Neural Networks:-

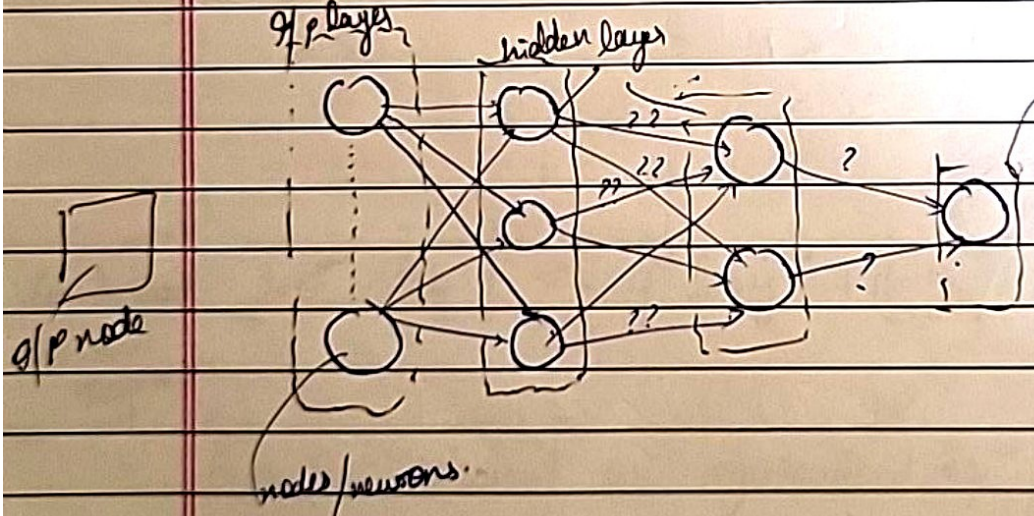
gradient descent here is termed as back propagation.

- every M.L model be it regression, classification, neural n/w etc will give probability not the exact value.

→ we can put as many hidden layers as we want in neural n/w.

→ neural n/w with 1 or 2 hidden layer is shallow neural n/w

→ neural n/w with more than 2 hidden layer is deep neural n/w.



we can feed the i/p as a matrix converted into 1-d array (each row wise) at each node of the i/p layer.

Size of i/p layer = Size of i/p (matrix)

Computation Graph.

$$J(a, b, c) = 3(a + bc)$$

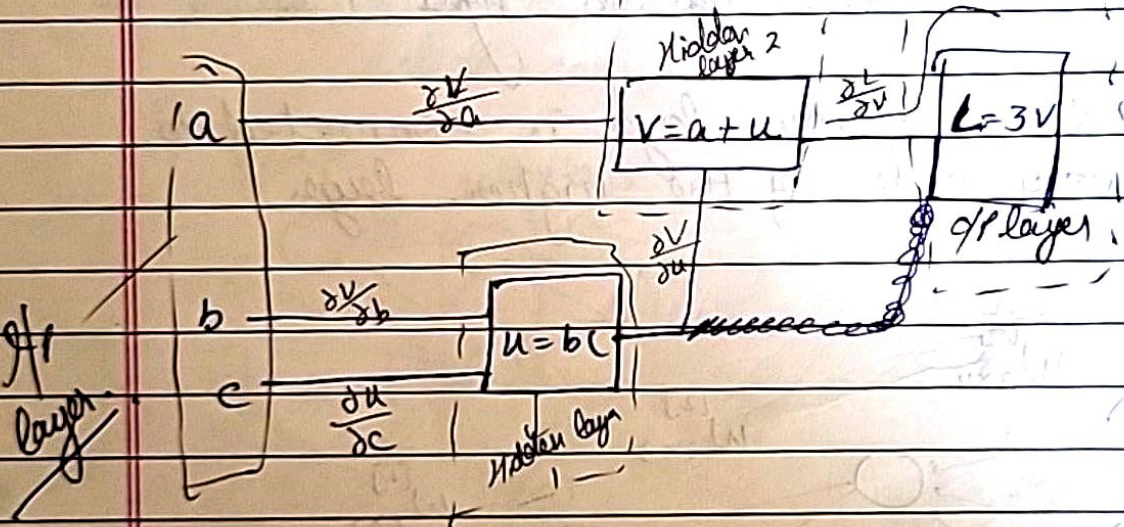
Calculate $\frac{\partial L}{\partial a}$, $\frac{\partial L}{\partial b}$, $\frac{\partial L}{\partial c}$??

Let $(a + bc) = v$.

Let $bc = u$.

- Direct calc = $\frac{\partial L}{\partial a} = 3$ $\quad \frac{\partial L}{\partial b} = 3c$ $\quad \frac{\partial L}{\partial c} = 3b$

- Computational Graph method :-



$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial v} \times \frac{\partial v}{\partial a} = 3 \cdot 1 = 3$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial v} \times \frac{\partial v}{\partial u} \times \frac{\partial u}{\partial b} = 3 \cdot 1 \cdot c = 3c$$

$$\frac{\partial L}{\partial c} = \frac{\partial L}{\partial v} \times \frac{\partial v}{\partial u} \times \frac{\partial u}{\partial c} = 3 \cdot 1 \cdot b = 3b$$

$v + w$.

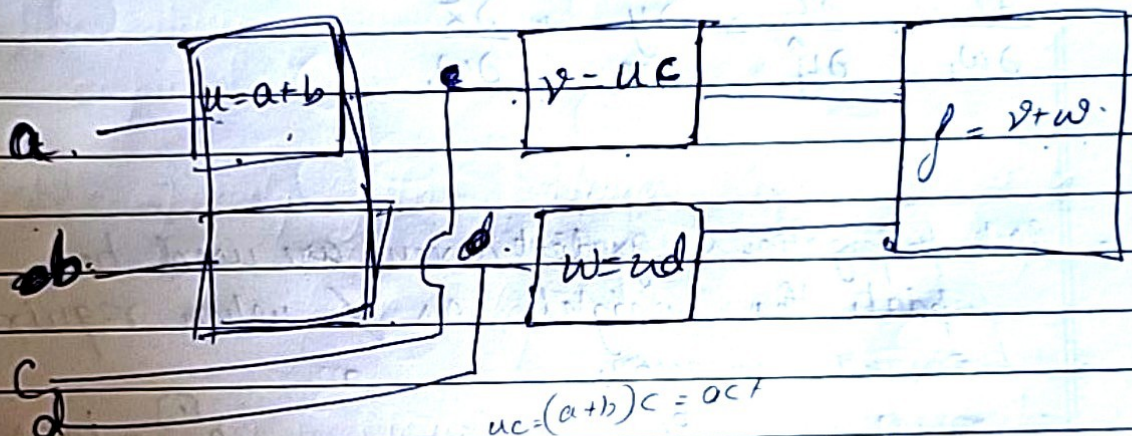
uc (u) $\quad ud$ (u)

RANKA DATE / / PAGE

Eg $f = (a+b)\,c + (a+b)\,d$

Calculate $\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}, \frac{\partial f}{\partial c}, \frac{\partial f}{\partial d} = ?$

Let $a+b = u$.
Let $uc = v$
Let $ud = w$.
Let $v + w = x$.



$uc = (a+b)c = ac + bc$

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial v} \times \frac{\partial v}{\partial u} \times \frac{\partial u}{\partial a} + \frac{\partial f}{\partial w} \times \frac{\partial w}{\partial u} \times \frac{\partial u}{\partial a}$$

$$= 1 * c * 1 + 1 \times d * 1 = c + d.$$

$$\frac{\partial f}{\partial d} = \frac{\partial f}{\partial x} \times \frac{\partial x}{\partial d} = 1.v = a+b.$$

$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial u} \times \frac{\partial u}{\partial c} = 1.v = a+b$$

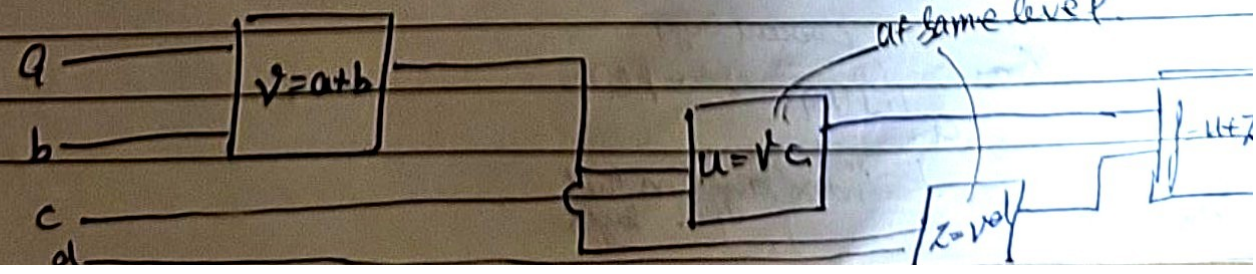
Acc to Sir :-
$a+b = v$.
$vc = u$.
$vd = x$.

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial u} \times \frac{\partial u}{\partial v} \times \frac{\partial v}{\partial a} + \frac{\partial f}{\partial x} \times \frac{\partial x}{\partial v} \times \frac{\partial v}{\partial a}$$

$$= 1 * d * 1 + 1 \times c \times 1 = c + d$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial x} \times \frac{\partial x}{\partial v} \times \frac{\partial v}{\partial b} + \frac{\partial f}{\partial u} \times \frac{\partial u}{\partial v} \times \frac{\partial v}{\partial b}$$

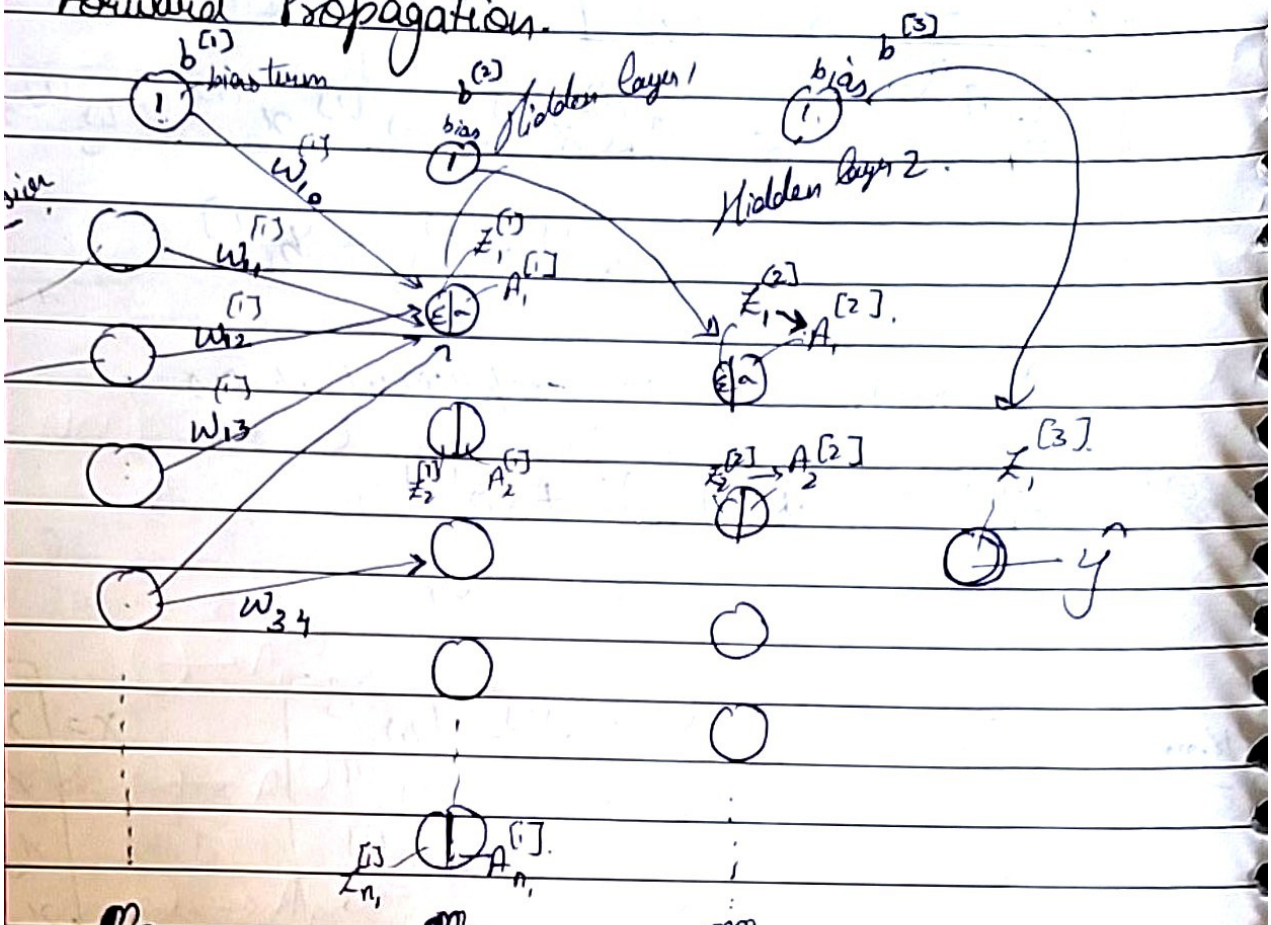
$$= 1 * d * 1 + 1 * c * 1 = c + d$$



$28 \begin{bmatrix} 28 \\ \vdots \\ \end{bmatrix} \xrightarrow{1\text{-}D} \begin{bmatrix} \\ \vdots \\ \end{bmatrix}_{784 \times 1}$

RANKA
DATE / /
PAGE

Forward Propagation.



n_0 ⇓ 784+1

n_1 It will have $A^{[1]}$ $Z^{[1]}$ $W^{[1]}$

n_2 It will have $A^{[2]}$ $Z^{[2]}$ $W^{[2]}$

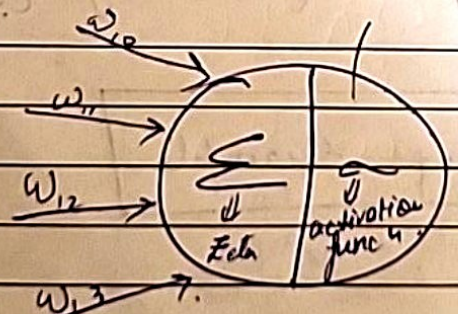
$W^{[1]}$ to prom. layer.

non-linearity

$x_0 = 1$

So $x_0 * w_0 =$ bias term.

(for each layer (i/p) a bias term will be added where the weight will change)



(The bias term will be used in the next layer)

$$\Sigma = w_1 x_1 + w_2 x_2 \ldots\ldots w_m x_m + w_0 (x_0)$$

$x_0 \rightarrow 1$

$$a\left(\sum_{j=0}^{n_0} w_j x_j\right)$$

DATE / /
PAGE

$$W^{[1]} \in R^{m_1 \times (m_0+1)}$$

$$\begin{bmatrix} w_{10}^{[1]} & w_{11}^{[1]} & \cdots & w_{1m_0}^{[1]} \\ w_{20}^{[1]} & w_{21}^{[1]} & \cdots & w_{2m_0}^{[1]} \\ \vdots & & & \\ w_{m_1 0} & w_{m_1 1} & \cdots & w_{m_1 m_0} \end{bmatrix} = w^{[1]}$$

(rows 1, 2, …, m_1 → neurons of first hidden layer)

$Z_1^{[1]}$ = 1st neuron of 1st hidden layer ke saare parameters

$$= x_1 w_{11}^{[1]} + x_2 w_{12}^{[1]} + x_3 w_{13}^{[1]} + \cdots 1 \cdot w_{10}^{[1]}$$

$$Z_1^{[1]} = \sum_{j=1}^{m_0} w_{1j}^{[1]} x_j + b_1^{[1]}$$

$$Z_3^{[1]} = \sum_{j=1}^{m_0} w_{3j}^{[1]} x_j + b_3^{[1]}.$$

So generalised k^{th} neuron of 1st hidden layer ke saare parameter

$$Z_k^{[1]} = \sum_{j=1}^{m_0} x_j w_{kj}^{[1]} + b_k^{[1]}$$

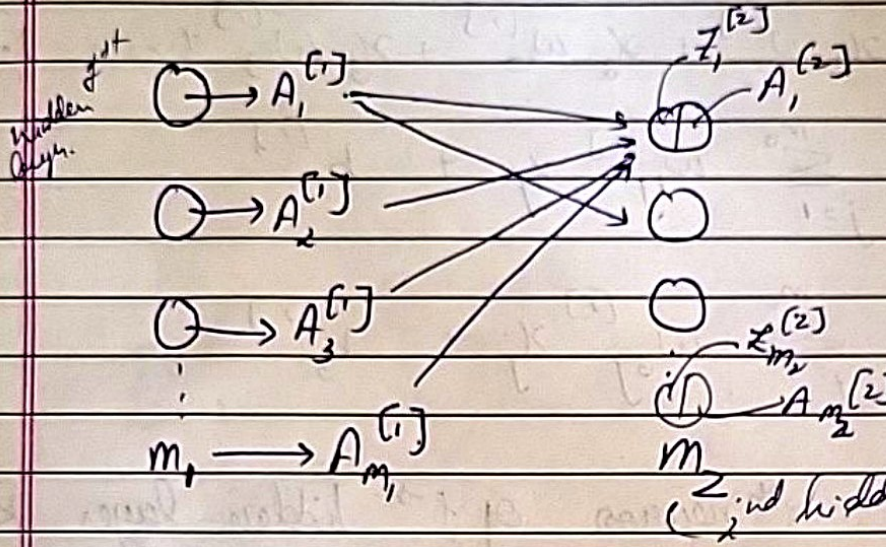
For multiple training samples or image.

1st image = $Z_1^{[1](1)} = \sum_{j=1}^{m_0} x_j^{(1)} w_{1j}^{[1]} + b_1^{[1]}$

2nd image = $Z_1^{[1](2)} = \sum_{j=1}^{m_0} x_j^{(2)} w_{1j}^{[1]} + b_1^{[1]}$

So $Z_k^{[1](i)}$ = 1st hidden layer ke k^{th} neuron of i^{th} image / example

$$= \sum_{j=1}^{m_0} x_j^{(i)} w_{kj}^{[1]} + b_k^{[1]}$$



So $Z_1^{[2]} = w_{1,1}^{[2]} A_1^{[1]} + w_{1,2}^{[2]} A_2^{[1]} + w_{1,3}^{[2]} A_3^{[1]} + \cdots$

$\cdots w_{1,m_1}^{[2]} A_{m_1}^{[1]} + b_1^{[2]}$.

$Z_2^{[2]} = w_{2,1}^{[2]} A_1^{[1]} + w_{2,2}^{[2]} A_2^{[1]} + \cdots w_{2,m_1}^{[2]} A_{m_1}^{[1]} + b_2^{[2]}$

So generalised $Z_k^{[2]} = \sum_{j=1}^{m_1} w_{kj}^{[2]} A_j^{[1]} + b_k^{[2]}$

for multiple images $Z_k^{[2](i)} = \sum_{j=1}^{m_1} w_{kj}^{[2]} A_j^{[1](i)} + b_k^{[2]}$

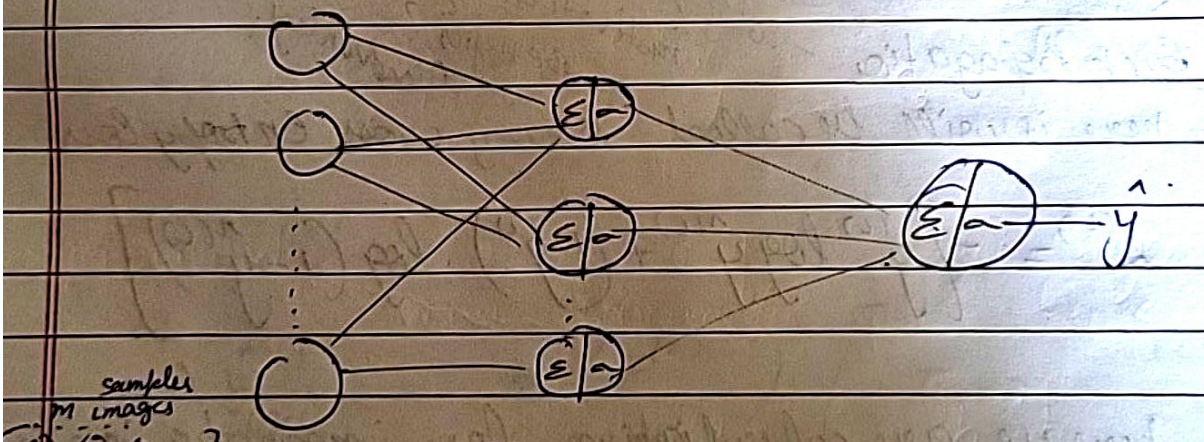
$$A_k^{[2](i)} = a\left(Z_k^{[2](i)}\right)$$

for Output or 3rd layer

$$Z_\beta^{[3](i)} = \sum_{j=1}^{m_2} w_{\beta j}^{[3]} A_j^{[2](i)} + b_\beta^{[3]}$$

$$A_\beta^{[3](i)} = a\left(Z_\beta^{[3](i)}\right) = \hat{y}$$

unvectorised (1 eg by 1 eg.)



$$x^{(i)} \in \mathbb{R}^{n^{[0]} \times 1}$$

$$w^{[1]} \in \mathbb{R}^{n^{[1]} \times n^{[0]}} \qquad b^{[1]} \in \mathbb{R}^{n^{[1]} \times 1}$$

$$w^{[2]} \in \mathbb{R}^{1 \times n^{[1]}} \qquad b^{[2]} \in \mathbb{R}^{1 \times 1}$$

epoch in range(1,10):
i in range (1, m):

$$z^{[1](i)} = \underbrace{w^{[1]} x^{(i)}}_{(n^{[1]} \times n^{[0]})(n^{[0]} \times 1)} + \underbrace{b^{[1]}}_{n^{[1]} \times 1}$$

$$n^{[1]} \times 1$$

RANKA
DATE / /
PAGE

any non-linear activation func ^{n}.

$$A = \begin{bmatrix} A_1^{[1](i)} \\ A_2^{[1](i)} \\ \vdots \\ A_{n^{[1]}}^{[1](i)} \end{bmatrix}_{n^{[1]} \times 1}$$

$$A^{[1](i)} = g\left(\underbrace{Z^{[1](i)}}_{n^{[1]} \times 1}\right)$$

$A^{[1](i)}$: $n^{[1]} \times 1$

So now

$$Z^{[2](i)} = \underset{1 \times n^{[1]}}{w^{[2]}} \; \underset{n^{[1]} \times 1}{A^{[1](i)}} + \underset{1 \times 1}{b^{[2]}}$$

$Z^{[2](i)}$: 1×1

$$\underset{1 \times 1}{A^{[2](i)}} = a\left(\underset{1 \times 1}{Z^{[2](i)}}\right) = \hat{y}^{(i)}$$

Back Propagation - why? motivation? benefits? Disadvantage?

Here it will be called Binary cross entropy loss

$$\mathcal{L}^{(i)} = -\left[y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log (1-\hat{y}^{(i)})\right]$$

here we are calculating loss image by image.

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}^{(i)}$$

So here we have to calculate.

$$w^{[1]} = w^{[1]} - \alpha \frac{\partial \mathcal{L}}{\partial w^{[1]}} \qquad b^{[1]} = \frac{\partial \mathcal{L}}{\partial b^{[1]}}$$

$$w^{[2]} = w^{[2]} - \alpha \frac{\partial \mathcal{L}}{\partial w^{[2]}} \qquad b^{[2]} = \frac{\partial \mathcal{L}}{\partial b^{[2]}}$$

RANKA
DATE / /
PAGE

∵ $w^{[2]}$ & $b^{[2]}$ are nearest to the o/p so we will calculate them first.

$$\frac{\partial L^{(i)}}{\partial w^{[2]}} = \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \cdot \frac{\partial \hat{y}^{(i)}}{\partial z^{[2](i)}} \cdot \frac{\partial z^{[2](i)}}{\partial w^{[2]}}$$

$$\text{So } \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} = -\left[y^{(i)} \times \frac{1}{\hat{y}^{(i)}} + (1-y^{(i)}) * \frac{1}{1-\hat{y}^{(i)}} (-1) \right]$$

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

$$= \frac{-y^{(i)} + y^{(i)}\hat{y}^{(i)} + \hat{y}^{(i)} - y^{(i)}\hat{y}^{(i)}}{\hat{y}^{(i)}(1-\hat{y}^{(i)})}$$

$$= \frac{\hat{y}^{(i)} - y^{(i)}}{\hat{y}^{(i)}(1-\hat{y}^{(i)})} \qquad \left(A^{[2](i)} = \hat{y}^{(i)}\right)$$

$$\frac{\partial \hat{y}^{(i)}}{\partial z^{[2](i)}} = a(z^{[2](i)}) \left[1 - a(z^{[2](i)})\right]$$

$$= \hat{y}^{(i)}(1-\hat{y}^{(i)})$$

$$\frac{\partial z^{[2](i)}}{\partial w^{[2]}} = A^{[1](i)}$$

$$\frac{\partial L^{(i)}}{\partial z^{[2](i)}} = \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \times \frac{\partial \hat{y}^{(i)}}{\partial z^{[2](i)}}$$

$$= dz^{[2](i)} \text{ (notation)}$$

$$\text{So. } \frac{\partial L^{(i)}}{\partial w^{[2]}} = \frac{\hat{y}^{(i)} - y^{(i)}}{\cancel{\hat{y}^{(i)}(1-\hat{y}^{(i)})}} * \cancel{\hat{y}^{(i)}(1-\hat{y}^{(i)})} * A^{[1](i)}$$

$$= \underset{(1\times 1)}{(\hat{y}^{(i)} - y^{(i)})} * \underset{(1\times 1)}{A^{[1](i)}}$$

$$\downarrow n^{[1]} \times 1$$

but $w^{[2]} \in \mathbb{R}^{1 \times n^{[1]}}$ So we need to take $(A^{[1](i)})^T$

$$dw^{[2]} = \frac{\partial \mathcal{L}^{(i)}}{\partial w^{[2]}} = \underset{1\times 1}{\left(\hat{y}^{(i)} - y^{(i)}\right)} \underset{1\times n^{[1]}}{\left(A^{[1](i)}\right)^T}$$

$$\underset{1\times n^{[1]}}{w^{[2](new)}} = \underset{1\times n^{[1]}}{w^{[2](old)}} - \alpha \underset{1\times n^{[1]}}{\left(\hat{y}^{(i)} - y^{(i)}\right)} \left(A^{[1](i)}\right)^T$$

$$\Rightarrow \frac{\partial \mathcal{L}^{(i)}}{\partial b^{[2]}} = \frac{\partial \mathcal{L}^{(i)}}{\partial \hat{y}^{(i)}} * \frac{\partial \hat{y}^{(i)}}{\partial z^{[2](i)}} * \frac{\partial z^{[2](i)}}{\partial b^{[2]}}$$

$$db^{[2]} = \frac{\partial \mathcal{L}^{(i)}}{\partial b^{[2]}} = \left(\hat{y}^{(i)} - y^{(i)}\right) = d\left(z^{[2]}\right)$$

$$\Rightarrow \frac{\partial \mathcal{L}^{(i)}}{\partial w^{[1]}} = \frac{\partial \mathcal{L}^{(i)}}{\partial \hat{y}^{(i)}} * \frac{\partial \hat{y}^{(i)}}{\partial z^{[2](i)}} * \frac{\partial z^{[2](i)}}{\partial A^{[1](i)}} * \frac{\partial A^{[1](i)}}{\partial z^{[1](i)}} * \frac{\partial z^{[1](i)}}{\partial w^{[1]}}$$

$dw^{[2]} =$

$$\frac{\partial Z^{[2](i)}}{\partial A^{[1](i)}} = W^{[2]}$$

$$\frac{\partial A^{[1](i)}}{\partial Z^{[1](i)}} = g'\left(Z^{[1](i)}\right)$$

$$\frac{\partial Z^{[1]X^{(i)}}}{\partial W^{[1]}} = X^{(i)}$$

$$\frac{\partial L^{(i)}}{\partial w^{[1]}} = \left(\hat{y}^{(i)} - y^{(i)}\right) * \frac{\hat{y}^{(i)}\left(1-\hat{y}^{(i)}\right)}{\hat{y}^{(i)}\left(1-\hat{y}^{(i)}\right)} * w^{[2]} * g'(z^{[1](i)}) * x^{(i)}$$

$n^{[1]} \times n^{[0]}$; 1×1 ; $1 \times n^{[1]}$; $n^{[1]} \times 1$; 1×1

we need to take transpose

$$\frac{\partial L^{(i)}}{\partial w^{[1]}} = \left(\hat{y}^{(i)} - y^{(i)}\right) * w^{[2]} * g'(z^{[1](i)}) * \left(x^{(i)}\right)^T$$

$$dw^{[1]} = \frac{\partial L^{(i)}}{\partial w^{[1]}} = \left(w^{[2]}\right)^T \left(\hat{y}^{(i)} - y^{(i)}\right) \odot g'(z^{[1](i)}) * \left(x^{(i)}\right)^T$$

$n^{[1]} \times 1$; 1×1 ; $n^{[1]} \times 1$; $1 \times n^{[0]}$

$\odot$ element by element multiplication of the 2 matrix

$$\Rightarrow \frac{\partial L^{(i)}}{\partial b^{[1]}} = \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} * \frac{\partial \hat{y}^{(i)}}{\partial z^{[2](i)}} * \frac{\partial z^{[2](i)}}{\partial A^{[1](i)}} * \frac{\partial A^{[1](i)}}{\partial z^{[1](i)}} * \frac{\partial z^{[1](i)}}{\partial b^{[1]}}$$

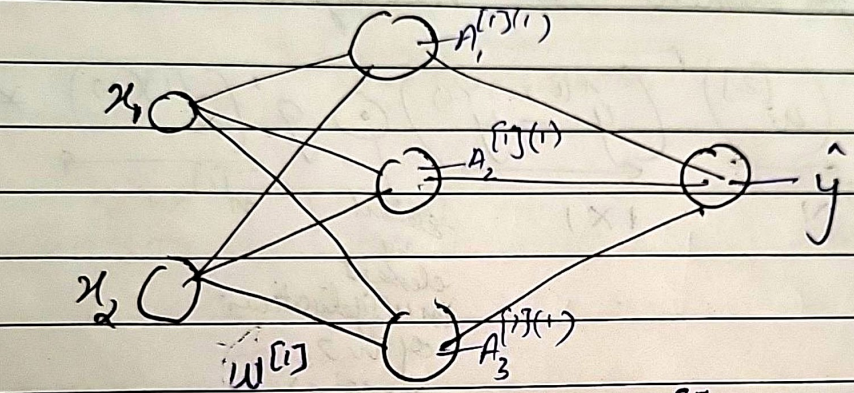
$$db^{[1]} = \left(\hat{y}^{(i)} - y^{(i)}\right) * w^{[2]} * g'(z^{[1](i)}) * 1$$

$$= \left(w^{[2]}\right)^T \left(\hat{y}^{(i)} - y^{(i)}\right) \odot g'(z^{[1](i)})$$

So Gradient update of the parameters in the neural n/w

$$w^{[2]} = w^{[2]} - \alpha\, dw^{[2]}$$
$$w^{[1]} = w^{[1]} - \alpha\, dw^{[1]}$$
$$b^{[1]} = b^{[1]} - \alpha\, db^{[1]}$$
$$b^{[2]} = b^{[2]} - \alpha\, db^{[2]}$$

Q



$y_{actual} = 1$

$$X = \begin{bmatrix} 0.5 = x_1 \\ 0.3 = x_2 \end{bmatrix}$$

$W^{[1]} \in \mathbb{R}^{3 \times 2}$.

$b^{[1]} \in \mathbb{R}^{3 \times 1}$

$W^{[2]} \in \mathbb{R}^{1 \times 3}$

$b^{[2]} \in \mathbb{R}^{1 \times 1}$

$$W^{[1]} = \begin{bmatrix} 0.2 & 0.4 \\ 0.1 & 0.3 \\ 0.5 & -0.2 \end{bmatrix}$$ will be given in exam.

$$b^{[1]} = \begin{bmatrix} 0.1 \\ -0.1 \\ 0.05 \end{bmatrix} \qquad W^{[2]} = \begin{bmatrix} 0.3 & -0.2 & 0.4 \end{bmatrix}$$

$$b^{[2]} = \begin{bmatrix} -0.2 \end{bmatrix}$$

Target → get new value of w from old values.

$$z^{[1](1)} = W^{[1]} x^{(1)} + b^{[1]}$$

$$= \underset{3\times 2}{\begin{bmatrix} 0.2 & 0.4 \\ 0.1 & 0.3 \\ 0.5 & -0.2 \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} 0.5 \\ 0.3 \end{bmatrix}} + \underset{3\times 1}{\begin{bmatrix} 0.1 \\ -0.1 \\ 0.05 \end{bmatrix}}$$

$$= \begin{bmatrix} 0.22 \\ 0.14 \\ 0.19 \end{bmatrix} + \begin{bmatrix} 0.1 \\ -0.1 \\ 0.05 \end{bmatrix} = \underset{3\times 1}{\begin{bmatrix} 0.32 \\ 0.04 \\ 0.24 \end{bmatrix}}$$

$a = \frac{1}{1+e^{-x}}$.

$$A^{[1](1)} = a\left(Z^{[1](1)}\right) = a \begin{bmatrix} 0.32 \\ 0.04 \\ 0.24 \end{bmatrix} = \begin{bmatrix} a(0.32) \\ a(0.04) \\ a(0.24) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{1+e^{-(0.32)}} \\ \frac{1}{1+e^{-(0.04)}} \\ \frac{1}{1+e^{-(0.24)}} \end{bmatrix} = \begin{bmatrix} 0.579 \\ 0.510 \\ 0.559 = 0.560 \end{bmatrix}$$

RANKA
DATE / /
PAGE

$$z^{[2](1)} = w^{[2]} A^{[1](1)} + b^{[2]}$$

$$= \underset{1\times3}{\begin{bmatrix} 0.3 & -0.2 & 0.4 \end{bmatrix}} \underset{3\times1}{\begin{bmatrix} 0.579 \\ 0.510 \\ 0.560 \end{bmatrix}} + \begin{bmatrix} -0.2 \end{bmatrix}$$

$$= \begin{bmatrix} .1737 + (-.102) + 0.224 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2957 \end{bmatrix} + \begin{bmatrix} -0.2 \end{bmatrix}$$

$$= 0.0957$$

$$A^{[2](1)} = a\left(z^{[2](1)}\right)$$

$$= a(0.0957) = \frac{1}{1+e^{-(0.0957)}}$$

$\hat{y} = 0.0524$. (it is not close to 1 (our goal))

how =

So now

$$dz^{[2](1)} = A^{[2](1)} - y$$

$$= 0.524 - 1 = (0.4476)$$

$$dw^{[2]} = d\left(z^{[2](1)}\right)\left(A^{[1](1)}\right)^T$$

$$= -0.476 \begin{bmatrix} 0.579 & 0.510 & 0.560 \end{bmatrix}$$

$$\frac{\partial l^{(1)}}{\partial w^{[2]}} = dw^{[2]} = \begin{bmatrix} -0.275 & -0.243 & -0.267 \end{bmatrix}$$

RANKA
DATE / /
PAGE

$$db^{[2]} = dz^{[2](1)} = (-0.476)$$

$$dw^{[2]} = (w^{[2]})^T \cdot dz^{[2](1)} \odot a'(z^{[1](1)})(x^{(1)})^T$$

$$= \begin{bmatrix} 0.3 \\ -0.2 \\ 0.4 \end{bmatrix} (-0.476) \odot \begin{bmatrix} a(0.32)(1-a(0.32)) \\ a(0.04)(1-a(0.04)) \\ a(0.24)(1-a(0.24)) \end{bmatrix} \begin{bmatrix} 0.5 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1428 \\ 0.0952 \\ -0.1904 \end{bmatrix} \odot \begin{bmatrix} \frac{1}{1+e^{-(0.32)}}\left(1-\frac{1}{1+e^{-(0.32)}}\right) \\ \frac{1}{1+e^{-(0.04)}}\left(1-\frac{1}{1+e^{-(0.04)}}\right) \\ \frac{1}{1+e^{-(0.24)}}\left(1-\frac{1}{1+e^{-(0.24)}}\right) \end{bmatrix} \begin{bmatrix} 0.5 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1428 \\ 0.0952 \\ -0.1904 \end{bmatrix} \odot \begin{bmatrix} 0.579(1-0.579) \\ 0.510(1-0.510) \\ 0.560(1-0.560) \end{bmatrix} \begin{bmatrix} 0.5 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1428 \\ 0.0952 \\ -0.1904 \end{bmatrix} \odot \begin{bmatrix} 0.244 \\ 0.250 \\ 0.246 \end{bmatrix} \begin{bmatrix} 0.5 & 0.3 \end{bmatrix}$$

3×1 3×1 1×2

$$= \begin{bmatrix} -0.0348 \\ 0.0238 \\ -0.0468 \end{bmatrix} \begin{bmatrix} 0.5 & 0.3 \end{bmatrix}$$

3×1 1×2

$$= \begin{bmatrix} -0.0175 & 0.0105 \\ 0.0120 & 0.0072 \\ -0.0235 & 0.0141 \end{bmatrix}$$

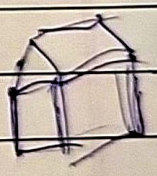
RANKA
DATE / /
PAGE

$$db^{[1]} = (w^{[2]})^T (\hat{y}^{(1)} - y^{(1)}) \odot g'(z^{[1](1)})$$

$dz^{[2](1)}$

$$= \begin{bmatrix} 0.3 \\ -0.2 \\ 0.4 \end{bmatrix} (-0.4476) \odot a(z^{[1](1)})$$

$$= \begin{bmatrix} -0.0342 \\ 0.0233 \\ -0.0468 \end{bmatrix}$$



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