Backpropagation

<u>Note</u>: $z_j^\ell = \left(\sum_k w_{jk}^\ell \, a_k^{\ell-1} \right) + b_j^\ell$

$$a_k^{\ell-1} = \sigma\left(z_k^{\ell-1} \right)$$

To compute ∇L, we must compute:

$$\boxed{\frac{\partial L}{\partial w_{jk}^\ell} = \frac{\partial L}{\partial z_j^\ell} \cdot \frac{\partial z_j^\ell}{\partial w_{jk}^\ell} = d_j^\ell \, a_j^{\ell-1}}$$

$$\boxed{\frac{\partial L}{\partial b_j^\ell} = \frac{\partial L}{\partial z_j^\ell} \cdot \frac{\partial z_j^\ell}{\partial b_j^\ell} = d_j^\ell}$$

$$d_j^\ell = \frac{\partial L}{\partial z_j^\ell}$$

sometimes called "error"

Final Layer ($\ell = F$)

Consider L as a fxn of the nodes in the final layer $a_1^F, \ldots, a_n^F$.

$$a_j^F = \sigma(z_j^F)$$

$$d_j^F = \frac{\partial L}{\partial z_j^F} = \sum_k \frac{\partial L}{\partial a_k^F} \frac{\partial a_k^F}{\partial z_j^F}$$

$$= \frac{\partial L}{\partial a_j^F} \frac{\partial a_j^F}{\partial z_j^F}$$

$$= \underbrace{\frac{\partial L}{\partial a_j^F}}_{\text{depends on choice of loss fxn}} \underbrace{\sigma'}_{\text{depends on choice of activation}}(z_j^F)$$

$$\vec{d}^F = \begin{bmatrix} d_1^F \\ d_2^F \\ \vdots \\ d_n^F \end{bmatrix}$$

n nodes in the final layer

$$\boxed{\vec{d}^F = \nabla_{a^F} L \odot \sigma'(\vec{z}^F)}$$

Intermediate Layers ($\ell = 2, 3, \ldots, F-1$)

$$d_j^\ell = \frac{\partial L}{\partial z_j^\ell}$$

Question: Where do z_j^ℓ's appear?

→ In a_j^ℓ's, which appear in $z_j^{\ell+1}$'s

sum over nodes in $\ell+1$ layer

$$d_j^\ell = \frac{\partial L}{\partial z_j^\ell} = \sum_k \frac{\partial L}{\partial z_k^{\ell+1}} \cdot \frac{\partial z_k^{\ell+1}}{\partial z_j^\ell}$$

$$= \sum_k d_k^{\ell+1} \cdot \frac{\partial z_k^{\ell+1}}{\partial z_j^\ell}$$

sum over nodes in ℓ^{th} layer

$$z_k^{\ell+1} = \sum_j w_{kj}^{\ell+1} a_j^\ell + b_k^{\ell+1}$$

$$= \sum_j w_{kj}^{\ell+1} \sigma(z_j^\ell) + b_k^{\ell+1}$$

$$\frac{\partial z_k^{\ell+1}}{\partial z_j^\ell} = w_{kj}^{\ell+1} \sigma'(z_j^\ell)$$

$$\Rightarrow d_j^{\ell} = \sum_k w_{kj}^{\ell+1} \, d_k^{\ell+1} \, \sigma'(z_j^{\ell})$$

$$\Rightarrow \boxed{\vec{d}^{\,\ell} = \left((W^{\ell+1})^T \vec{d}^{\,\ell+1} \right) \odot \sigma'(\vec{z}^{\,\ell})}$$

#8 from WS 5

$$L_T = \frac{1}{n}\sum_{i=1}^{n} L_i = \frac{1}{n}\sum_{i=1}^{n} L(\vec{y}_i, \tilde{y}_i)$$

$\vec{y}_i$ = one training sample

$\tilde{y}_i = \vec{a}_i^F$ is model prediction

n = # samples in training data

<u>Backprop</u>: Gives us an algorithm for computing

$$\frac{\partial L_i}{\partial w_{jk}^l} \quad \& \quad \frac{\partial L_i}{\partial b_j^l}$$

* For Grad. Desc., we need

$$\frac{\partial L_T}{\partial w_{jk}^l} \quad \& \quad \frac{\partial L_T}{\partial b_j^l}$$

$$L_T = \frac{1}{n}\sum_{i=1}^{n} L_i$$

$$\frac{\partial L_T}{\partial w_{jk}^{l}} = \sum_{i=1}^{n} \frac{\partial L_T}{\partial L_i} \cdot \underbrace{\frac{\partial L_i}{\partial w_{jk}^{l}}}$$

know from back prop

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial L_i}{\partial w_{jk}^{l}}$$

$$\frac{\partial L_T}{\partial b_j^{l}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial L_i}{\partial b_j^{l}}$$

$L = \frac{1}{2}\|\vec{y} - \tilde{y}\|^2$

$$\boxed{\nabla_{a^F} L = \vec{a}^F - \vec{y}}$$

#10 WS 5

#4 JN3

$$\nabla_{a^F} L = \begin{bmatrix} \frac{\partial L}{\partial a_1^F} \\ \frac{\partial L}{a_2^F} \\ \vdots \\ \frac{\partial L}{\partial a_K^F} \end{bmatrix} \overset{\star}{=} \begin{bmatrix} a_1^F - y_1 \\ a_2^F - y_2 \\ \vdots \\ a_K^F - y_K \end{bmatrix} = \vec{a}^F - \vec{y}$$

$$L(\vec{y}, \tilde{y}) = \frac{1}{2}\|\vec{y} - \tilde{y}\|^2$$

$$= \frac{1}{2}\|\vec{y} - \vec{a}^F\|^2$$

$$= \frac{1}{2}\left(\sqrt{(y_1 - a_1^F)^2 + (y_2 - a_2^F)^2 + \cdots + (y_K - a_K^F)^2}\right)^2$$

$$= \frac{1}{2}\left[(y_1 - a_1^F)^2 + (y_2 - a_2^F)^2 + \cdots + (y_K - a_K^F)^2\right]$$

$$\frac{\partial L}{\partial a_K^F} = \frac{1}{2}\left[(-2) \cdot (y_K - a_K^F)\right]$$

$$\overset{\star}{=} \boxed{a_K^F - y_K}$$