Lecture - Logistic Regression

https://colab.research.google.com/drive/1c21GqeA5S0do0JZ2H6_Uc0DRi8WF1SgD?usp=sharing

→ First classification model = LogReg ~~Regression~~

→ LinReg $\xrightarrow{t}$ LogReg

Notation

no. of samples - m
no. of features - d
no. of classes - n

First axis

X - (m, d) ↓

ith sample - $X[i]$, $X^{(i)}$

jth feature - $X[:, j]$, X_j

Output → y -

Nominal

One hot Encoding

Binary classification (0, 1)

$y \rightarrow (m, 1)$

Multiclass classif. n classes

$y \rightarrow (m, n)$

2 classes

1
0
1
0
1
0

mx1

3 classes

0	0	1
0	0	1
1	0	0
0	1	0
1	0	1

mx3

Class1 001
Class2 010
Class3 100

Binary

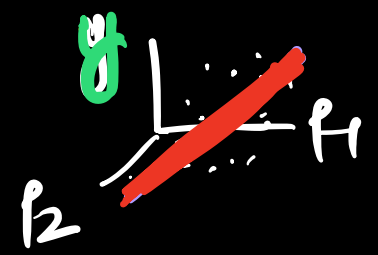
Multiclass

Parameters - $[w_0, w_1, w_2\ w_3 - - w_d]$

Summary: LinReg $h: w^T x + w_0$ ★★

Linear model - Best line of fit passing through data

Assume d-features.



line $\in$ (d+1) dimensional space.

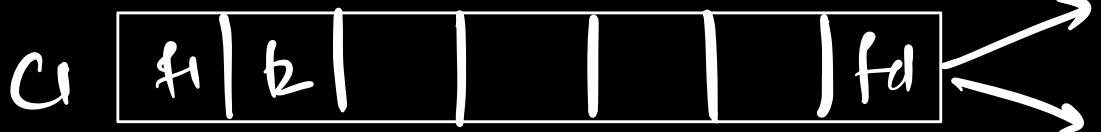
$$\text{loss} = \text{MSE} = \frac{1}{m}\sum_{i}^{m}(y^i - \hat{y}^i)^2$$

$$y \in (-\infty, \infty), \; y \in R$$

Intro to Businesscase - Verizon/Airtel/VI

Customer Churn - leaving subscription.

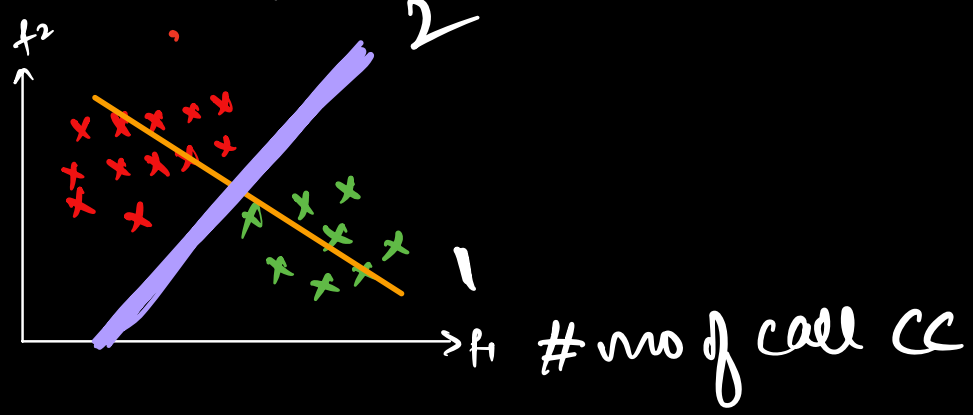
Plan Rate, Network, statelcity, # disconnection
Customer Servie Feedback, # customer care.



- Binary Classification - 1, 0 — not leave
- ↳ leave/churn
- Historical data with labels (1/0) - Supervised
- $y^i \notin R,\ y^i \in \{0, 1\}$

~~Class1 0 1~~ 0
~~Class2 1 0~~ 1

Charges for extra service



y is represent with colors not with axis.

1 line of best fit
d+1 space
one axis for y
$y = w^T x + w_0$

2 Decision Boundary
Best line of seperation
d space

Linear Regression-2

Agenda

- → LinReg → LogReg - Sigmoid
- → Geometric Interpretation
  - → likelihood vs Prob - MLE
  - → Error Fn. LogReg - ~~MSE~~ NLL/ Logloss
  - → Gradient Descent for LogReg ✓ - 2.e.x

Disclaimer: Content "may" look difficult
DONT NEED REMEMBER ★★★
THUMB RULES/RESULTS

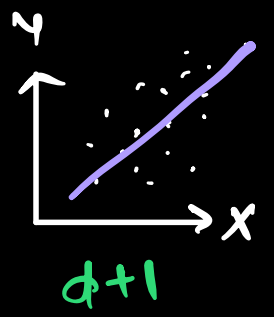
LinReg ⟶ LogReg

LinReg

$$y = W^T X + W_0$$

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots w_d x_d$$

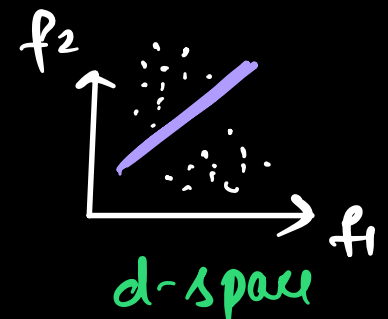
d+1 variable



LogReg

$$0 = W^T X + W_0$$

$$0 = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots w_d x_d$$



Equation of a plane

d variables.

⇒ Seperating Hyperplane is present d-space

⇒ $W^T X + W_0 = 0$

$y = w^T x + w_0 \; ; \; y \in R \quad (-\infty, \infty)$

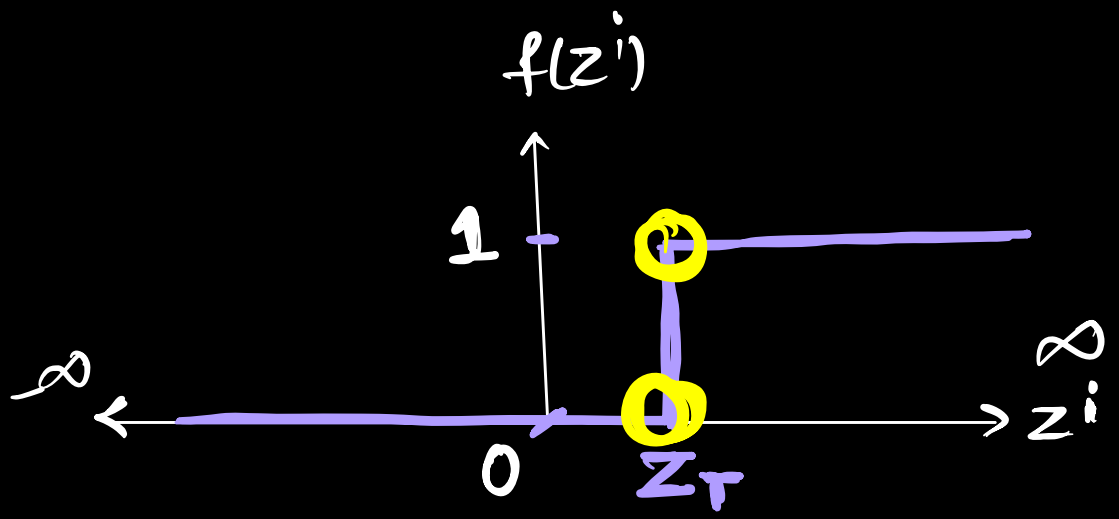
Goal: $y = f(w^T x + w_0) \in \{0, 1\} \; \{-1, 1\}$

Step1: find $z^i = w^T x^i + w_0$

Step2: $f(z^i) \rightarrow (0, 1)$ or $(-1, 1)$

↑ find this function.

Solution) Step Function + Thresholding



$z^i = w^T x^i + w_0$ - Trivial

If $z^i > z_T \rightarrow 1$

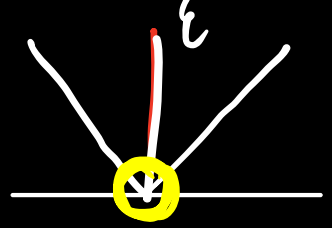
else if $z^i < z_T \rightarrow 0/-1$

else $z^i = z_T \rightarrow$ Unclassified.

$$y = SF(w^T x^i + w_0)$$

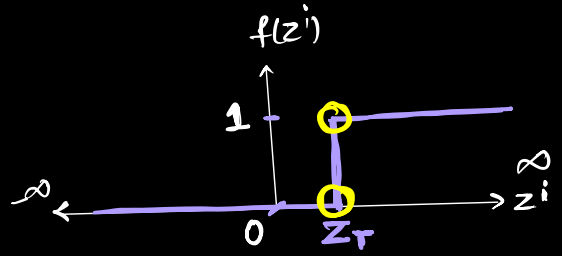
Problems with SF ?

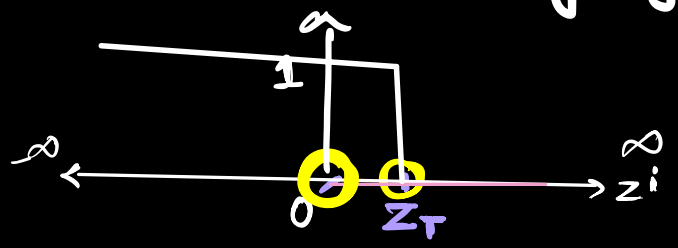
SF has sharp edges



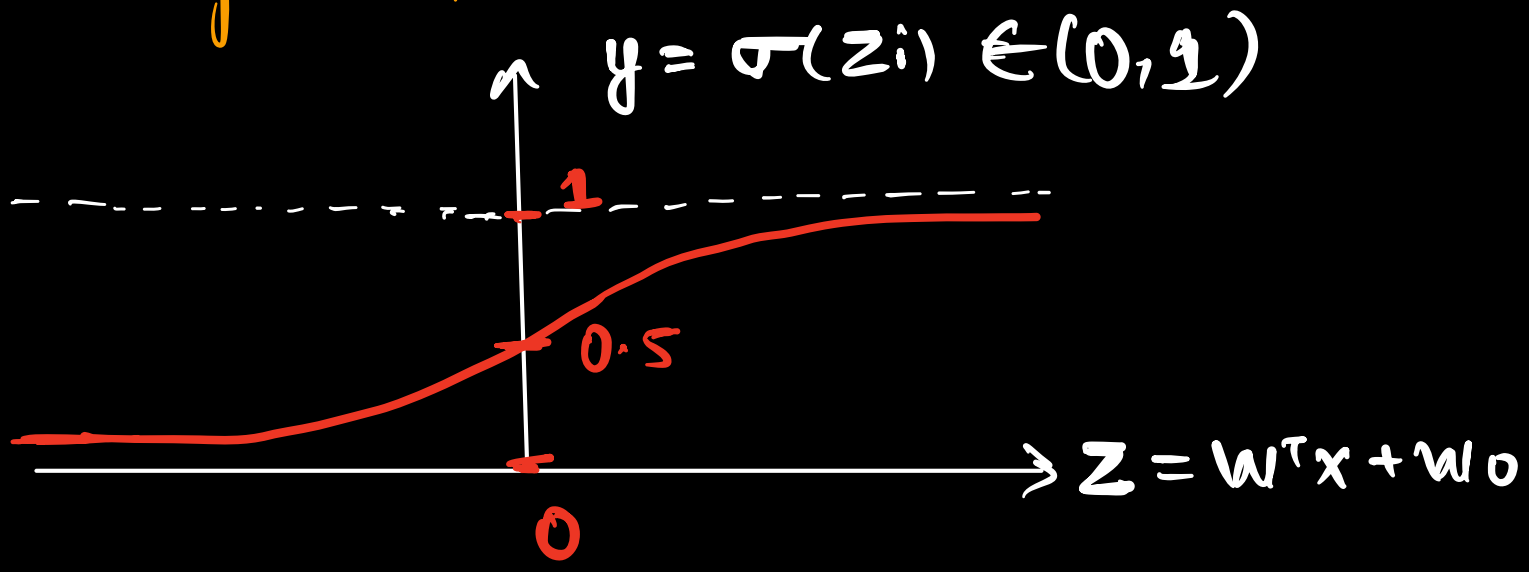
- Not differentiability

$E = |y - y^i|$





Sigmoid Function.



$y \in (0,1)$

① Squash $z \in (-\infty, \infty) \longrightarrow (0,1)$

② Differentiability ✓

③ Probabilistic Angle to classification

$y \in (0,1)$ ~~$y \in \{0,1\}$~~

$y = 0$ or 1 $\quad P(y=1|x^i) = \sigma(z^i)$

Case1) $z^i \to \infty$; $\sigma(z) \to 1$

Case2) $z^i \to -\infty$; $\sigma(z) \to 0$

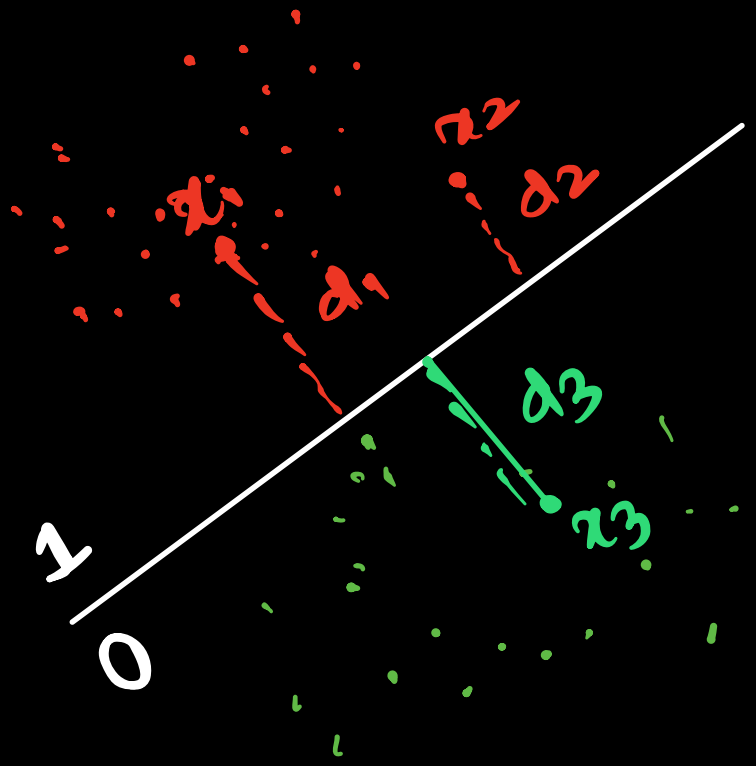
Case3) $z^i \to 0$, $\sigma(z) = 0.5$

Case 4) $z^i = +ve$, $\sigma(z) > 0.5$

Case 4) $z^i = -ve$, $\sigma(z) < 0.5$.

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

Geometric Interpretation.



$\pi : W^T x + W_0 = 0$ Decision Boundary

$$d_1 = \frac{W^T x^1 + W_0}{\|W\|}$$

$$d_2 = \frac{W^T x^2 + W_0}{\|W\|}$$

$$d_3 = \frac{W^T x^3 + W_0}{\|W\|}$$

Just a constant

Numerator

$x_1 \in$ Class 1
$x_2 \in$ Class 1
$x_3 \in$ Class 0

$$d \propto W^T x + W_0 = z$$

→ $|z|$ tells us how far is x from π
→ +/- of z tells us which class does it belong

MLE. (Maximum Likelihood Estimation)

p^i $\quad P(y^i=1 \mid x^i) = \sigma(\underbrace{w^T x^i + w_0}_{z^i}) \in [0,1)$ — ①

$1-p^i$ $\quad P(y^i=0 \mid x^i) = 1 - P(y^i=1 \mid x^i)$ — ②

$= 1-p$

↑ Observed Data.

Combine ① and ②:

Ground Truth

$$(p^i)^{y^i}(1-p^i)^{(1-y^i)}$$

Common eq.

If $y^i=1$, $\frac{\text{—}}{\text{—}} \cdot \frac{1}{\text{—}} = p^i$ — ①

If $y^i=0$, $\frac{1}{\text{—}} \cdot \frac{\text{—}}{\text{—}} = 1-p^i$ — ②

Independent Event A, B $P(A,B) = P(A) \cdot P(B)$

$P(A,B,C) = P(A) \cdot P(B) \cdot P(C)$

$P(A,B,C,\ldots Z) = P(A) \cdot P(B) \cdots P(Z)$

m samples, independent

$P(Y^1=1, Y^2=1, Y^3, \ldots Y^m=1 \mid X) = P^1 \cdot P^2 \cdot P^3 \cdot P^4 \cdot \ldots P^m$

Σ Summation
Π Multiplication

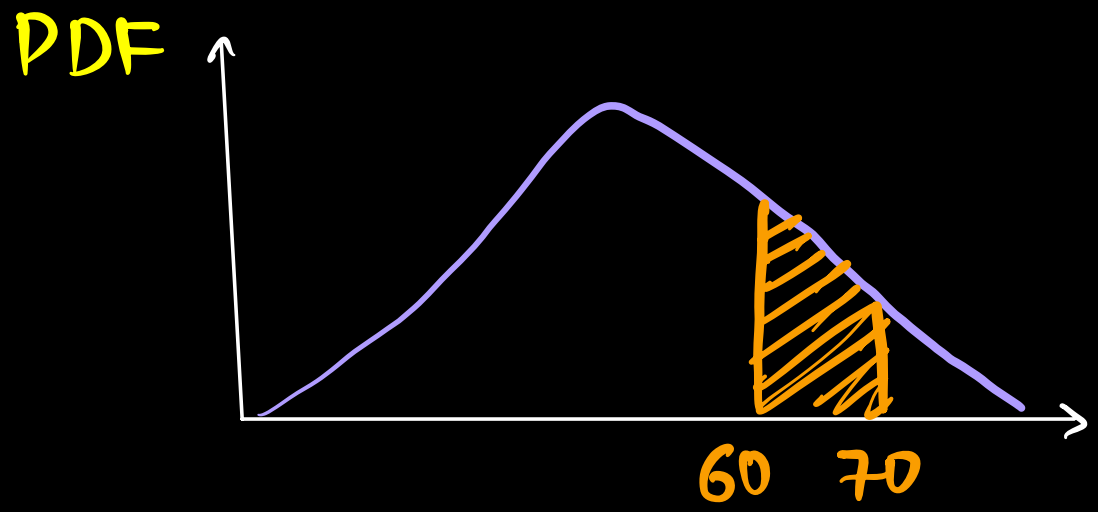
$$= \prod_{i=1}^{m} \left[(p^i)^{y^i} \times (1-p^i)^{(1-y^i)} \right]$$

= PDF

= Joint Distribution of data

= Likelihood.

Prob Vs Likelihood



AUC b/w 60 and 70 of a PDF

$$P(60 < x \leq 70 \mid \mu, \sigma)$$

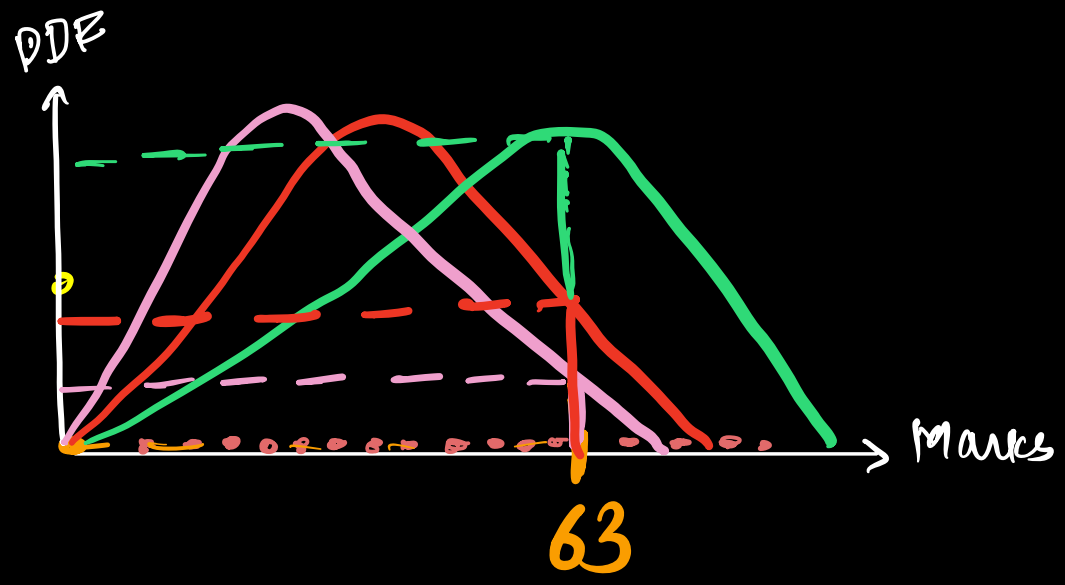
↑ Distribution is given

BUT in our case,

- We have Data
- We don't know the distribution.

$= P(\mu, \sigma \mid \text{Data})$

$= P(\mu, \sigma \mid m_1 = 63)$



You only one data pt. $m_1 = 63$, which distrib. is it most likely coming from. - Green

You have m data pt → Which dist. is most likely?

P(Distribution | Data)

↓

Likelihood. ★★★

Maximum Likelihood Estimation

$$\text{likelihood} = \prod_{i=1}^{m} (p^i)^{y^i} \cdot (1-p^i)^{1-y^i}$$

↑ maximise

$$\log(\text{likelihood}) = \log\left(\prod_{i=1}^{m} (p^i)^{y^i} \cdot (1-p^i)^{1-y^i} \right)$$

$$\log(A \cdot B \cdot C) = \log(A) + \log(B) + \log(C)$$

$$= \sum_{i=1}^{m} \log\left((p^i)^{y^i} \cdot (1-p^i)^{1-y^i}\right)$$

$$= \sum_{i=1}^{m} \log p^{i\,y^i} + \log(1-p^i)^{(1-y^i)}$$

$$\log A^B = B \log A$$

$$= \sum_{i=1}^{m} y^i \cdot \log p^i + (1-y^i) \log(1-p^i)$$

Maximise

Class 1 | Class 0

Case 1) $y^i = 1 \rightarrow \log p^i$

Case 2) $y^i = 0 \rightarrow \log(1-p^i)$

loss fn for training a model has to be minized.

$$NLL = -\sum_{i=1}^{m} y^i \cdot \log p^i + (1-y^i)\log(1-p^i)$$

Minimize

Negative Log likelihood.

③ ② ①

Step1) Calculate likelihood of m samples occurring

Step2) Take log both sides

Step3) Take negative eq

$$\text{log loss} \begin{cases} -\log p^i & \text{if } y^i = 1 \\ -\log(1-p^i) & \text{if } y^i = 0 \end{cases}$$

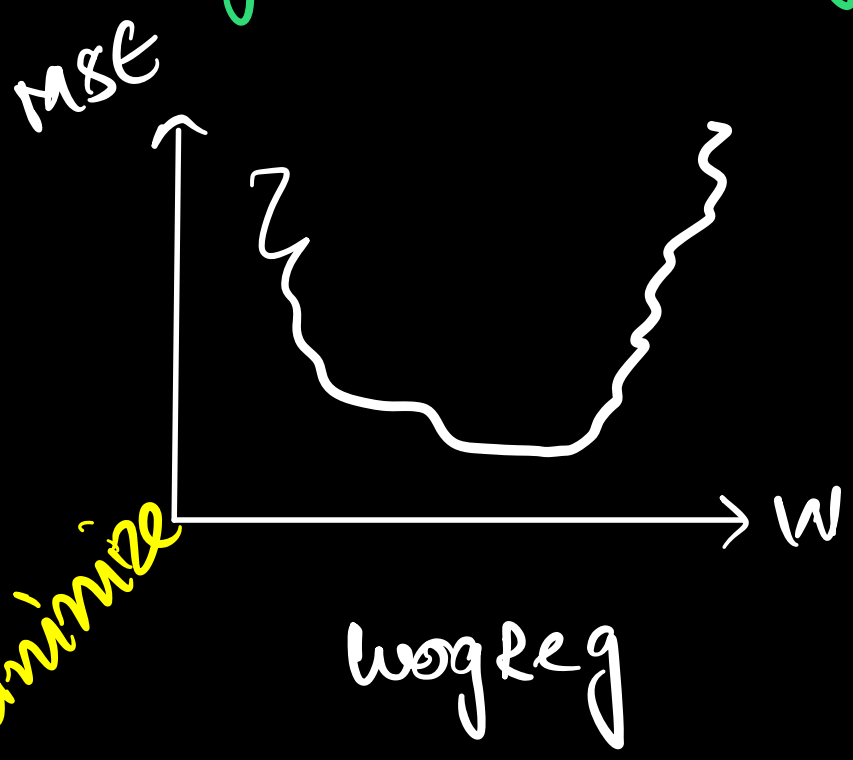
$p^i \rightarrow$ Output of log Reg model.

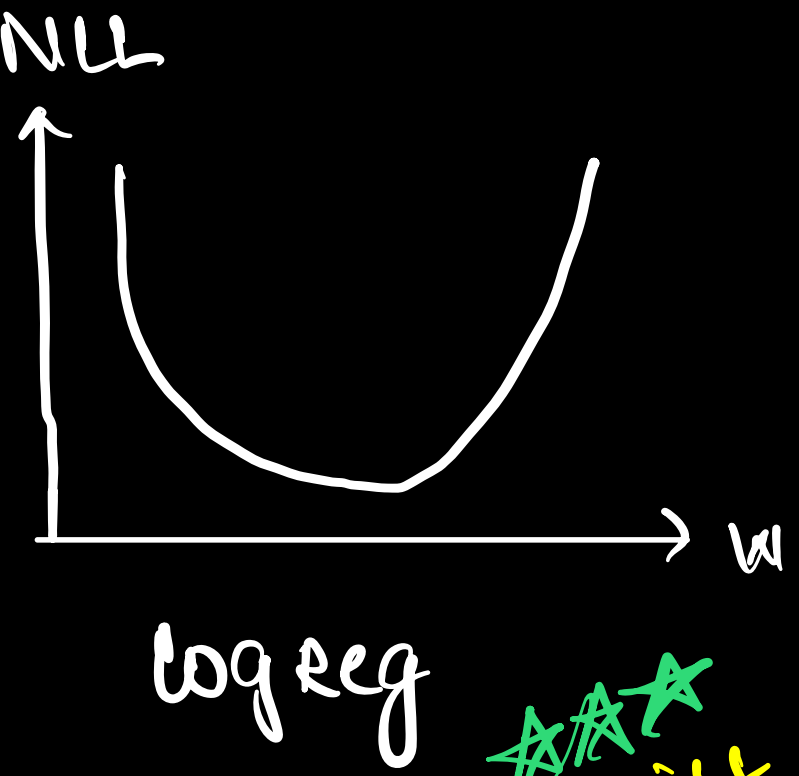
$$\text{log loss} \begin{cases} -\log \hat{y}^i & \text{if } y^i = 1 \\ -\log(1-\hat{y}^i) & \text{if } y^i = 0 \end{cases}$$

$$NLL = -\sum_{i=1}^{m} y^i \log \hat{y}^i + (1-y^i) \log (1-\hat{y}^i)$$

↑ Minimize.

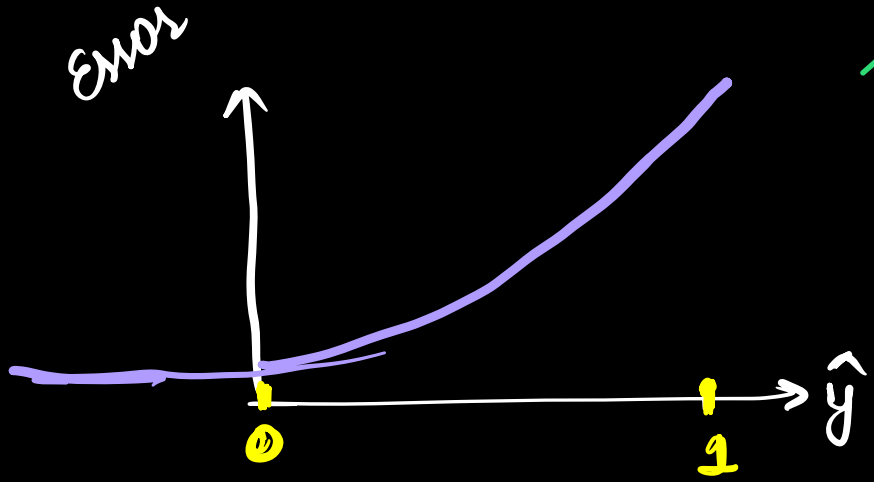
Q why not simply MSE ? $\frac{\sum (\hat{y}^i - y^i)^2}{m}$

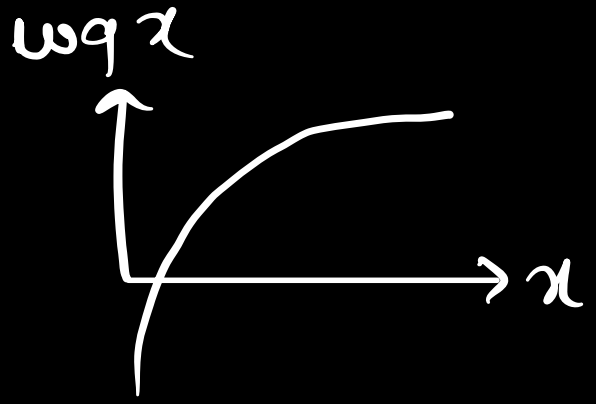




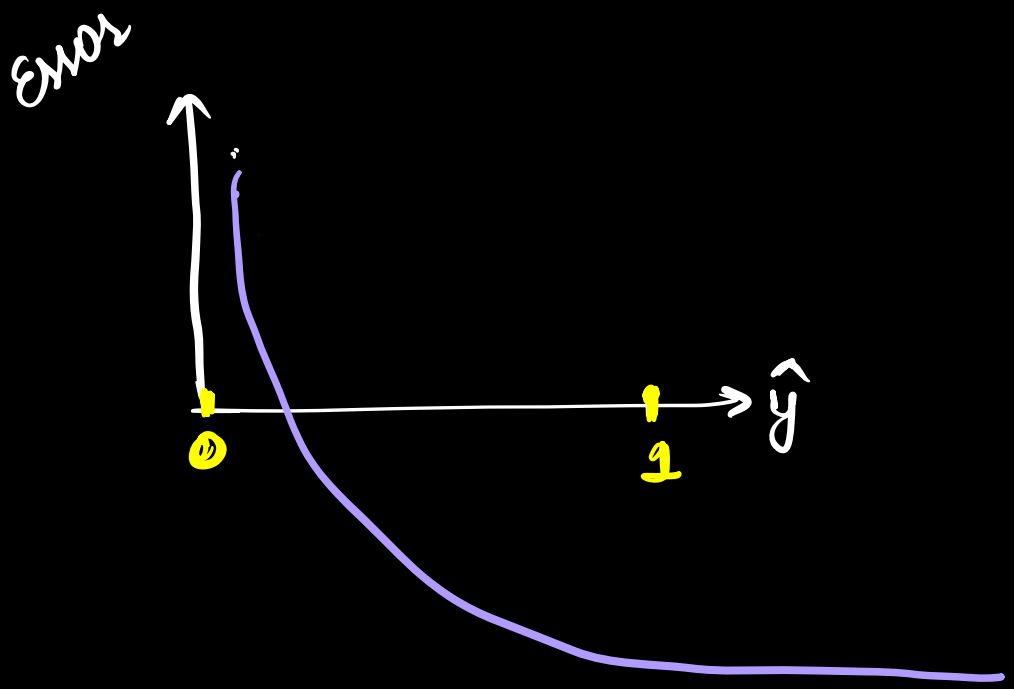
$$NLL = -\sum_{1}^{m} y^i \log \hat{y}^i + (1-y^i) \log(1-\hat{y}^i)$$

case1) $GT = 0$; $y^i = 0$; $NLL = -\log(1-\hat{y})$





Case 2) $GT = 1$, $y^i = 1$, $NLL = -\log(y^i)$



Thing to remember:

$$NLL = -\sum_{1}^{m} y^i \log \hat{y}^i + (1-y^i) \log(1-\hat{y}^i)$$