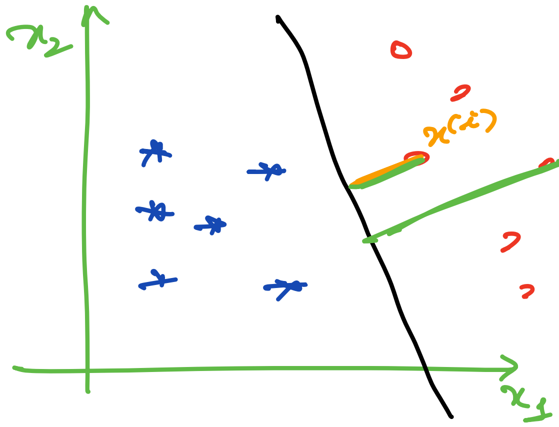
Logistic Regression

↳ Classification



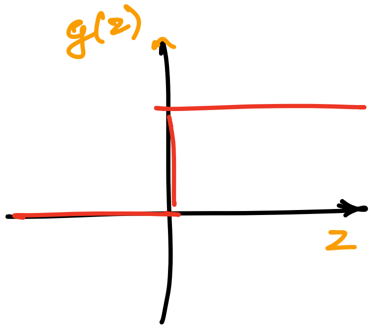
Linear Regression

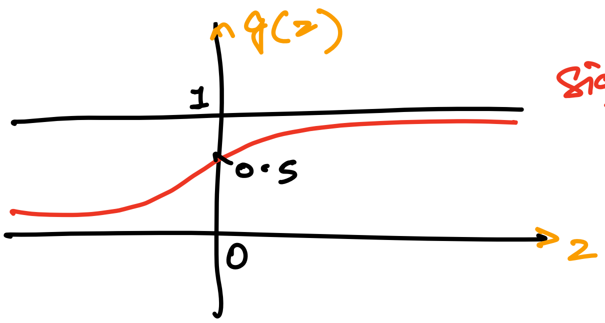
$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Real no.

$$h_\theta(x^{(i)}) = \underbrace{\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}}$$

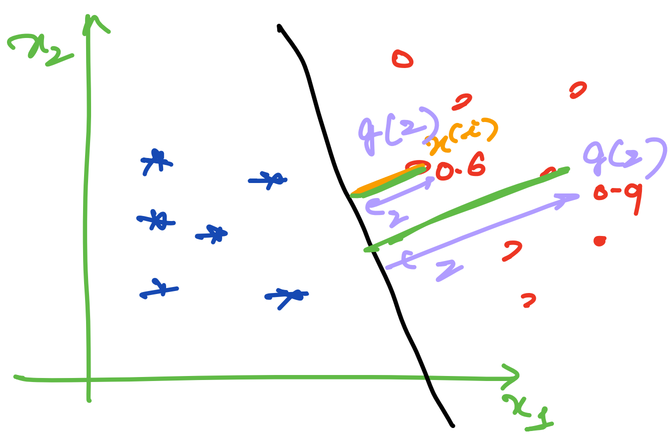
$$= \underbrace{\theta^T x^{(i)}}_{z}$$





Sigmoid fxn

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



$$g(z) \rightsquigarrow [0 - 1]$$

$$\boxed{\begin{aligned} h_\theta(x) &= g(z) = g(\theta^T x) \\ &= \frac{1}{1+e^{-\theta^T x}} \quad \text{Probability that o/p belongs to class 1} \\ \hat{y} &= 1 \quad \text{if } h_\theta(x) \geq 0.5 \\ &= 0 \quad \text{if } h_\theta(x) < 0.5 \end{aligned}}$$

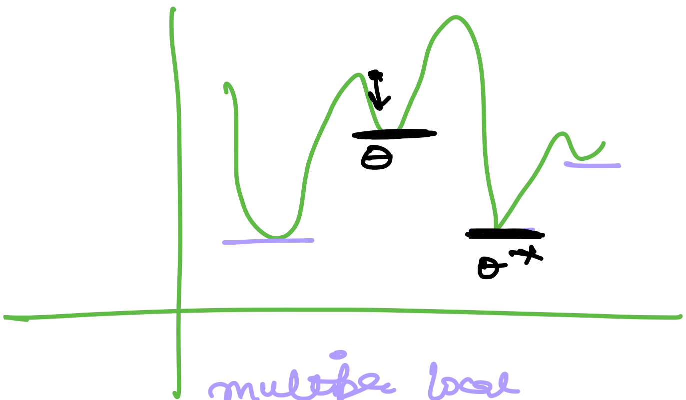
Loss fx^n

Linear Regression: $\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right)^2$ → Mean squared Error loss

$$\frac{1}{1+e^{-\theta^T x}}$$

$(1 - 0.8)^2$
$+$
$(1 - 0.2)^2$
$+$
$\vdots$
$(1 - 0.6)^2$

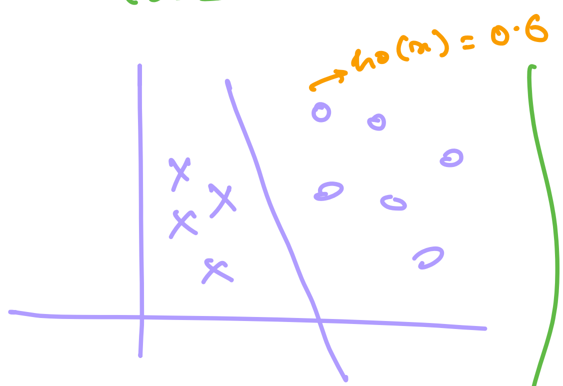
Loss fx^n → non convex fx^n ⟹ ✗



multiple local minimas

BINARY CROSS ENTROPY OR LOG LOSS

$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ → used to tell the prob. that point x belongs to class 1.



$P(y=1 \mid x;\theta) = h_\theta(x)$

$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$

(Probability Mass fx^n) ⟹

$P(y \mid x;\theta) = [h_\theta(x)]^{y}\,[1-h_\theta(x)]^{1-y}$

(0.6) 60% → class 1
(0.4) 40% → class 0

put $y=1$ | put $y=0$

$h_\theta(x)$ | $1-h_\theta(x)$

Bernoulli Distribution

10 tosses $H = 1/3$

$T = 2/3$

$= {}^{10}C_6 \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)^4$

<u>likelihood</u>

$$P(y^{(1)} y^{(2)} \ldots y^{(m)} \mid x^{(1)} x^{(2)} \ldots x^{(m)} ; \theta)$$

$$P(y^{(1)} \mid x^{(1)}; \theta) \cdot P(y^{(2)} \mid x^{(2)}; \theta) \ldots\ldots P(y^{(m)} \mid x^{(m)}; \theta)$$

$$\prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta)$$

likelihood of the data y.

$$\prod_{i=1}^{m} \left[h_\theta(x^{(i)})\right]^{y^{(i)}} \left[1-h_\theta(x^{(i)})\right]^{1-y^{(i)}}$$

<u>log likelihood</u>

$$\log \prod_{i=1}^{m} \left[h_\theta(x^{(i)})\right]^{y^{(i)}} \left[1-h_\theta(x^{(i)})\right]^{1-y^{(i)}}$$

$\log(a \cdot b \cdot c \cdot d) = \log a + \log b + \log c + \log d$

$$\sum_{i=1}^{m} \log\left(\left[h_\theta(x^{(i)})\right]^{y^{(i)}} \left[1-h_\theta(x^{(i)})\right]^{1-y^{(i)}}\right)$$

$$\sum_{i=1}^{m} \log\left(h_\theta(x^{(i)})^{y^{(i)}}\right) + \log\left(\left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}}\right)$$

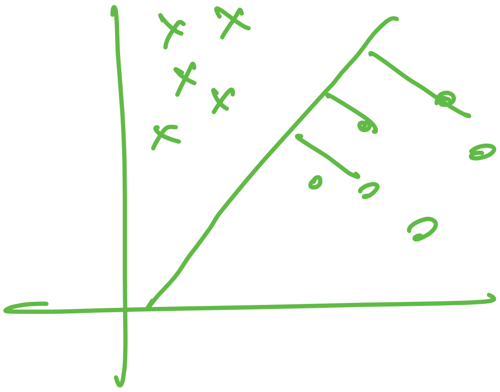
$$\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log\left(1-h_\theta(x^{(i)})\right)$$

$$\sum_{i=1}^{m} y^{(i)} \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(1 - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right)$$

$$\sum_{i=1}^{m} y^{(i)} \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right)$$

→ log likelihood → $LL(\theta)$

→ maximized



→ optimisation

→ GRADIENT DESCENT

(differentiate)

maximize log likelihood

minimize − log likelihood

$$LL(\theta) = \sum_{i=1}^{m} y^{(i)} \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right)$$

$$\frac{\partial}{\partial\theta} LL(\theta) = \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{m} y^{(i)} \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right)\right)$$

$$= \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{m} y^{(i)} \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(e^{-\theta^T x^{(i)}}\right)\right)$$

$$= \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{m} \log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right) + (1-y^{(i)}) \log\left(e^{-\theta^T x^{(i)}}\right)\right)$$

$$= \frac{\partial}{\partial \theta}\left(\sum_{i=1}^{m} -\log\left(1+e^{-\theta^T x^{(i)}}\right) + (1-y^{(i)})(-\theta^T x^{(i)})\right)$$

$$= \sum_{i=1}^{m}\left(\frac{\partial}{\partial \theta}\left(-\log\left(1+e^{-\theta^T x^{(i)}}\right)\right) + (1-y^{(i)})\frac{\partial}{\partial \theta}(-\theta^T x^{(i)})\right)$$

$$= \sum_{i=1}^{m}\left(-\frac{1}{1+e^{-\theta^T x^{(i)}}}(e^{-\theta^T x^{(i)}})(-x^{(i)}) + (1-y^{(i)})(-x^{(i)})\right)$$

$$= \sum_{i=1}^{m}\left((1-y^{(i)}) - \frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right)(-x^{(i)})$$

$$= \sum_{i=1}^{m}\left(1 - \frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}} - y^{(i)}\right)(-x^{(i)})$$

$$\frac{\partial (LL(\theta))}{\partial \theta} = \sum_{i=1}^{m}\left(\frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)}\right)(-x^{(i)})$$

$$\frac{\partial (LL(\theta))}{\partial \theta} = \sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right)(x^{(i)})$$

Gradient of log likelihood seen as a parameter of θ

Actual label

$h_\theta(x^{(i)})$

$$\frac{1}{m}\frac{\partial (LL(\theta))}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right)(x^{(i)})$$

Instead of optimizing LL

$\rightarrow$ optimize average LL

$$\frac{\partial}{\partial \theta} LL(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \underline{\underline{h_\theta(x^{(i)})}} \right) x^{(i)}$$

Linear Regression

$\theta^T x^{(i)}$

Logistic Regression

$$\frac{1}{1+e^{-\theta^T x^{(i)}}}$$