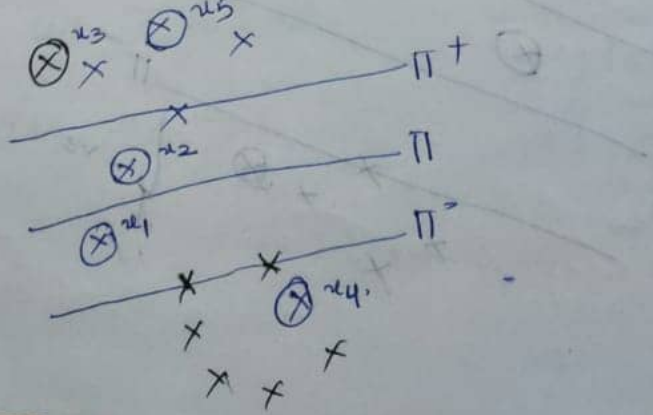
S.T $y_i[w^T x_i + b] \geq 1$ → is called

Hard-Margin S.V.M.

Very strong saying that



→ Can we find a w & b, which helps in the problems & useful for almost linearly seperable pts.



→ For x_1, → $y_i = 1$

$w^T x_i + b$

x_1 is b/n Π & Π^-

$\Pi : w^T x_i + b = 0$

$\Pi^+ : w^T x_i + b = +1$

$\Pi^- : w^T x_i + b = -1$

So, For x_1 $y_i(w^T x_i + b) = -0.5$ (b/n 1 & -1)

y_i = +1

↓ half way b/n Π & Π^-

So, $y_i(w^T x_i + b) = 1 - (1.5)$

↓ ξ_1

This 1 because, as it's +ve pt, & so, $y_i(w^T x_i + b) = 1$ for Π^+, actually x_1 need to be above Π^+, but it's not there so, we'll write it as follows

→ For, x_4 → $y_i = 1$

$y_i(w^T x_i + b) = -1.5$

$y_i(w^T x_i + b) = 1 - (2.5)$

ξ_4

→ For x_2, → $y_i = 1$

$y_i(w^T x_i + b) = 0.5$ → half way b/n Π & Π^+ so 0.5

$y_i(w^T x_i + b) = 1 - (0.5)$

ξ_2

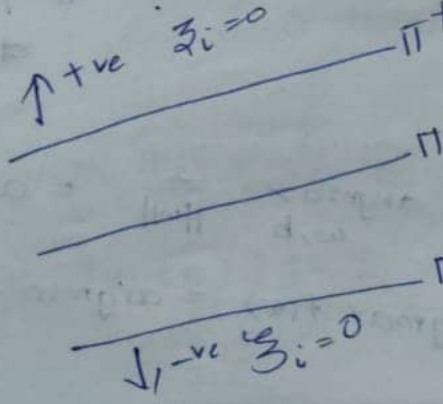
→ For x_5 → $y_i = 1$.

$y_i(w^T x_i + b) = 1.5$ → It's >1 ; so no worries

(ie, y_i $w^T x + b = 1$ for +ve pt)

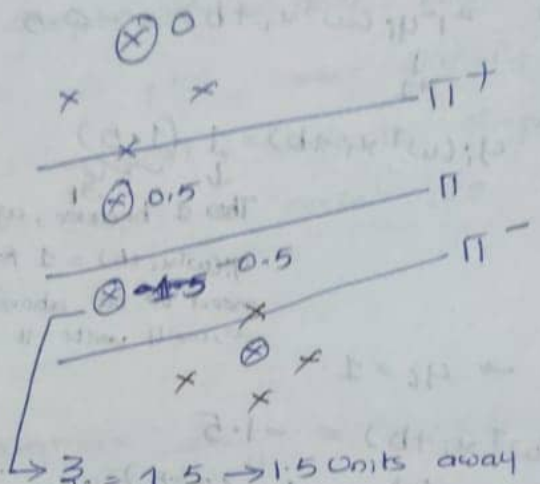
→ We create a new variable called ξ_i S.T if +ve pt lies above Π^+ & -ve pt lies below Π^-

$\xi_i = 0$.



So, if +ve / -ve pt lies anywhere except in the regions showed above it

$\xi_i \geq 0$



$\xi_i = 1.5 \rightarrow 1.5$ units away from right hyperplane

as $\xi_i \uparrow$; pt is farther away from the correct hyperplane in the incorrect reg direction.

So, For every pt $x_i \longrightarrow \xi_i$

$\xi_i = 0$, if $y_i(w^T x_i + b) \geq 1$ → Correctly classified

$\xi_i \geq 0$ & it is equal to the some units of dist. away from the correct hyperplane either Π^+ / Π^- in the incorrect direction.

So, now,

$$(w^*, b^*) = \underset{w,b}{\text{argmax}} \frac{2}{\|w\|} = \underset{w,b}{\text{argmin}} \frac{\|w\|}{2}$$

$\text{argmax } f(x) = \text{argmin } -f(x) / \frac{1}{f(x)}$

So, now

$$w^*, b^* = \underset{w,b}{\operatorname{argmin}} \underbrace{\frac{\|w\|}{2}}_{\text{margin}} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

$\sim \frac{1}{\text{margin}}$

S.T

$$y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i > 0.$$

Previously we saw for every misclassified pt, i can write $y_i(w^T x_i + b) = 1 - \xi_i$; where ξ_i is +ve

$\xi_i > 0.$

So, For all misclassified pt

$y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i$ } Incorrect classified pt

where $\xi_i > 0$

For correctly classified pt $\xi_i = 0$.

We want to minimise errors / misclassifications

It means, $\xi_i > 0$, For misclassified pts.

$\downarrow\downarrow$

$\min \sum \xi_i \rightarrow$ we want to minim. sum of ξ_i

as per Π^+, Π^-

$\rightarrow C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i \rightarrow$ dist.

avg.

avg dist of misclassified pts from correct hyperplanes

hyper parameter.

$\rightarrow (w^*, b^*) = \underset{w,b}{\operatorname{argmin}} \underbrace{\frac{\|w\|}{2}}_{\sim \frac{1}{\text{margin}}} + C \cdot \underbrace{\frac{1}{n} \sum_{i=1}^{n} \xi_i}_{\text{avg. dist for misclassified pts.}}$

S.T $y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i$ } $\rightarrow$ Constraint.

$\xi_i \geq 0$

We can think it as,

$$(w^*, b^*) = \text{argmin} \; \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

regularization (like L_2) — loss

In Log. regr we had

$$\min_{w} \; (\text{log-loss}) + \lambda (\text{reg})$$

Here,

$$\min_{w,b} \; C \cdot \text{hingeloss} + \text{Regularization}$$

C ↓ hyper-parameter.

as $C \uparrow$, loss ↑; So - tendency to make mistakes ↓

$\Rightarrow$ Overfit

$C \downarrow$, tendency to make mistakes ↑

$\Rightarrow$ Underfit

$\therefore$ $C \uparrow$; C-dominates; Overfit (high-Var)

$C \downarrow$; reg-dominates; Underfit (high-bias)

$\rightarrow \therefore (w^*, b^*) = \underset{w,b}{\text{argmin}} \; \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i$

S.T $y_i (w^T x_i + b) \geq 1 - \xi_i \; \forall i$

$\xi_i > 0$

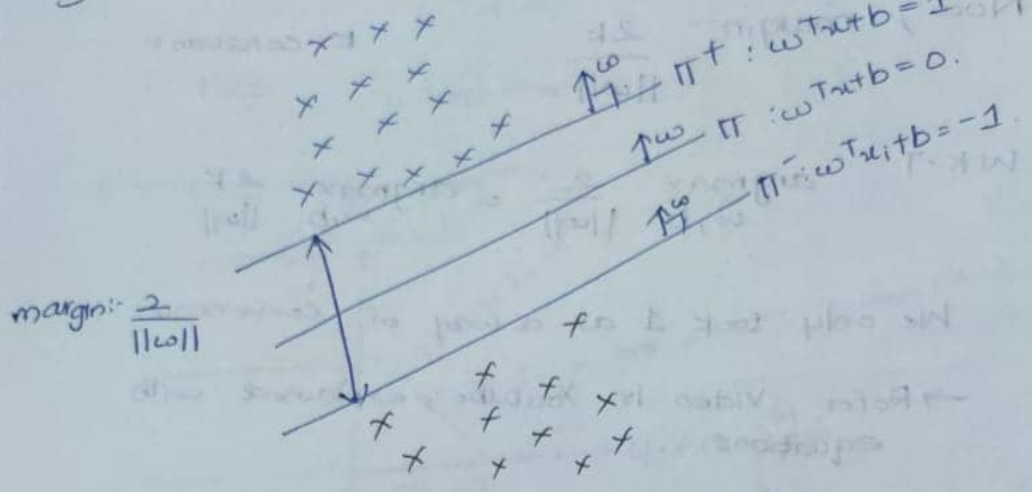
It's Soft margin-SVM

→ Hard-Margin doesn't allow errors.

But Soft-Margin allows errors but it says minimize errors.

→ Why take +1 & -1 :-

SVM:-



Q) Why +1 & -1 on R.H.S of $\pi^+ \& \pi^-$

Here, we're not saying $\|w\| \neq 1$; w could be any vector it need not be a unit vector.

$w \perp \pi, \pi^+, \pi^-$

→ margin :- $\frac{2}{\|w\|}$

Our whole task is to maximise the margin

$$w^*, b^* = \text{argmax} \frac{2}{\|w\|}$$

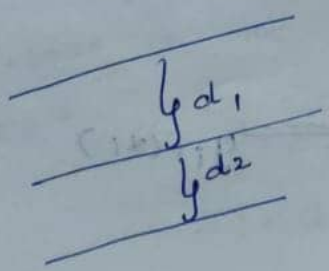
s.t $y_i(w^T x_i + b) \geq 1$.

① Let's assume,

$\pi^+ : w^T x + b = K$

$\pi^- : w^T x + b = -K$ $\quad K > 0$

we want dist shd be same



$d_1 = d_2$, that's why we took $K \& -K$, but not $K_1, -K_2$.

So, +ve, -ve hyperplane are equivalently far away from Π

Now, margin :- $\frac{2k}{\|w\|}$ — k: constant

W.K.T $\underset{w,b}{\text{argmax}} \frac{2}{\|w\|} = \underset{w,b}{\text{argmax}} \frac{2k}{\|w\|}$

We only took 1 as a way of convenience.

→ Refer Video in Youtube, explained with equations.

② Π: $w^T x + b = k$

$\left(\frac{w}{k}\right)^T x + \left(\frac{b}{k}\right) = 1$. $w \perp \Pi$

$(w')^T x + b' = 1$. $\|w\|$ need not be 1.

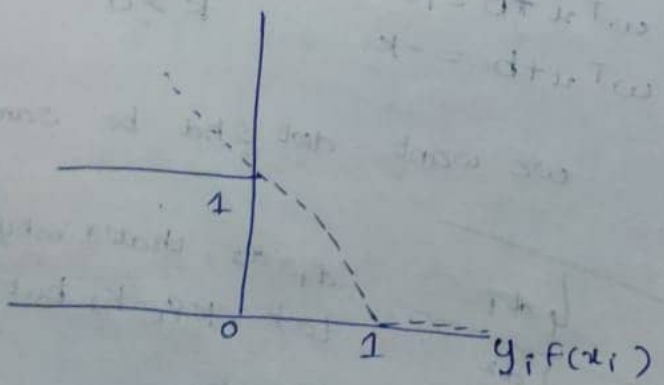
The reason to take +1 & -1 is to simplify math.

It's possible to take any value

All is because $\|w\| \neq 1$.

If $\|w\| = 1$; margin = 2. It's not possible

→ Loss-minimization : Hinge-loss :-



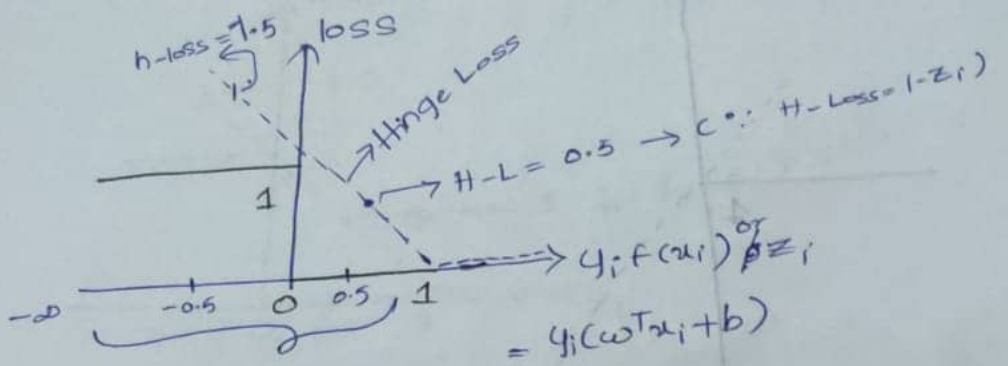
log-Reg:- min log-loss + reg.

↓

0-1-loss

lr. Reg:- lr. loss + reg.

SVM:- hinge-loss + reg



when z_i is +ve; x_i is correctly classified

$z_i > 0$; x_i correctly classified

↳ $y_i(w^T x_i + b)$

when z_i is -ve

$z_i < 0$; x_i is incorrectly classified.

Hence in 0-1, Loss we'll give 1 (Incorrect)

0 (Correct)

∵ 0-1 loss is not continuos at 0, so 0-1 loss is not differentiable.

Hinge-Loss is a S.L from $-\infty$ to 1, & from 1 it is 0.

Hinge-Loss is not differentiable at 1.

→ Hinge-Loss:- if $z_i \geq 1$; hinge-loss = 0

$z_i < 1$; hinge-loss = $1 - z_i$

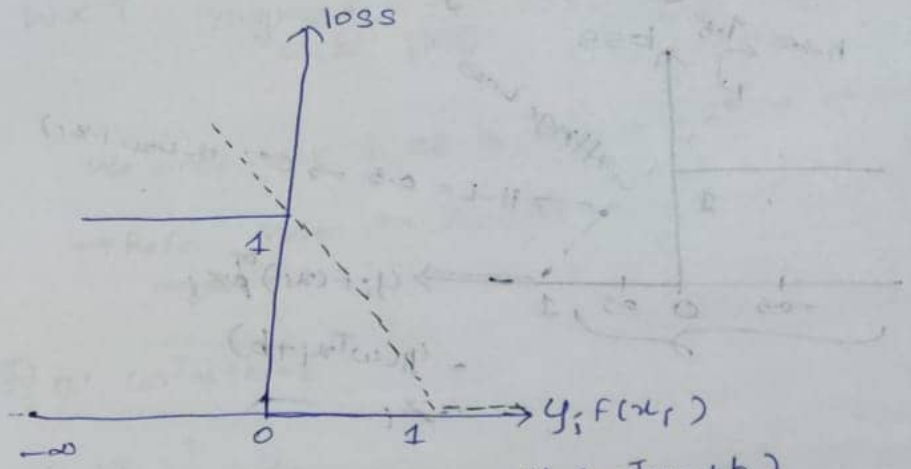
24/12 :-

→ Loss-Minimization : Hinge-loss

log-reg in loss point of view :- log-loss + reg

Lr.reg :- lr-loss + reg

SVM :- Hinge-loss + reg



$y_i(w^T x_i + b)$

$= z_i$

when z_i is +ve; $z_i > 0$; x_i is coredtly classified

$z_i < 0$; x_i is incorrectly classified

In 0-1 loss ; loss 1 if 1 incorrectly classified.

0-1 loss is not continuos, so not differentiable.

1

0

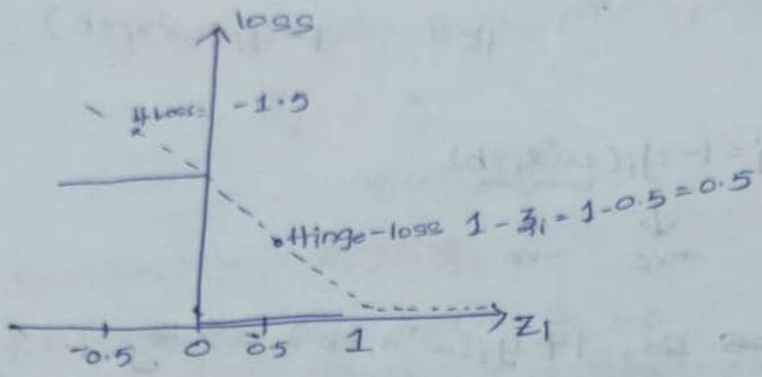
→ Hinge-loss is s.t from $-\infty$ to 1 & from 1 it's 0.

H-L is continuos, but its not diff.

→ Let's look at how hinge loss behaves from z_i point of view

Hinge-Loss :-

if $z_i > 1$; hinge-loss = 0
if $z_i < 1$; hinge-loss = $1 - z_i$ } ①



The alternative way of hinge-loss is $\max(0, 1 - z_i)$

Case 1 :- $z_i \geq 1$

$1 - z_i$ = -ve

So, max(0, -ve value) = 0.

case 2 :- $z_i \leq 1$

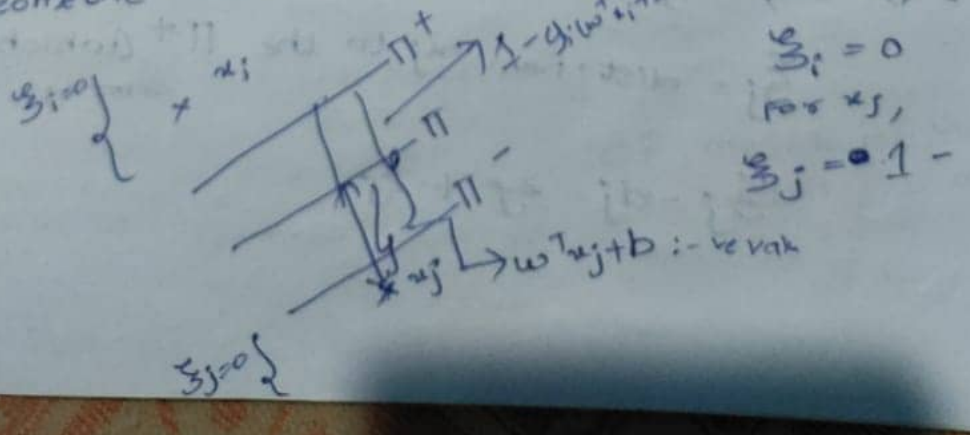
$1 - z_i$ = +ve (> 0)

So, max (0, +ve value) = +ve value

$\Rightarrow 1 - z_i$

These two cases are exactly same as ①

→ Geometric-Formulation & loss-min :-



For x_i,
$\xi_i = 0$
For x_j,
$\xi_j = 1 -$