Proof of Backpropagation Algorithm

Sterling Jeppson

February 5, 2022

Nomenclature

$\sigma(x)$	$\sigma(x) = 1/(1+e^{-x})$
w^l_{jk}	Weight connecting k neuron in layer $l-1$ to j neuron in layer l
b^l_j	Bias of j neuron in layer l
a^l_j	Activation of j neuron in layer l, $a^l_j = \sigma\left(\sum_{i=1}^k w^l_{ji} a^{l-1}_i + b^l_j\right)$
w^l	Weight matrix where jth row and kth column is w^l_{jk}
a^l	Activation matrix where jth row is a^l_j
b^l	Bias matrix where jth row is b^l_j
$f(\mathbf{A})$	Function applied to a matrix is the function applied to every element
z^l_j	Weighted input to the activation function for neuron j in layer l, $z^l_j = \sum_{i=1}^k w^l_{ji} a^{l-1}_i + b^l_j$
z^l	Weighted input to the neurons in layer l, $z^l = w^l a^{l-1} + b^l$
t_j	Target activation of neuron j in the output layer
C	Cost function for a single training input, $C = \frac{1}{2}\sum_{j=1}^k (a^L_j - t_j)^2$
$\mathbf{A} \odot \mathbf{B}$	Hadamard product is an elementwise product of two matrices with the same dimensions
L	Number of layers in the neural net
δ^l_j	Error of neuron j in layer l, $\delta^l_j = \partial C / \partial z^l_j$
δ^l	Error matrix where jth row is δ^l_j

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

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

$$\delta^L = \begin{bmatrix} \dfrac{\partial C}{\partial z^L_1} \\ \vdots \\ \dfrac{\partial C}{\partial z^L_k} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial z^L_1}\left(\dfrac{1}{2}\displaystyle\sum_{j=1}^k (\sigma(z^L_j) - t_j)^2\right) \\ \vdots \\ \dfrac{\partial C}{\partial z^L_k}\left(\dfrac{1}{2}\displaystyle\sum_{j=1}^k (\sigma(z^L_j) - t_j)^2\right) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial z^L_1}\dfrac{1}{2}\left(\sigma(z^L_1) - t_1\right)^2 \\ \vdots \\ \dfrac{\partial C}{\partial z^L_k}\dfrac{1}{2}\left(\sigma(z^L_k) - t_k\right)^2 \end{bmatrix} = \begin{bmatrix} \left(\sigma(z^L_1) - t_1\right)\sigma'(z^L_1) \\ \vdots \\ \left(\sigma(z^L_k) - t_k\right)\sigma'(z^L_k) \end{bmatrix}$$

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

$$\delta^L = \begin{bmatrix} \dfrac{\partial C}{\partial a^L_1}\sigma'(z^L_1) \\ \vdots \\ \dfrac{\partial C}{\partial a^L_k}\sigma'(z^L_k) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial C}{\partial a^L_1} \\ \vdots \\ \dfrac{\partial C}{\partial a^L_k} \end{bmatrix} \odot \begin{bmatrix} \sigma'(z^L_1) \\ \vdots \\ \sigma'(z^L_k) \end{bmatrix} = \nabla_a C \odot \sigma'(z^L) \qquad \square$$

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

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

In order to evaluate $\frac{\partial z_j^{l+1}}{\partial z_i^l}$ 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.

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

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

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

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 $\frac{\partial z_j^l}{\partial b_j^l} = 1$. Hence $\frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} = \delta_j^l$. □

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

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 $\frac{\partial z_j^l}{\partial w_{jk}^l} = a_k^{l-1}$. Hence $\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$. □

Derivative of Sigmoid Function

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