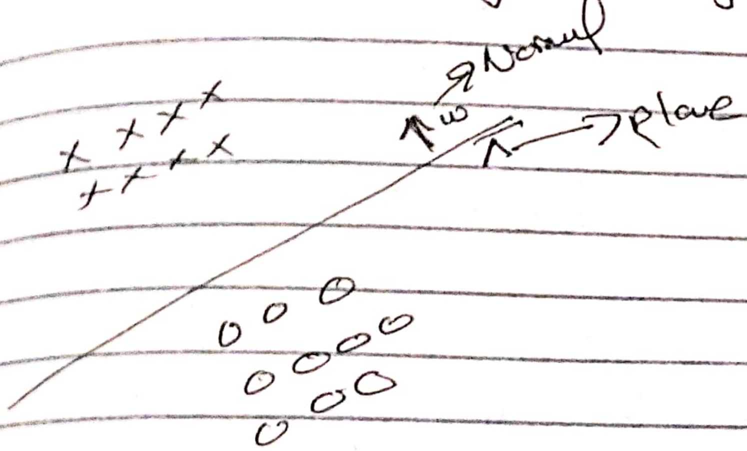
regression



Apply when data is linearly seprable

- only for binary classification

$$\boxed{w^* = \arg\max_{w} \sum_{i=1}^{n} y_i w^T x_i}$$

optimization problem

objective is to find optimal value for "w^*"

<u>Squashing & Sigmoid function.</u>

$\pi : w_i$: Best / optimal $w \rightarrow w^*$

$$\arg\max_{w} \sum_{i=1}^{n} \underbrace{y_i w^T x_i}_{\text{signed distance}}$$

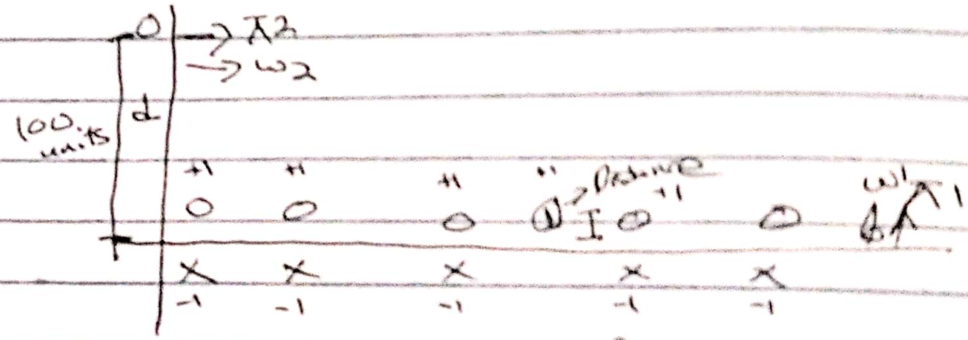
$w^T x_i$: distance from "x_i" to "π" (w is unit vector) if

$y_i w^T x_i$: +ve ⇒ π as defined by w correctly classifies "x_i"

→ -ve ⇒ incorrectly classifies.

Example:

assume distance of all the points from plane is "1"



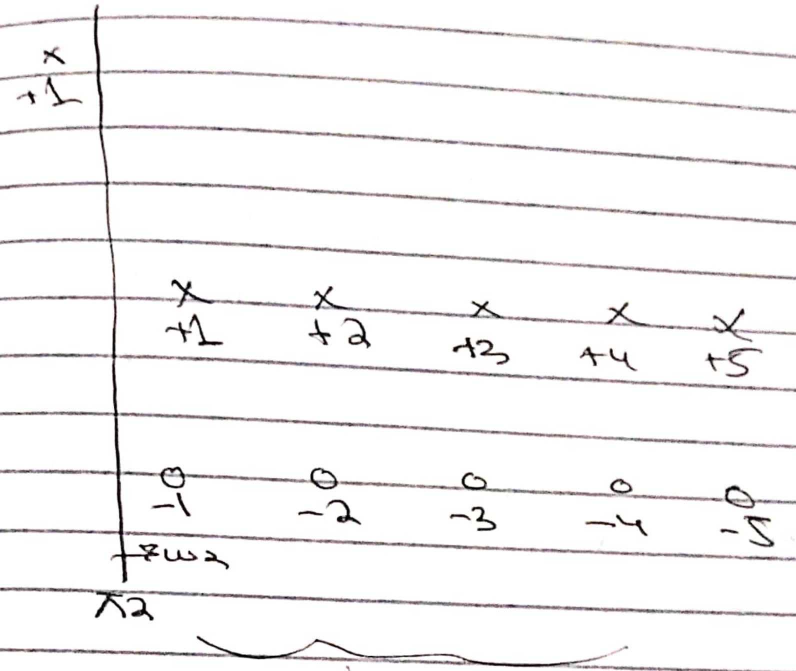
Case 1:- π_1 is class separating hyperplane

$$\sum_{i=1}^{n} y_i w_1^T x_i = 1+1+1+1+1 \; (+ve)$$
$$1+1+1+1+1 \; (-ve)$$
$$-100$$
$$= -96$$

Case 2: π_2 is class separating plane

Obj: is to find best 'π' which maximize the sum of signed distance

One single outlier is changing whole model which is very terrible situation.



distance from all points is 1 unit

Obj: find w that maximize signed distance

1+2+3+4+5, +ve
-1,-2-3-4-5 -ve
+1 outlier

= +1

we get value of +1 using w_2

+1 > -90

π_2 is my classifier

max. sum of signed dist

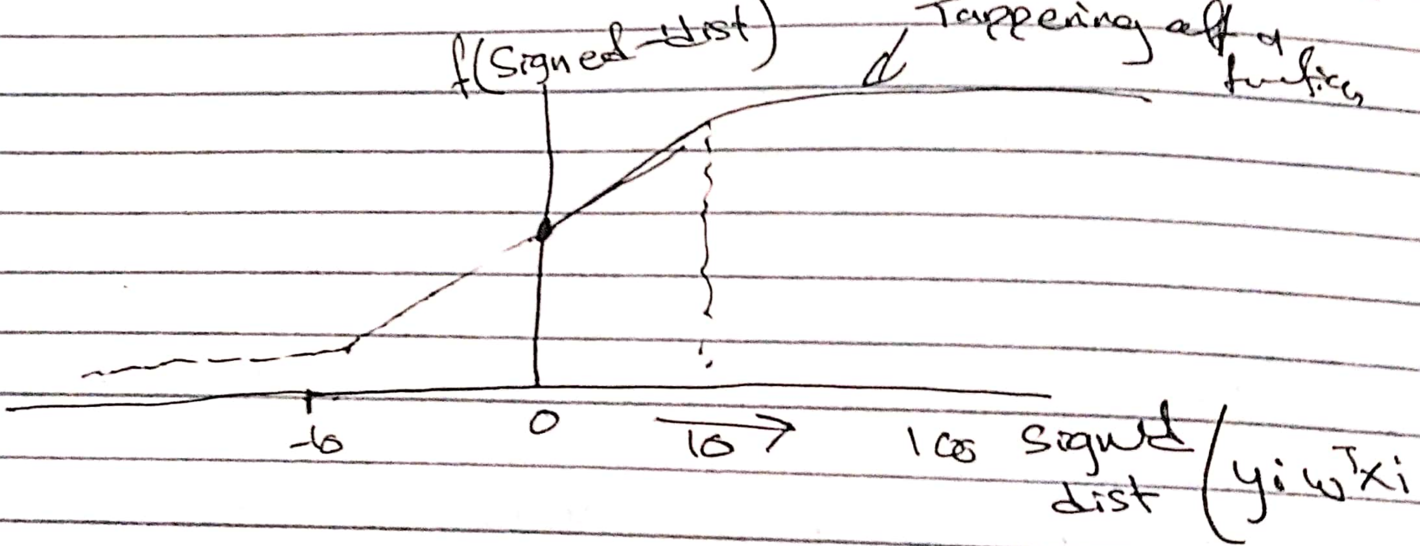
↓

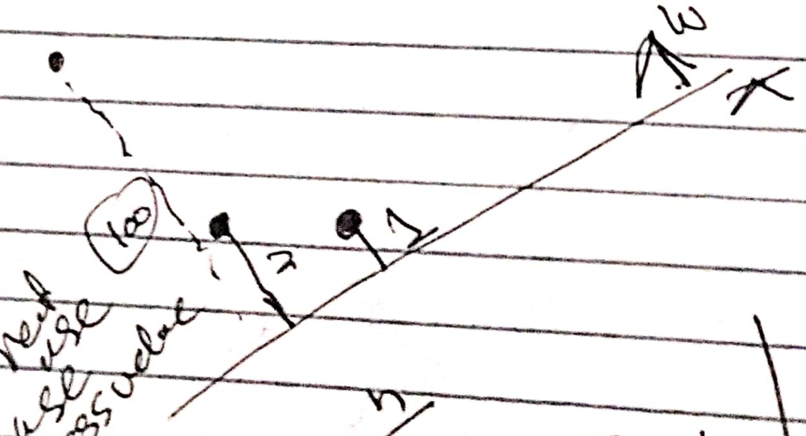
Prone to outlier

Squashing:-

iDea:- instead of using simple signed distance we here to modify 'little bit'

if signed distance is small:- use it as it is
if " " " large:- Make it a smaller value





$$\underset{w}{\arg\max} \sum_{i=1}^{n} y_i w^T x_i \quad \Bigg| \quad \underset{w}{\arg\max} \sum_{i=1}^{n} f(y_i w^T x_i)$$

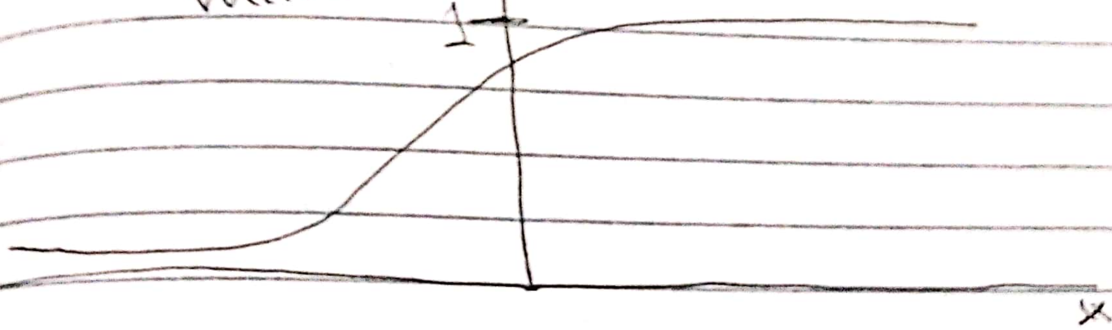
one of such function is called sigmoid function

Sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

max value = 1
min value = 0

x = Signed distance

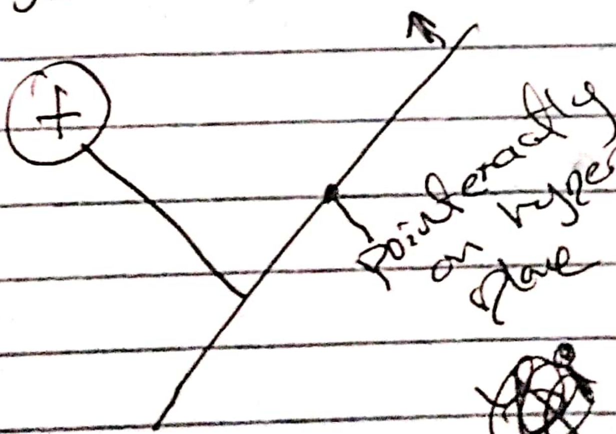


$$\sigma(0) = 0.5$$

The reason we choose sigmoid function because it has nice probabilistic interpretation.

$w^T x_i$ = very large

$\sigma(w^T x_i) = 1$



$w^T x_i = 0$

$P(y_i = 1) = 0.5$

Deriving Logistic regression using Geometry.

max. Sum. of Signed dist → outlier problem

↓

Easy to differentiate. ⇐ $\sigma(x)$ → Sigmoid

→ tapering, linear

→ Nice probabilistic interpretation.

max. Sum of transformed signed dist.

~~$w^* =$ argmax~~

$$\boxed{w^* = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \sigma(y_i w^T x_i) \quad \downarrow \quad w^* = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \frac{1}{1+\exp(y_i w^T x_i)}}$$

This function is called squashing because it squeses dis value which is $(-\infty, +\infty)$ to $(0,1)$

Scanned with CamScanner

Optimization Problem:

$$w^* = \arg\max_{w} \sum_{i=1}^{n} \frac{1}{1+\exp(-y_i w^T x_i)}$$

optimization problem.

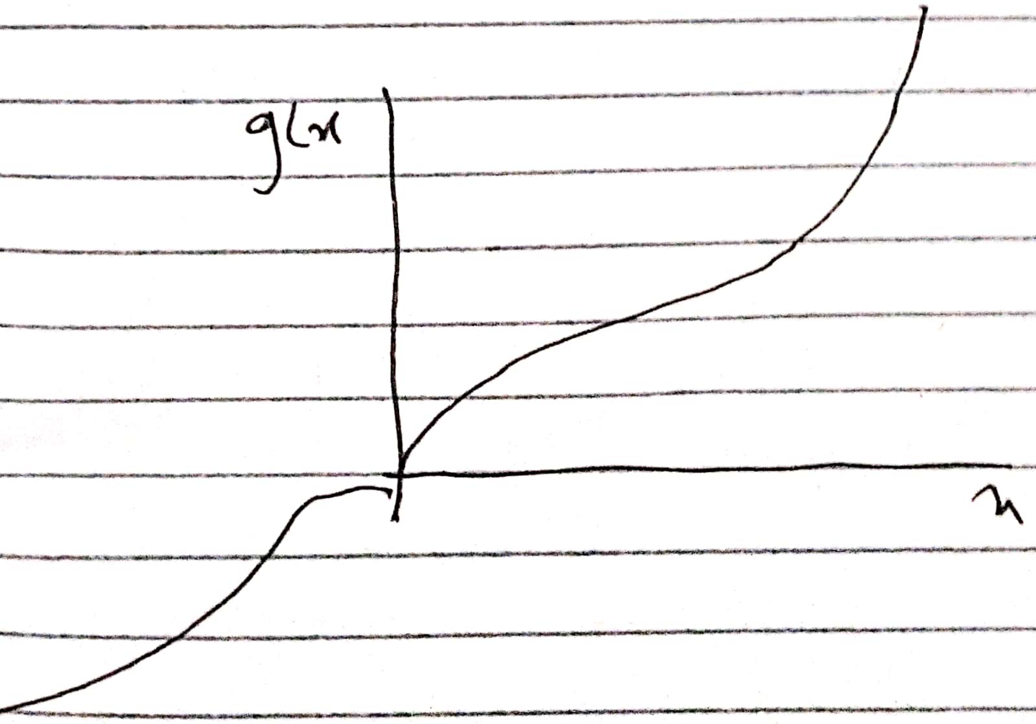
→ Monotonic function:

we modify this problem to solve easily

a function $f(x)$ is said to be monotonic increasing

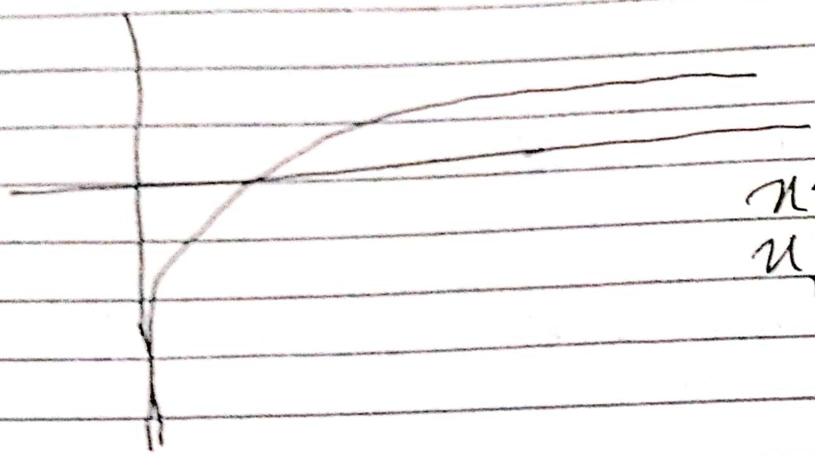
if x increase $f(x)$ increase

if $x_1 > x_2$ then $g(x_1) > g(x_2)$
then $g(x)$ is called monotonically increasing function.



$\log(x)$ $x>0$

$\log(x)$; $x>0$



$x\uparrow$ $\log(x)\uparrow$
$x\downarrow$ $\log(x)\downarrow$

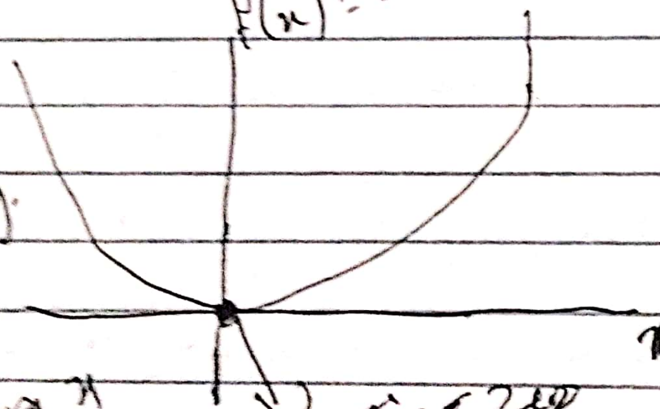
opt-prob:-

$x^* = \arg\min_x x^2$

"we want to find x which minimize x^2"

① → dx.

$f(x)=x^2$



x

x^2 is monotonically increasing when $x\geq 0$ and monotonically decreasing when $x<0$

monotonically increasing

$\Uparrow$

$g(x)=\log x$

minimum value is zero

$x^* = \arg\min_x f(x)$; $f(x)=x^2$ Problem ①

$x' = \arg\max g(f(x))$; $f(x)=x^2$ Problem ②

$g(x) = \log(x)$

$f(x) = x^2$

$x^* = \arg\min_x f(x)$

$x' = \arg\min_x g(f(x)) = \log(x^2)$

x^0 and x' is same because $g(x)$ is monotonically increasing $f(x)$.

Theorm:

if $g(x)$ is monotonic function.

$\arg\min_x f(x) = \arg\min_x g(f(x))$

$\arg\max_x f(x) = \arg\max_x g(f(x))$

Scanned with CamScanner

Original Problem.

$$w^* = \arg\max_{w} \sum_{i=1}^{n} \frac{1}{1+e(-y_i w^T x_i)}$$

$g(x) = \log(x)$

$$w^* = \arg\max_{w} \sum_{i=1}^{n} \log(\sigma(y_i w^T x_i))$$

$$w^* = \arg\max_{w} \sum_{i=1}^{n} \log\left[\frac{1}{1+\exp(-y_i w^T x_i)}\right]$$

$\log(1/x) = -\log(x)$

$$w^* = \arg\max_{w} \sum_{i=1}^{n} -\log(1+\exp(-y_i w^T x_i))$$

$\arg\max_{x} f(x) = \arg\min_{x} -f(x)$

$\arg\max_{x} -f(x) = \arg\min_{x} f(x)$

[-1 or +1)

$$\boxed{w^* = \arg\min_{w} \sum_{i=1}^{n} \log(1+\exp(-y_i w^T x_i))}$$

Solution of optimization problem

$$= \arg\min_{w} \sum_{i=1}^{n} \cancel{\log}(\cancel{1+\exp}(-y_i w^T x_i))$$

$$w^* = \arg\min_{w} \sum_{i=1}^{n} -y_i w^T x_i$$

$$\boxed{w^* = \arg\max_{w} \sum_{i=1}^{n} y_i w^T x_i}$$

↓ huge outlier problem

$$w^* = \arg\min_{w} \sum_{i=1}^{n} \log(1+\exp(-\underset{+1,-1}{y_i} w^T x_i))$$

$$w^* = \arg\min_{w} \sum -\overset{+1,0}{y_i} \log p_i - (1-y_i)\log(1-p_i)$$

$$p_i = \sigma(w^T x_i)$$

↓ Probablistic method.

Weight vectors:-

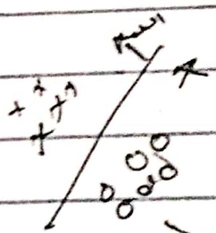
$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \log(1+\exp(-y_i w^T x_i))$$

It is called weight vector.

$$w = \langle w_1, w_2, w_3 \dots w_d \rangle$$

$\in \mathbb{R}^d \rightarrow$ d-features

$$w = \langle w_1, w_2, w_3 \dots w_d \rangle$$
$$\;\; f_1, f_2, f_3 \dots f_d$$



for every feature we have 'weight' associated with it

decision: $x_q \rightarrow y_q$ if "$w^T x_q$" > 0 than $y_q = +1$
if "$w^T x_q$" < 0 than $y_q = -1$

Probabilistic:

$$\sigma(w^T x_q) = p(y_q = +1)$$

interpretation of w:

(1) if w_i is +ve, $x_{qi} \uparrow \Rightarrow (w_i x_{qi}) \uparrow$
f_i

$$\Rightarrow \sum_{i=1}^{d} w_i x_{qi} \uparrow$$

$$\sigma(w^T x_q) \uparrow \text{ then}$$

interpretation of w.

eg. Case 1:

if w_i is +ve : if x_{qi} is $\uparrow$

$\Downarrow$

f_i

$= (w_i x_{qi}) \uparrow$

$= \left(\sum_{i=1}^{d} w_i x_{qi}\right) \uparrow$

$= \sigma(w^T x_q) \uparrow$

$= P(y_q = 1) \uparrow$

Case 2:

if w_i is -ve : if $x_{qi} \uparrow$

$\Rightarrow w_i x_{qi} \downarrow$

$\Rightarrow \sum_{i=1}^{d} w_i x_{qi} \downarrow$

$\Rightarrow \sigma(w^T x_q) \downarrow$

$\Rightarrow P(y_q = 1) \downarrow$

$\Rightarrow P(y_q = -1) \uparrow$