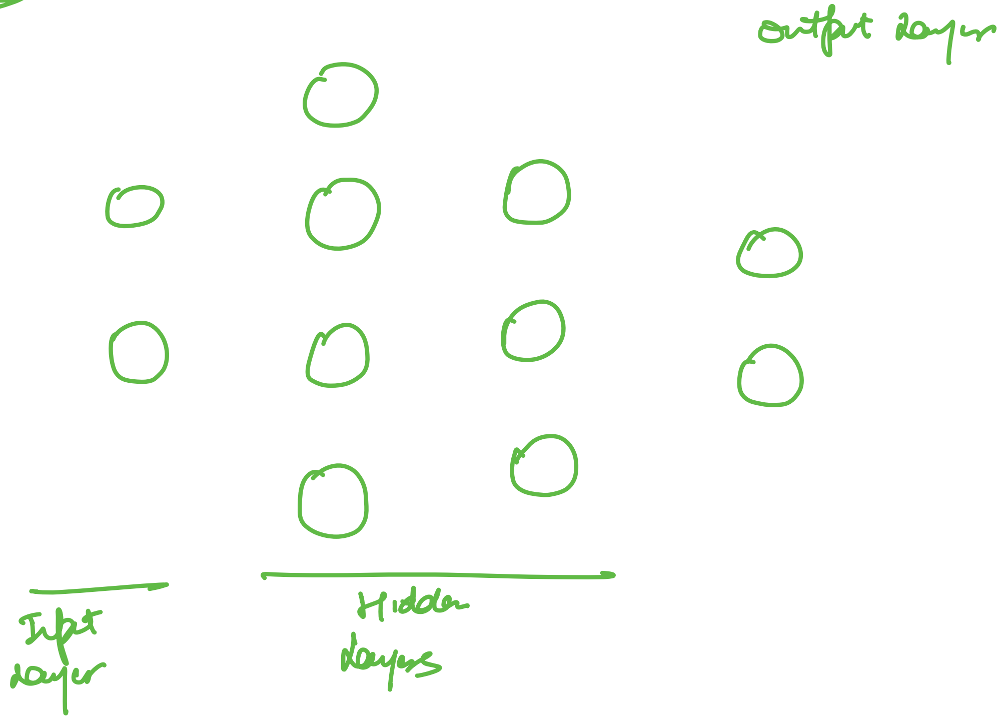
Weights bias

Loss



2

a) Output (L)

$$\frac{\partial L}{\partial w^{L}_{i,j}}$$

∂L wrt w in last layer

$$\frac{\partial L}{\partial b^{L}}$$

∂L wrt bias in last layer

b) Hidden (ℓ)

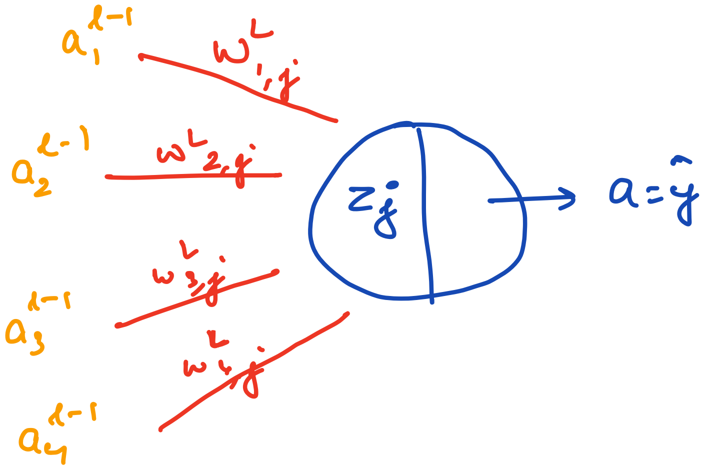
$$\frac{\partial L}{\partial w^{\ell}_{i,j}}$$

∂L wrt w in hidden layers

$$\frac{\partial L}{\partial b^{\ell}}$$

∂L wrt bias in hidden layers

Output layer

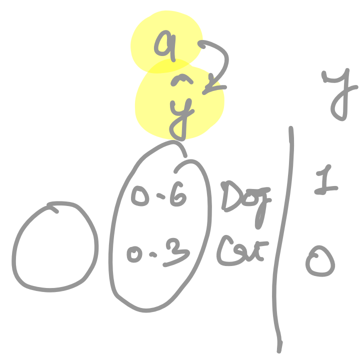


$$z_j = \sum_i w_{i,j}^L \, a_i^{l-1} + b_j^L$$

i iterates over all neurons in l-1 layer

$\frac{\partial L}{\partial w}$ $\quad w_{i,j} \to z_j \to a_j \to L(a,y)$

?



$$\frac{\partial L}{\partial w_{i,j}} = \underbrace{\frac{\partial L}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j}}_{\delta_j^L} \cdot \frac{\partial z_j}{\partial w_{i,j}}$$

(1) (2) (3)

$$\delta_j^L = \frac{\partial L}{\partial z_j}$$

(1) $\frac{\partial L}{\partial a_j}$

$$L = \frac{1}{2} \sum_i (y_i^L - a_i^L)^2$$

L	a^L	y^L
○ →	0.3	0
○ →	0.5	1

$$\frac{\partial L}{\partial a_j} = \frac{1}{2} \cdot 2 \cdot (y_j - a_j)(-1)$$

$0 \rightarrow 0.2 \mid 0$

$$\frac{\partial L}{\partial a_j} = -(y_j - a_j)$$

(2) $\dfrac{\partial a_j}{\partial z_j}$

$$a_j = \sigma(z_j)$$

$$\frac{\partial a_j}{\partial z_j} = \sigma'(z_j)$$

$$\frac{\partial a_j}{\partial z_j} = (1 - \sigma(z_j))\,\sigma(z_j)$$

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

(Sigmoid fxn)

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

$$= \frac{-1\,(e^{-z})(-1)}{(1+e^{-z})^2}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

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

$$= \sigma(z)\,(1-\sigma(z))$$

(3) $\dfrac{\partial z_j}{\partial w_{i,j}}$

$$z_j^{l} = \sum_i w_{i,j}\, a_i^{l-1} + b_j^{l}$$

$$\frac{\partial z_j}{\partial w_{i,j}} = a_i^{l-1}$$

$$\frac{\partial (w_{1j}\, a_1^{l-1} + w_{2j}\, a_2^{l-1} + \cdots)}{\partial w_{2j}}$$

Combine

$$\frac{\partial L}{\partial w_{i,j}} = \delta_j^L \cdot \frac{\partial z_j}{\partial w_{i,j}}$$

$$\boxed{\frac{\partial L}{\partial w_{i,j}} = \delta_j^L \cdot a_i^{l-1} \qquad \delta_j^L = -(y_j - a_j)\,\sigma'(z_j)}$$

Bias

$$\frac{\partial L}{\partial b_j} = \underbrace{\frac{\partial L}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j}}_{\delta_j^L} \cdot \frac{\partial z_j}{\partial b_j}$$

$$z_j^L = \sum_i w_{i,j}\, a_i^{l-1} + b_j^L$$

$$\frac{\partial z_j}{\partial b_j} = 1$$

$$\frac{\partial L}{\partial b_j} = \delta_j^L \cdot 1$$

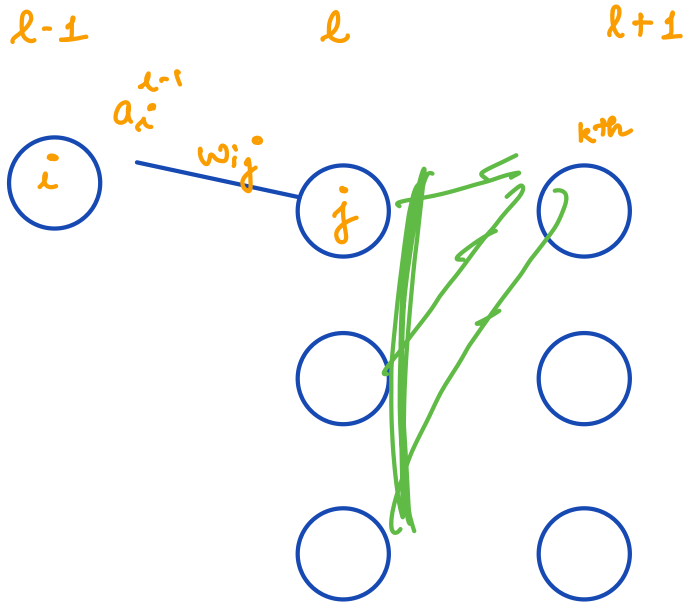
for output layer:

$$\frac{\partial L}{\partial w_{i,j}} = \delta_j^L \cdot a_i^{l-1}$$

$$\frac{\partial L}{\partial b_j} = \delta_j^L$$

$$\delta_j^L = (a_j - y_j)\,\sigma'(z_j)$$

Hidden Layer



$w_{ij} \to z_j \to a_j \to z_k \to L$

$$\frac{\partial L}{\partial w_{ij}^{l}} = \underline{\underline{\frac{\partial L}{\partial z_k^{l+1}}}} \cdot \underbrace{\frac{\partial z_k^{l+1}}{\partial a_j^{l}}}_{1} \cdot \underbrace{\frac{\partial a_j^{l}}{\partial z_j^{l}}}_{2} \cdot \underbrace{\frac{\partial z_j^{l}}{\partial w_{ij}}}_{3}$$

1 $\quad \dfrac{\partial z_k^{l+1}}{\partial a_j}$

$$z_k^{l+1} = \sum_j w_{jk}\, a_j^{l} + b^{l+1}$$

$$\frac{\partial z_k^{l+1}}{\partial a_j^{l}} = w_{jk}$$

2 $\quad \dfrac{\partial a_j}{\partial z_j}$

$$a_j = \sigma(z_j)$$

$$\frac{\partial a_j}{\partial z_j} = \sigma'(z_j)$$

3 $\dfrac{\partial z_j}{\partial w_{ij}}$

$$z_j = w_{ij} \; a_i^{l-1}$$

$$\frac{\partial z_j}{\partial w_{ij}} = a_i^{l-1}$$

$$\frac{\partial L}{\partial w_{ij}^{l}} = \sum_k \frac{\partial L}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}}$$

$$= \underbrace{\boxed{\sum_k \delta_k^{l+1} \cdot w_{jk} \cdot \sigma'(z_j^{l})}}_{\delta_j^{l}} \cdot a_i^{l-1}$$

$$\frac{\partial L}{\partial w_{ij}^{l}} = \delta_j^{l} \cdot a_i^{l-1}$$

Bias Update Rule

$$\frac{\partial L}{\partial b} = \underbrace{\boxed{\sum_k \frac{\partial L}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j}}}_{\delta_j^{l}} \cdot \underbrace{\frac{\partial z_j}{\partial b}}_{1}$$

$$\frac{\partial L}{\partial b} = \delta_j^{l} \cdot 1$$

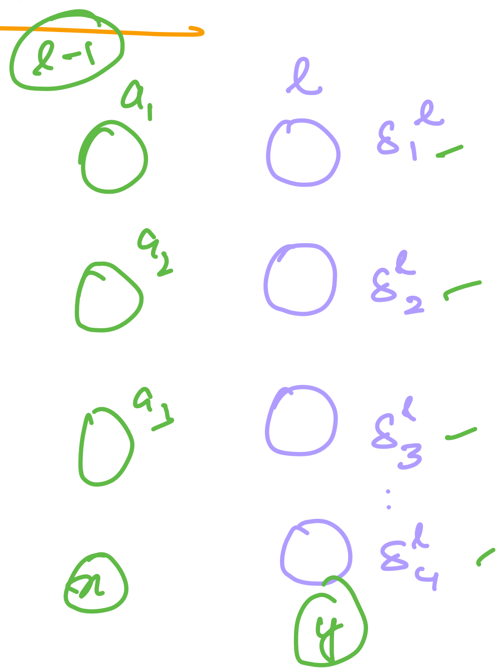
Final Results:

$$\frac{\partial L}{\partial w_{ij}^{l}} = \delta_j^{l} \cdot a_i^{l-1}$$

$$\frac{\partial L}{\partial b} = \delta_j^{l} \cdot 1$$

$$\delta_j^{l} = \sum_k \frac{\partial L}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j}$$

Matrix Representation:



$$\frac{\partial L}{\partial w_{ij}} \longrightarrow \frac{\partial L}{\partial w^{L}}$$

$$\frac{\partial L}{\partial w^{L}} = \underbrace{a_i^{l-1}}_{(x \times 1)} \cdot \underbrace{(\delta^{L})^{T}}_{(y \times 1)}$$

$$(1 \times y)$$

$$x \times y$$

$$\begin{bmatrix} a_1^{l-1} \\ a_2^{l-1} \\ \vdots \\ a_x^{l-1} \end{bmatrix} \cdot \begin{bmatrix} \delta_1^{l} & \delta_2^{l} & \delta_3^{l} & \delta_4^{l} & \cdots & \delta_y^{l} \end{bmatrix}$$

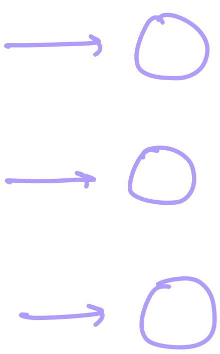
$$\frac{\partial L}{\partial w_{ij}} \leftarrow \begin{bmatrix} a_1^{l-1}\delta_1^{l} & a_1^{l-1}\delta_2^{l} & a_1^{l-1}\delta_3^{l} & \cdots \\ a_2^{l-1}\delta_1^{l} & a_2^{l-1}\delta_2^{l} & \cdots & \\ & & & \end{bmatrix}$$

$$\frac{\partial L}{\partial w^{L}} = a^{L-1} \cdot (\delta^{L})^{T}$$

For bias:

$$\frac{\partial L}{\partial b_j^{l}} = \delta_j^{l}$$

$$\frac{\partial L}{\partial b^{l}} = \delta^{L}$$



$$b = b - \eta \frac{\partial L}{\partial b}$$

$$\begin{bmatrix} b_1^{l} \\ b_2^{l} \\ b_3^{l} \end{bmatrix} \qquad \frac{\partial L}{\partial b} = \begin{bmatrix} \delta_1^{L} \\ \delta_2^{L} \\ \delta_3^{L} \end{bmatrix}$$

Loss fxn: Square error

Binary Classification: Output layer:

Dog/cat

z

$z \rightarrow a = \sigma(z) = \hat{y}$

$< 0.5 \rightarrow$ Dog

$\geq 0.5 \rightarrow$ Cat

k classes $\quad k > 2$

k neurons in output layer.

Binary Classification:

↳ Binary Cross Entropy:

$$= -\sum_{i=1}^{m} \left(y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i) \right)$$

y_i = True Label

$\hat{y}_i$ = Predicted Label

$$\delta_L = \frac{\partial L}{\partial z^l} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z}$$

$\partial a \rightarrow \hat{y}$

$\frac{\partial a}{\partial z} \rightarrow$?

$z \rightarrow \sigma(z) \xrightarrow{\hat{y}} L(y, \hat{y})$

$$\delta_L = \left(\frac{-y_i}{\hat{y}_i} + \frac{1-y_i}{1-\hat{y}_i} \right)$$

$$a = \sigma(z)$$

$$\frac{\partial a}{\partial z} = \sigma(z)\left(1-\sigma(z)\right)$$

$$= a\,(1-a)$$

$$= \hat{y}\,(1-\hat{y})$$

$$\delta_L = \left(\frac{-y_i}{\hat{y}_i} + \frac{1-y_i}{1-\hat{y}_i}\right)\left(\hat{y}_i\,(1-\hat{y}_i)\right)$$

$$= \left(\frac{-y_i + \cancel{y_i\hat{y}_i} + \hat{y}_i - \cancel{y_i\hat{y}_i}}{\cancel{\hat{y}_i\,(1-\hat{y}_i)}}\right)\cancel{\left(\hat{y}_i\,(1-\hat{y}_i)\right)}$$

$$\delta_L = \left(\hat{y}_i - y_i\right)$$

δ_L in cross Entropy loss

Multiple Examples

c: no. of classes in o/p layer

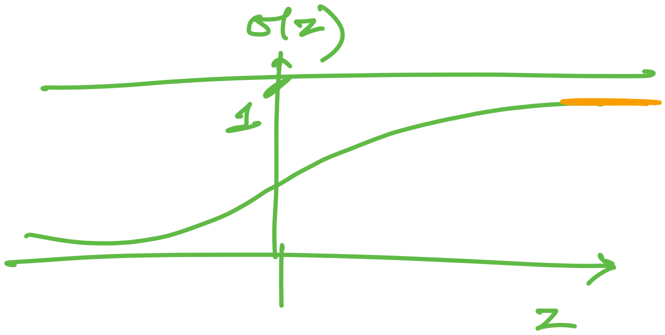
$$a_c = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \longrightarrow \begin{bmatrix} \text{——} & a^{(1)} & \text{——} \\ \text{——} & a^{(2)} & \text{——} \\ \text{——} & a^{(3)} & \text{——} \\ & \vdots & \\ \text{——} & a^{(m)} & \text{——} \end{bmatrix}$$

$$a_i = \begin{bmatrix} 0.1 \\ 0.8 \\ 0.2 \end{bmatrix} \begin{array}{l} \rightarrow \text{Dog} \\ \rightarrow \text{cat} \\ \rightarrow \text{flower} \end{array}$$

$$\begin{array}{r} 1^{st} \\ 2^{nd} \\ \vdots \\ m^{th} \end{array} \begin{bmatrix} 0.1 & 0.8 & 0.2 \\ & & \\ & & \\ & & \end{bmatrix}_{m \times c}$$

Vanishing Gradient

Sigmoid



$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$

Z is large

$$\sigma'(z) = \sigma(z)(1-\underbrace{\sigma(z)}_{1})$$

$$\underbrace{}_{0}$$

Z is small

$$\sigma'(z) = \underbrace{\sigma(z)}_{0}(1-\sigma(z))$$

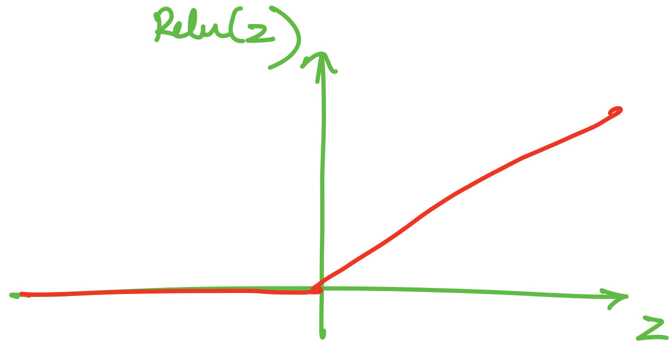
$$= 0$$

gradient

$$W = w - \eta \left(\frac{\partial L}{\partial w}\right) \rightarrow 0$$

no change in wts.

Relu



$$\text{Relu}(z) = \begin{cases} z & z \geq 0 \\ 0 & z < 0 \end{cases}$$

① Overfit?

Dropout