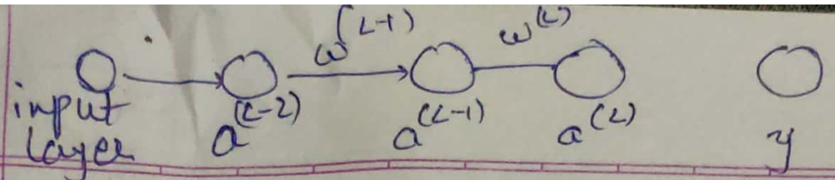
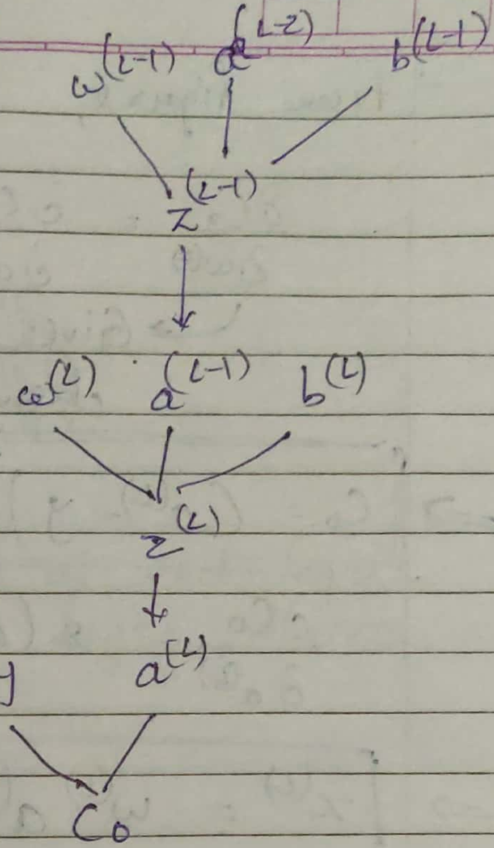


$a^{(L)}$ is predicted o/p
y is desired o/p.

cost of a single training ex. Cost$(C_0) = (a^{(L)} - y)^2$

$$a^{(L)} = \sigma\big(\underbrace{a^{(L-1)}\omega^{(L)} + b^{(L)}}_{z^{(L)}}\big)$$

$$\therefore a^{(L)} = \sigma(z^{(L)})$$



→ Goal is to know how each ~~eac~~ weights & bias affects the cost,
so, we want following quantities

$$\frac{\partial C_0}{\partial \omega^{(L)}}, \quad \frac{\partial C_0}{\partial a^{(L-1)}} \text{ and } \frac{\partial C_0}{\partial b^{(L)}}$$

→ from figure,

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial C_0}{\partial a^{(L)}} \cdot \frac{\partial a^{(L)}}{\partial z^{(L)}} \cdot \frac{\partial z^{(L)}}{\partial w^{(L)}}$$ (This is chain rule)

↳ Gives change in cost C_0 w.r.t small change in $w^{(L)}$ (weight)

⇒ $\boxed{C_0 = (a^{(L)} - y)^2}$ [as defined previously]

$$\frac{\partial C_0}{\partial a^{(L)}} = 2(a^{(L)} - y)$$

⇒ $\boxed{z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}}$ ⇒ $\boxed{a^{(L)} = \sigma(z^{(L)})}$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$ [Derivative of sigmoid f^n]

and $$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)}$$

$$\therefore \underbrace{\frac{\partial C_0}{\partial w^{(L)}}}_{\text{for a single training ex.}} = 2(a^{(L)} - y)\, \sigma'(z^{(L)})\, a^{(L-1)}$$

∴ Total cost over all examples is given by

$$\frac{\partial C}{\partial w^{(L)}} = \frac{1}{n} \sum_{k=0}^{n-1} \frac{\partial C_k}{\partial w^{(L)}}$$ [mean of all individual costs]

↳ One component of Gradient vector ∇C.

$$\nabla C = \begin{bmatrix} \frac{\partial C}{\partial w^{(1)}} \\ \frac{\partial C}{\partial b^{(1)}} \\ \vdots \\ \partial C / \partial w^{(L)} \end{bmatrix}$$

$$\frac{\partial C_0}{\partial b^{(L)}} = \frac{\partial C_0}{\partial a^{(L)}} \cdot \frac{\partial a^{(L)}}{\partial z^{(L)}} \cdot \frac{\partial z^{(L)}}{\partial b^{(L)}}$$

$$= 2(a^{(L)} - y)\, \sigma'(z^{(L)}) \cdot 1$$

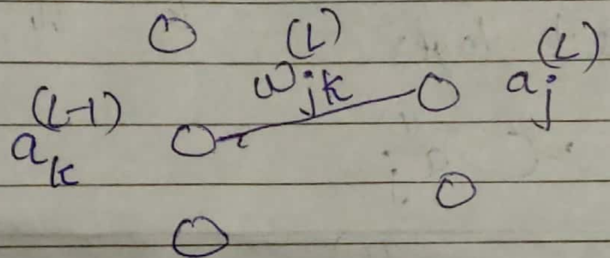
and

$$\frac{\partial C_0}{\partial a^{(L-1)}} = \frac{\partial C_0}{\partial a^{(L)}} \cdot \frac{\partial a^{(L)}}{\partial z^{(L)}} \cdot \frac{\partial z^{(L)}}{\partial a^{(L-1)}}$$

$$= 2(a^{(L)} - y)\, \sigma'(z^{(L)})\, w^{(L)}$$

→ This is for single ~~lay~~ neuron at each layer

→ For multiple neurons at each layer, following is the notation:-

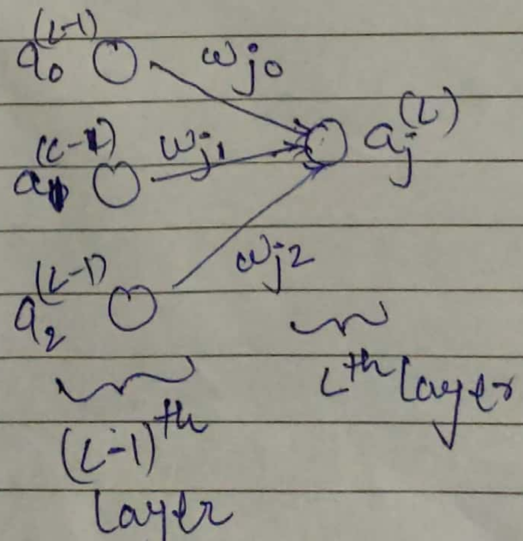


→ Equations are essentially the same.

$$\Rightarrow C_0 = \sum_{j=0}^{n_L - 1} (a_j^{(L)} - y_j)^2$$

$$\Rightarrow z_j^{(L)} = w_{j0}^{(L)} a_0^{(L-1)} + w_{j1}^{(L)} a_1^{(L-1)} + w_{j2}^{(L)} a_2^{(L-2)} + b_j^{(L)}$$

that is,



$$\Rightarrow a_j^{(L)} = \sigma(z_j^{(L)})$$

$$\frac{\partial C_0}{\partial w_{jk}^{(L)}} = \frac{\partial C_0}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L)}}$$

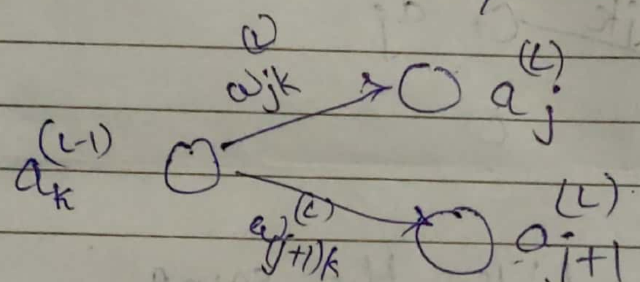
$$\frac{\partial C_0}{\partial b_j^{(L)}} = \frac{\partial C_0}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial b_j^{(L)}}$$

Same as previous

and

$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{n_L-1} \frac{\partial C_0}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}}$$

→ Summation is because each $a_k^{(L-1)}$ affects multiple neurones in next layer.



→ So, total change in C_0 is sum of all affected neurons.