Proof of Backpropagation Algorithm

Sterling Jeppson

February 5, 2022

Theorem 1. $\delta^L = \nabla_a C \odot \sigma'(z^L)$

Proof. Suppose that the output layer has k nodes. Then

$$\delta^L = \begin{bmatrix} \dfrac{\partial C}{\partial z_1^L} \\ \vdots \\ \dfrac{\partial C}{\partial z_k^L} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial z_1^L}\left(\dfrac{1}{2}\displaystyle\sum_{j=1}^{k}(\sigma(z_j^L)-t_j)^2\right) \\ \vdots \\ \dfrac{\partial C}{\partial z_k^L}\left(\dfrac{1}{2}\displaystyle\sum_{j=1}^{k}(\sigma(z_j^L)-t_j)^2\right) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial z_1^L}\dfrac{1}{2}\left(\sigma(z_1^L)-t_1\right)^2 \\ \vdots \\ \dfrac{\partial C}{\partial z_k^L}\dfrac{1}{2}\left(\sigma(z_k^L)-t_k\right)^2 \end{bmatrix} = \begin{bmatrix} \left(\sigma(z_1^L)-t_1\right)\sigma'(z_1^L) \\ \vdots \\ \left(\sigma(z_k^L)-t_k\right)\sigma'(z_k^L) \end{bmatrix}$$

The third matrix follows from the second because in the jth row the derivative of the cost is being performed with respect to z_j^L which only occurs in one term in the sum. Thus the derivative of all other terms is 0. Since $\sigma(z_j^L) = a_j^L$ and $C = \frac{1}{2}\sum_{j=1}^{k}(a_j^L - t_j)^2$ and $\dfrac{\partial C}{\partial a_j^L} = (a_j^L - t_j)$, we can replace the 1st term in the jth row of δ^L with $\dfrac{\partial C}{\partial a_j^L}$ for $1 \le j \le k$. Finally we have

$$\delta^L = \begin{bmatrix} \dfrac{\partial C}{\partial a_1^L}\sigma'(z_1^L) \\ \vdots \\ \dfrac{\partial C}{\partial a_k^L}\sigma'(z_k^L) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial a_1^L} \\ \vdots \\ \dfrac{\partial C}{\partial a_k^L} \end{bmatrix} \odot \begin{bmatrix} \sigma'(z_1^L) \\ \vdots \\ \sigma'(z_k^L) \end{bmatrix}$$

It is shown that $\delta^L = \nabla_a C \odot \sigma'(z^L)$. □

Theorem 2. $\delta^l = ((w^{l+1})^{\mathrm{T}}\delta^{l+1}) \odot \sigma'(z^l)$

Proof. Suppose that the $l+1$ layer has q nodes and the l layer has k nodes. Then

$$\delta^l = \begin{bmatrix} \dfrac{\partial C}{\partial z_1^l} \\ \vdots \\ \dfrac{\partial C}{\partial z_k^l} \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{j=1}^{q}\dfrac{\partial C}{\partial z_j^{l+1}}\dfrac{\partial z_j^{l+1}}{\partial z_1^l} \\ \vdots \\ \displaystyle\sum_{j=1}^{q}\dfrac{\partial C}{\partial z_j^{l+1}}\dfrac{\partial z_j^{l+1}}{\partial z_k^l} \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{j=1}^{q}\delta_j^{l+1}\dfrac{\partial z_j^{l+1}}{\partial z_1^l} \\ \vdots \\ \displaystyle\sum_{j=1}^{q}\delta_j^{l+1}\dfrac{\partial z_j^{l+1}}{\partial z_k^l} \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{j=1}^{q}\dfrac{\partial z_j^{l+1}}{\partial z_1^l}\delta_j^{l+1} \\ \vdots \\ \displaystyle\sum_{j=1}^{q}\dfrac{\partial z_j^{l+1}}{\partial z_k^l}\delta_j^{l+1} \end{bmatrix}$$

In order to evaluate $\dfrac{\partial z_j^{l+1}}{\partial z_i^l}$ recall the definition of z_j^{l+1}. That is

$$\frac{\partial z_j^{l+1}}{\partial z_i^l}\left(z_j^{l+1}\right) = \frac{\partial z_j^{l+1}}{\partial z_i^l}\left(\sum_{p=1}^{k} w_{jp}^{l+1}a_p^l + b_j^{l+1}\right) = \frac{\partial z_j^{l+1}}{\partial z_i^l}\left(\sum_{p=1}^{k} w_{jp}^{l+1}\sigma(z_p^l) + b_j^{l+1}\right)$$

We are differentiating with respect to z_i^l and so the derivative of all terms in the sum are 0 except when $p = i$. Hence we only need to differentiate $w_{ji}^{l+1}\sigma(z_i^l)$. By the chain rule and the product rule of Calculus,

$$\frac{\partial}{\partial z_i^l}\left(w_{ji}^{l+1}\sigma(z_i^l)\right) = \frac{\partial}{\partial z_i^l}(w_{ji}^{l+1})\cdot\sigma(z_i^l) + \frac{\partial}{\partial z_i^l}(\sigma(z_i^l))\cdot w_{ji}^{l+1} = w_{ji}^{l+1}\sigma'(z_i^l)$$

Now we substitute back into δ^l.

$$\begin{bmatrix} \sum_{j=1}^{q} w_{j1}^{l+1}\sigma'(z_1^l)\delta_j^{l+1} \\ \vdots \\ \sum_{j=1}^{q} w_{jk}^{l+1}\sigma'(z_k^l)\delta_j^{l+1} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{q} w_{j1}^{l+1}\delta_j^{l+1} \\ \vdots \\ \sum_{j=1}^{q} w_{jk}^{l+1}\delta_j^{l+1} \end{bmatrix} \odot \begin{bmatrix} \sigma'(z_1^l) \\ \vdots \\ \sigma'(z_k^l) \end{bmatrix} = \left(\begin{bmatrix} w_{11}^{l+1} & w_{21}^{l+1} & \dots & w_{q1}^{l+1} \\ w_{12}^{l+1} & w_{22}^{l+1} & \dots & w_{q2}^{l+1} \\ \vdots & \vdots & & \vdots \\ w_{1k}^{l+1} & w_{2k}^{l+1} & \dots & w_{qk}^{l+1} \end{bmatrix} \begin{bmatrix} \delta_1^{l+1} \\ \vdots \\ \delta_q^{l+1} \end{bmatrix} \right) \odot \begin{bmatrix} \sigma'(z_1^l) \\ \vdots \\ \sigma'(z_k^l) \end{bmatrix}$$

It is shown that $\delta^l = ((w^{l+1})^{\mathrm{T}}\delta^{l+1}) \odot \sigma'(z^l)$. □

Theorem 3. $\dfrac{\partial C}{\partial b_j^l} = \delta_j^l$.

Proof. Suppose that the $l-1$ layer has k nodes. Then

$$\frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial b_j^l}$$

Recall that $z_j^l = \sum_{i=1}^{k} w_{ji}^l a_i^{l-1} + b_j^l$ and so $\dfrac{\partial z_j^l}{\partial b_j^l} = 1$. Hence $\dfrac{\partial C}{\partial b_j^l} = \dfrac{\partial C}{\partial z_j^l} = \delta_j^l$. □

Theorem 4. $\dfrac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$.

Proof. Suppose that the $l-1$ layer has k nodes. Then

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

Recall that $z_j^l = \sum_{i=1}^{k} w_{ji}^l a_i^{l-1} + b_j^l$ and so $\dfrac{\partial z_j^l}{\partial w_{jk}^l} = a_k^{l-1}$. Hence $\dfrac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$. □

Derivative of Sigmoid Function

$$\begin{aligned} \frac{\mathrm{d}\sigma(x)}{\mathrm{d}x} &= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{1+e^{-x}}\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}x}(1+e^{-x})^{-1} \\ &= -(1+e^{-x})^{-2} \cdot (-e^{-x}) \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x) \cdot \sigma(-x) \end{aligned}$$