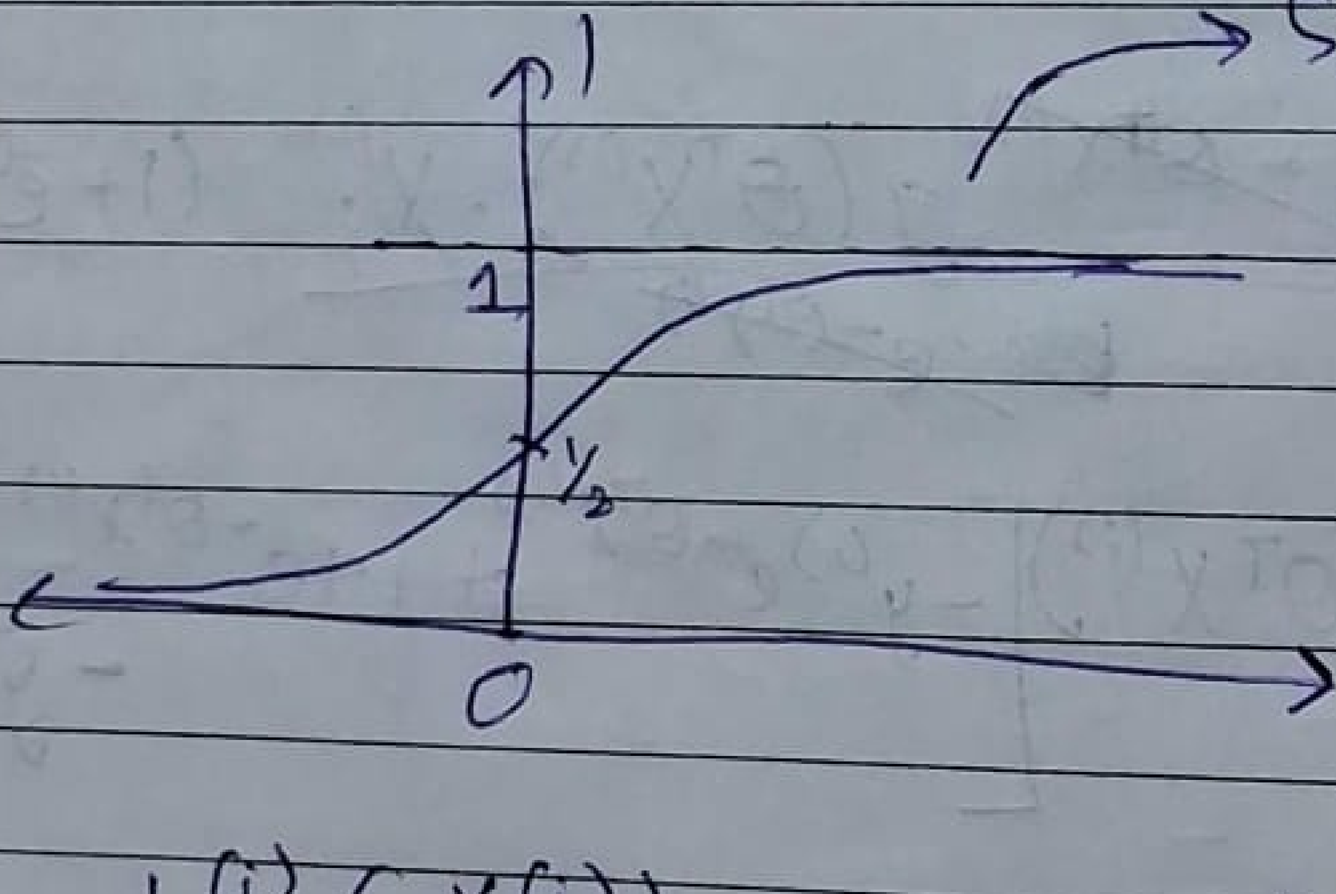
i^{th} training example (instance)

$$\begin{array}{cc} & X_j^{(i)} \\ X^{(i)} = & \begin{bmatrix} 20 \\ 95 \\ \vdots \\ \vdots \end{bmatrix} \end{array} \quad X_j^{(i)}$$

m instances having n features

$$\Theta^T X^{(i)} = z^{(i)} \rightarrow \text{"Logits"}$$

(n+1) parameters (n+1) features



Sigmoid Fn

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

$$\underbrace{h_\Theta^{(i)}(X^{(i)})}_{\text{estimated Probability}} = \sigma(z^{(i)})$$

Assume

$$\begin{cases} P(y=1 \mid x;\theta) = h_\theta(x). \\ P(y=0 \mid x;\theta) = 1-h_\theta(x). \end{cases}$$

$\because y \in \{0,1\}$

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

$$\mathcal{L}(\theta) = p(\vec{y} \mid X;\theta)$$

(Likelihood)

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

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

$l(\theta) = \log \mathcal{L}(\theta)$

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

<u>Goal</u>: maximize $l(\theta)$

Define

$$J(\theta) := -\frac{1}{m} l(\theta) \rightarrow \text{Negative Avg log-Likelihood}$$

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

$\rightarrow$ Cross-Entropy Loss

$$\Theta_j := \Theta_j - \alpha \frac{\partial J(\Theta)}{\partial \Theta_j}$$

$$\frac{\partial J}{\partial l} = -\frac{1}{m}$$

bcoz $h_\theta(x^{(i)}) = \sigma(z^{(i)})$

$$\frac{\partial l}{\partial \Theta_j} = \frac{\partial l}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial \Theta_j}$$

$$= \left[\frac{y^{(i)}}{h_\theta} - \frac{(1-y^{(i)})}{1-h_\theta} \right] \left[\sigma(z)(1-\sigma(z)) \right] x_j$$

$$= \left[\frac{y}{h} - \frac{(1-y)}{(1-h)} \right] \left[h(1-h) \right] x_j^{(i)}$$

$$= \frac{(y - \cancel{yh} - h + \cancel{yh})}{\cancel{h(1-h)}} \times \cancel{h(1-h)} \times x_j^{(i)} = (y-h) \cdot x_j^{(i)}$$

Date
Page

$$\Rightarrow \frac{\partial J}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

$$\nabla_\theta J(\theta) = \frac{1}{m} X^T \left(h_\theta(X) - y\right)$$

$$\cancel{-\frac{1}{m} X^T \left(y - h_\theta(X)\right)}$$

$$\theta \leftarrow \theta - \alpha \nabla_\theta J(\theta)$$

Learning Rate