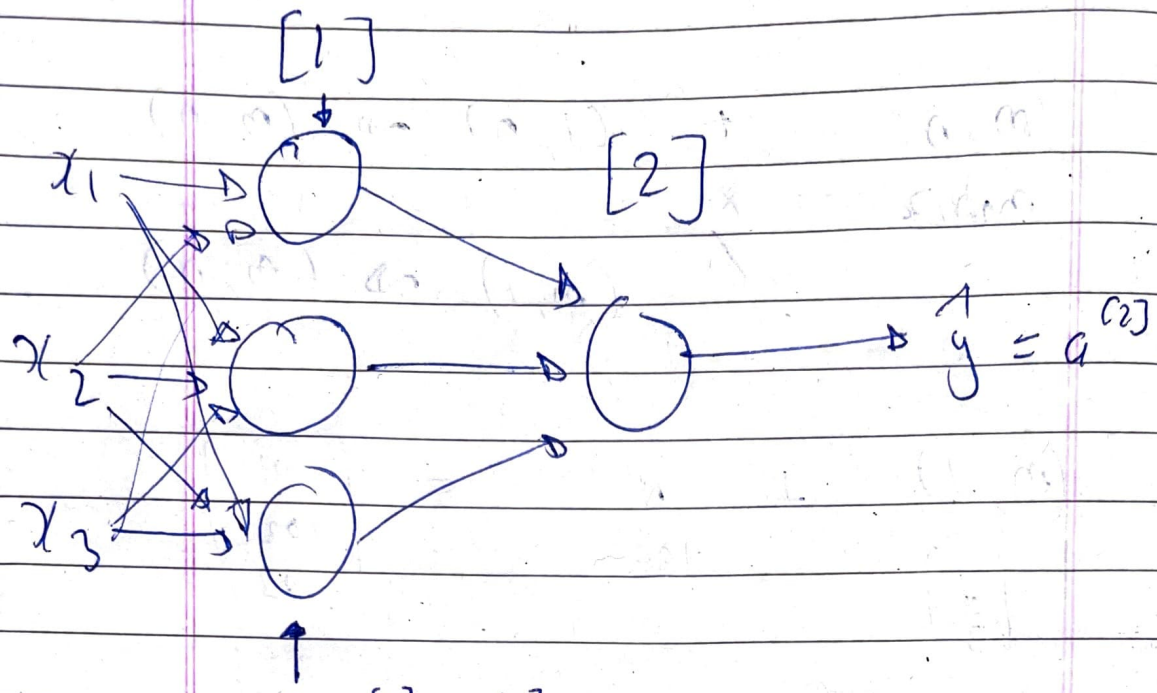
NN



$$z^{[1]} = W^{[1]} x + b^{[1]} \rightarrow a^{[1]} = \sigma(z^{[1]})$$

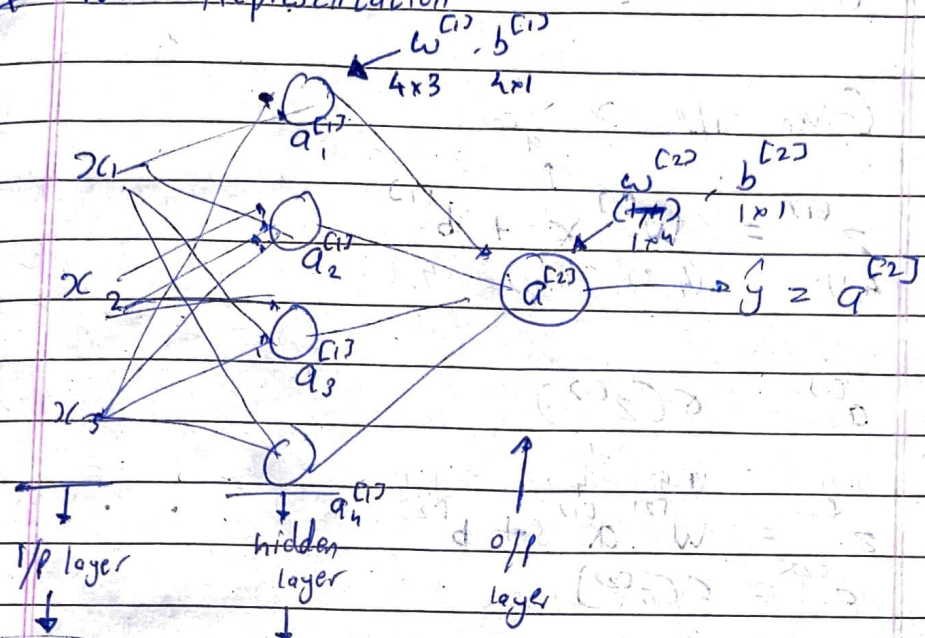
$Z^{[i]} \rightarrow i^{th}$ layer

$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = \sigma(z^{[2]})$$

$$L(a^{[2]}, y)$$

NN Representation



$\boxed{a^{[0]} = X}$

$$a^{[1]} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[2]} \\ \vdots \end{bmatrix}$$

→ 2 layer NN
(Don't count Ip layer)

$w^{[1]} \rightarrow 4 \times 3$

4 ↑ nodes in hidden layer; 3 ↑ I/p features

① On the first node

$$z_1^{[1]} = w_1^{[1]T} x + b_1^{[1]}$$

$$a_1^{[1]} = \sigma(z_1^{[1]})$$

Vectorized

$$\rightarrow \begin{bmatrix} z_1^{[1]} \\ z_2^{[2]} \\ z_3^{[1]} \\ z_4^{[1]} \end{bmatrix} = \underbrace{\begin{bmatrix} - & w_1^{[1]T} & - \\ - & w_2^{[1]T} & - \\ - & w_3^{[1]T} & - \\ - & w_4^{[1]T} & - \end{bmatrix}}_{w^{[1]}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \underbrace{\begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \\ b_4^{[1]} \end{bmatrix}}_{b^{[1]}}$$

$$a^{[1]} = \sigma(z^{[1]})$$

Given i/p x $\quad a^{[0]}$

$$\underset{(4,1)}{z^{[1]}} = \underset{(4,3)}{W^{[1]}} \underset{(3,1)}{x} + \underset{(4,1)}{b^{[1]}}$$

$$a^{[1]} = \sigma(\underset{(4,1)}{z^{[1]}})$$

$$\underset{(1,1)}{z^{[2]}} = \underset{(1,4)}{W^{[2]}} \underset{(4,1)}{a^{[1]}} + \underset{(1,1)}{b^{[2]}}$$

$$a^{[2]} = \sigma(z^{[2]})$$

$a^{[2](i)}$ ← example i

layer 2

$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & & | \end{bmatrix}$$

$$(n_x, m)$$

$$Z^{[1]} = \begin{bmatrix} | & | & & | \\ z^{[1](1)} & z^{[1](2)} & \cdots & z^{[1](m)} \\ | & | & & | \end{bmatrix}$$

$$A^{[1]} = \begin{bmatrix} | & | & & | \\ a^{[1](1)} & a^{[1](2)} & \cdots & a^{[1](m)} \\ | & | & & | \end{bmatrix}$$

hidden unit number

training examples

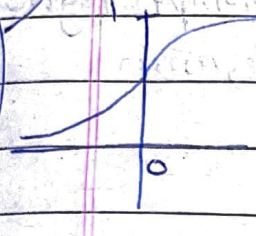
Activation Function

→ in general case

$$a^{[l]} = g(z^{[l]})$$

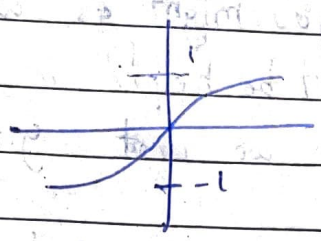
① sigmoid

when o/p 0 or 1



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

② tanh



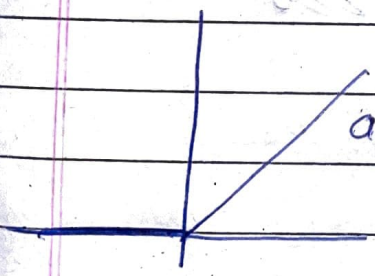
$$\tanh = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

- use tanh for all layers except o/p layer
- o/p layer needs to be 0 or 1 for binary classifn
- tanh better as you get close to zero mean ?

if you center your data, mean will be close to 0 than 0.5 makes learning next layer easier

③

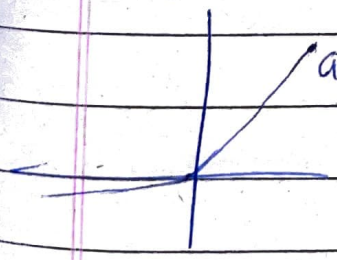
① ReLU



$a = \max(0, z)$

derivative is 0 when z is negative & 1 when z is positive

② Leaky ReLU



$a = \max(0.01z, z)$

↓

test this param

Adv → For a lot of space of Relu, derivative for activn is very diff from 0

tanh & sigmoid slow down at extremes

What if linear activation func^n ?

$$a^{[1]} = z^{[1]} = w^{[1]} x + b^{[1]}$$
$$a^{[2]} = z^{[2]} = w^{[2]} a^{[1]} + b^{[2]}$$
$$a^{[2]} = (w^{[2]} w^{[1]}) x + (w^{[2]} b^{[1]} + b^{[2]})$$
$$= w' x + b'$$

so you might as well not use multiple layers
- might be helpful in o/p layer of Regression cause we want $y \in R$

Derivatives of Activation function

① sigmoid $\frac{d}{dz} g(z) = g(z)(1 - g(z))$

② tanh $g'(z) = 1 - (g(z))^2$

③ ReLU $g'(z) = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } z \geq 0 \end{cases}$

④ Leaky ReLU $g'(z) = \begin{cases} 0.01 & z < 0 \\ 1 & z \geq 0 \end{cases}$

Gradient descent for NN

no. of features n_x or $n^{[0]}$

$n^{[1]} \rightarrow$ # hidden unit

$n^{[2]} \rightarrow$ # o/p units

Params.

$W^{[1]}$	$b^{[1]}$	$W^{[2]}$	$b^{(2)}$
$(n^{[1]}, n^{[0]})$	$(n^{[1]}, 1)$	$(n^{[2]}, n^{[1]})$	$(n^{[2]}, 1)$

Cost Func^n

$$J(w^{[1]}, w^{[2]}, b^{[1]}, b^{[2]})$$

$$= \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}, y) \qquad \hat{y} = a^{[2]}$$

Algo:

Repeat {

Compute predic^n $(\hat{y}^{(i)} \quad i \rightarrow 1 \ldots m)$

$$dw^{[1]} = \frac{dJ}{dw^{[1]}} \qquad db^{[1]} = \frac{dJ}{db^{[1]}} \qquad dw^{(2)} = \ldots \quad db^{(2)} = \ldots$$

$$w^{[1]} = w^{[1]} - \alpha\, dw^{[1]}$$

$$b^{[1]} = b^{[1]} - \alpha\, db^{[1]}$$

$$b^{[2]} = \ldots$$

$$w^{[2]} = \ldots$$

Formula:

$$Y = [y^{(1)}\ y^{(2)} \cdots y^{(m)}]$$

Forward Prop

$$Z^{[1]} = W^{[1]} X + b^{[1]}$$

$$A^{[1]} = g^{[1]}(Z^{[1]})$$

$$Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]}$$

$$A^{[2]} = g^{[2]}(Z^{[2]}) = \sigma(Z^{[2]})$$

Back Prop

$$dZ^{[2]} = A^{[2]} - Y$$

$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]T}$$

$$db^{[2]} = \frac{1}{m} np.sum(dZ^{[2]}, axis=1, keepdims=True)$$

element wise

$$dZ^{[1]} = \underbrace{W^{[2]T} dZ^{[2]}}_{(n^{[1]}, m)} * \underbrace{g^{[1]\prime}(Z^{[1]})}_{(n^{[1]}, m)}$$

$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^T$$

$$db^{[1]} = \frac{1}{m} np.sum(dZ^{[1]}, axis=1, keepdims=True)$$

$(n^{[1]}, 1)$ $(n^{[1]}, 1)$

$W^{[2]} - (n^{[2]}, n^{[1]})$

$Z^{[2]}, dZ^{[2]} - (n^{[2]}, 1)$

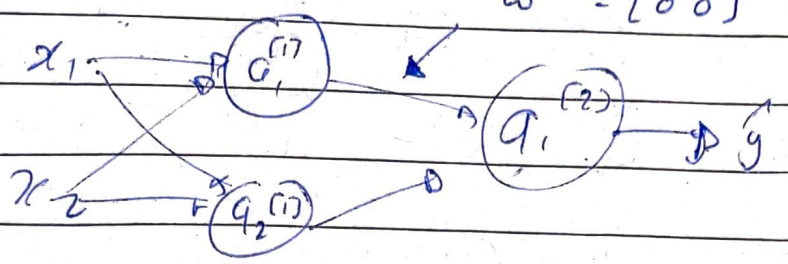
$Z^{[1]}, dZ^{[1]} - (n^{[1]}, 1)$

$$\underset{(n^{[1]},1)}{dZ^{[1]}} = \underbrace{\underset{(n^{[1]}, n^{[2]})}{W^{[2]T}} \underset{(n^{[2]},1)}{dZ^{[2]}}}_{(n^{[2]},1)} * \underset{(n^{[1]},1)}{g^{[1]\prime}(Z^{[1]})}$$

Random Initialization

→ If initialize wt to zero

$w^{[2]} = [0\ 0]$



$$w^{[1]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad b^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This will result in $a_1^{[1]} = a_2^{[1]}$

& $dz_1^{[1]} = dz_2^{[2]}$

→ Symmetric

hidden unit calculate same thing, no use

$W^{[1]} = np.random.randn((2,2)) * 0.01$

$b^{[1]} = np.zero((2,1))$

small rand value

if using sigmoid

large values

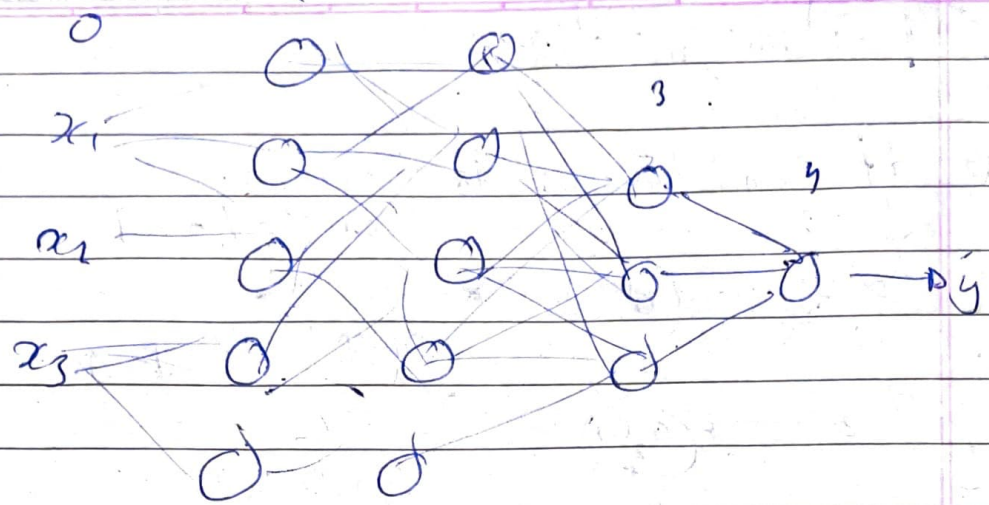
slow down descent

L-layer

PAGE NO.

DATE ...../...../.....

#



for $l = 1 \ldots 4$

$$Z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}$$

$$a^{[l]} = g^{[l]}(Z^{[l]})$$

$$a^{[0]} = X$$

$L = 5$

$Z^{[1]} = (n^{[1]}, 1)$ or $(n^{[1]}, m)$

$X = (n^{[0]}, 1)$ $(n^{[0]}, m)$

$W^{[1]} = (n^{[1]}, n^{[0]})$

$W^{[l]} = (n^{[l]}, n^{[l-1]})$

$b^{[l]} = (n^{[l]}, 1)$

Hyper params

Params: $W^{[l]}$, $b^{[l]}$ ...

Hyper param: α

iterations

hidden layer

hidden units $n^{[1]}$, $n^{[2]}$ ...

Activation funct^n.