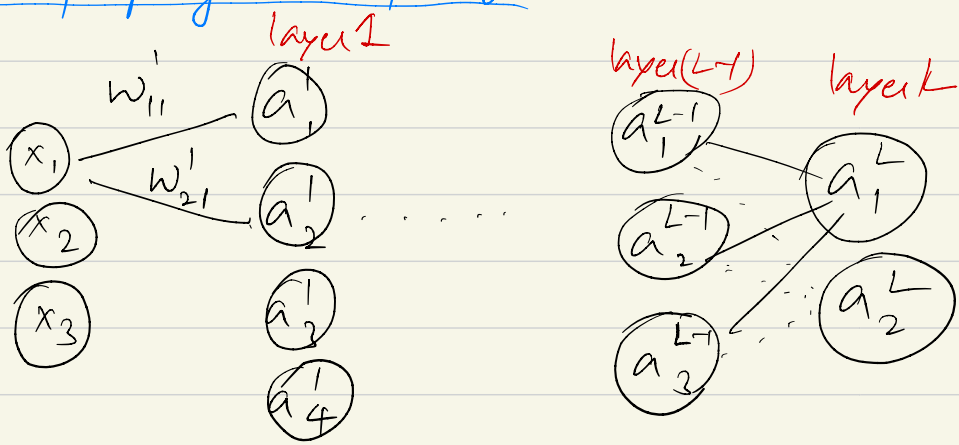
Back propagation prof (Intuition):



Example Network

what we know,

$$z^{[l]}_j = \sum_i w^{[l]}_{ji} \times a^{[l-1]}_i + b^{[l]}_j$$

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

(vectorized form) $\underset{(n_l \times 1)}{Z^{[l]}} = \underset{(n_l \times n_{l-1})}{W^{[l]}} \cdot \underset{(n_{l-1} \times 1)}{a^{[l-1]}} + \underset{(n_l \times 1)}{b^{[l]}}$

$$\underset{(n_l \times 1)}{A^{[l]}} = \sigma(\underset{(n_l \times 1)}{Z^{[l]}})$$

Cost Function : $C = \frac{1}{2} \sum (y_i^L - a_i^L)^2$

(Simple cost function is used for simplen illustration of back prop Eq proof)

As we know,

$$\frac{\partial C}{\partial w_{ji}^l} = \frac{\partial C}{\partial z_j^l} \times \frac{\partial z_j^l}{\partial w_{ji}^l} \quad \text{and} \quad \frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \times \frac{\partial z_j^l}{\partial b_j^l}$$

so, once we find $\left(\frac{\partial C}{\partial z_j^l}\right)$ it will be piece of cake

we define, $\frac{\partial C}{\partial z_j^L} = \delta_j^L$; $\frac{\partial C}{\partial z_j^l} = \delta_j^l$

what (δ_j^l) physically means ?

By how many times cost is changing, if we slightly change (z_j^l).

$$\frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} \cdot \frac{\partial a_j^L}{\partial z_j^L}$$

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

$$\boxed{\delta_j^L = (a_j^L - y_j^L) \cdot \sigma'(z_j^L)}$$

Now,

$$\delta_j^l = \frac{\partial C}{\partial z_j^l}$$

As we know change in z_j^l affects all the neurons in $(l+1) \dots (L)$ we need to keep in mind that fact.

In general,

Δz^l affects Δz^{l+1} affects C

So, we can use chain Rule here

$$\delta_j^l = \sum_K \frac{\partial C}{\partial z_K^{l+1}} \times \frac{\partial z_K^{l+1}}{\partial z_j^l}$$

$$z_j^l = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial z_j^l}$$

$$= \sum_k \delta_k^{l+1} \cdot \left(w_{kj}^{l+1} \sigma'(z_j^l) \right)$$

$$\frac{\partial z_k^{l+1}}{\partial z_j^l} \Rightarrow$$

$$z_k^{l+1} = \sum_i w_{ki}^{l+1} a_i^l + b_k^{l+1}$$

$$a_i^l = \sigma(z_i^l)$$

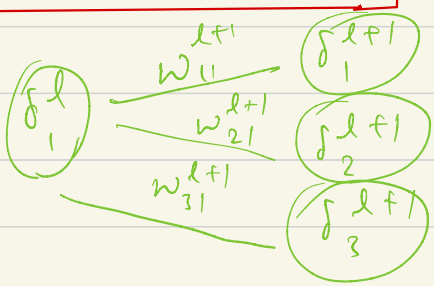
$$\frac{\partial z_k^{l+1}}{\partial z_j^l} = \frac{\partial \left[w_{kj}^{l+1} \sigma(z_j^l) + b_k^{l+1} \right]}{\partial z_j^l}$$

(b_k^{l+1} → 0)

$$= w_{kj}^{l+1} \sigma'(z_j^l)$$

$$\delta_j^l = \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \sigma'(z_j^l)$$

$$\delta_j^l = \left(\sum_k w_{kj}^{l+1} \delta_k^{l+1} \right) \cdot \sigma'(z_j^l)$$



vectorization :

$$\underset{(n_l \times 1)}{\delta^l} = \left(\underset{(n_l \times n_{l+1})}{W^{l+1}}{}^T \cdot \underset{(n_{l+1} \times 1)}{\delta^{l+1}} \right) \odot \underset{(n_l \times 1)}{\sigma'(z^l)}$$

$\odot$ ← Element wise Multiplication

$$\underset{(n_L \times 1)}{\delta^L} = \underset{(n_L \times 1)}{(A^L - y^L)} \odot \underset{(n_L \times 1)}{\sigma'(z^L)}$$

Finally,

$$\frac{\partial C}{\partial w_{ij}^{l}} = \left(\frac{\partial C}{\partial z_i^{l}}\right) \times \frac{\partial z_i^{l}}{\partial w_{ij}^{l}}$$

$$\downarrow$$

$$= \left(\delta_i^{l}\right) \times a_j^{l-1}$$

$$z_i^{l} = \sum_j a_j^{l-1} w_{ij}^{l} + b_i^{l}$$

For a particular (j) of $(l-1)^{th}$ layer.

$$\frac{\partial z_i^{l}}{\partial w_{ij}^{l}} = \frac{d\left[a_j^{l-1} \cdot w_{ij}^{l} + b_i^{l}\right]}{d\, w_{ij}^{l}}$$

$$= a_j^{l-1}$$

vectorization:

$$dw^{l}_{(n_l \times n_{l-1})} = \delta^{l}_{(n_l \times 1)} \cdot a^{l-1\,T}_{(1 \times n_{l-1})}$$

$$\frac{\partial C}{\partial b^l_j} = \frac{\partial C}{\partial z^l_j} \cdot \frac{\partial z^l_j}{\partial b^l_j} \quad (1)$$

{ For proof, look prev slide }

$$= \delta^l_j \cdot (1)$$

Vectorize :

$$\underset{(n_l \times 1)}{db^l} = \underset{(n_l \times 1)}{\delta^l}$$

All Vectorized eq at once :

$$\underset{(n_L \times 1)}{\delta^L} = (\underset{(n_L \times 1)}{A^L} - \underset{(n_L \times 1)}{y^L}) \odot \underset{(n_L \times 1)}{\sigma'(z^L)}$$

$$\underset{(n_l \times 1)}{\delta^l} = \left(\underset{(n_l \times n_{l+1})}{w^{l+1}}^T \cdot \underset{(n_{l+1} \times 1)}{\delta^{l+1}}\right) \odot \underset{(n_l \times 1)}{\sigma'(z^l)}$$

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

$$\underset{(n_l \times n_{l-1})}{\frac{\partial C}{\partial w^l}} = \underset{(n_l \times 1)}{\delta^l} \cdot \underset{(1 \times n_{l-1})}{a^{l-1}}^T$$

* Previous backprop eq are same for any activation function but not for any cost function. So, Let's create backprop eq for any cost function.

Let, $C \Rightarrow C(y^L, a^L)$

Ex1: $C(y^L, a^L) = \Sigma \frac{1}{2}(y^L - a^L)^2$

Ex2: $C(y^L, a^L) = -\frac{1}{m}\Sigma(y \log a + (1-y)\log(1-a))$

In this case everything remains same except $\frac{\partial C}{\partial z^L}$.

$$\delta^L = \frac{\partial C}{\partial z^L} = \frac{\partial C}{\partial a^L} \times \frac{\partial a^L}{\partial z^L}$$

$$\boxed{\delta^L = \frac{\partial C}{\partial a^L} \times \sigma'(z^L)}$$

Vectorized form:

$$\underset{(n_L \times 1)}{\delta L} = \underset{(n_L \times 1)}{da^L} \odot \underset{(n_L \times 1)}{\sigma'(z^L)}$$

Let's evaluate (δ^L) for log-likelihood cost func:

$$C(y^L, a^L) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} \log a^{(i)} + (1-y^{(i)}) \log(1-a^{(i)})\right)$$

$$\left.\frac{\partial C}{\partial a^L}\right|_{\substack{\text{for 1} \\ \text{Example}}} = -\left(\frac{y}{a} + \frac{(1-y)}{(1-a)} \times -1\right)$$

$$= \frac{1-y}{1-a} - \frac{y}{a} = \frac{a - ya - y + ya}{a(1-a)}$$

$$\boxed{\frac{\partial C}{\partial a^L} = \frac{a-y}{a(1-a)}}$$

If $\sigma(z)$ is sigmoid function, then

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

Finally,

$$\delta^L = \frac{\partial C}{\partial a^L} \times \sigma'(z^L)$$

$$= \frac{a-y}{a(1-a)} \times a(1-a) = a - y$$