Backwards Propagation

Benjamin S. Knight

March 5, 2017

The Derivative of the Sigmoid Function

$$\begin{aligned}
\frac{\delta}{\delta x}\sigma(x) &= \frac{\delta}{\delta x}\Big[\frac{1}{1+e^{-x}}\Big] \\
&= \frac{\delta}{\delta x}(1+e^{-x})^{-1} \\
&= (-1)(1+e^{-x})^{-2}\frac{\delta}{\delta x}(1+e^{-x}) \\
&= (-1)(1+e^{-x})^{-2}(e^{-x}) \\
&= -(1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} * \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} * \frac{(1+e^{-x})-1}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} * \Big(1 - \frac{1}{1+e^{-x}}\Big) \\
&= \sigma(x) * (1-\sigma(x))
\end{aligned} \tag{1}$$

Back Propagation from the Output to the Output Layer via the Sigmoid Function

$$\begin{aligned}
y &= 1 \\
\hat{y} &= 0.21 \\
\therefore E &= 0.79
\end{aligned} \tag{2}$$

$$\begin{aligned}
\delta^j &= (y-\hat{y}) * \sigma(x) * (1-\sigma(x)) \text{ where } x = \text{-1.28 and } \hat{y} = 0.21 \\
&= (1-0.21)\Big(\frac{1}{1+e^{-(-1.28)}}\Big)\Big(1-\frac{1}{1+e^{-(-1.28)}}\Big) \\
&= (1-0.21)(0.21)(1-0.21) \\
&= (0.79)(0.21)(0.79) \\
&= 0.131
\end{aligned} \tag{3}$$

A Note on Partial Derivatives

With partial derivatives, we are interested in specific change with respect to some other variable of interest. Take the following expression.

$$f(x,y) = x^2 + xy + y^2$$

If we take the partial derivative with respect to x, then we can drop all isolated instances of y. Thus,

$$\begin{aligned}\frac{\partial f}{\partial x}(x,y) &= \frac{\partial f}{\partial x}(x^2+xy+y^2) \\ &= 2x^1+(1)x^0y+0 \\ &= 2x+y\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\partial f}{\partial y}(x,y) &= \frac{\partial f}{\partial y}(x^2+xy+y^2) \\ &= 0+x(1)y^0+2y^1 \\ &= x+2y\end{aligned} \tag{5}$$

$$\begin{aligned}f(h1) = x_1(W^{h1}_{x1})+x_2(W^{h1}_{x2}) \quad &\rightarrow \quad \frac{\partial h_1}{\partial x_1} = W^{h1}_{x1} \quad \text{and} \quad \frac{\partial h_1}{\partial x_2} = W^{h1}_{x2} \\ f(h2) = x_1(W^{h2}_{x1})+x_2(W^{h2}_{x2}) \quad &\rightarrow \quad \frac{\partial h_2}{\partial x_1} = W^{h2}_{x1} \quad \text{and} \quad \frac{\partial h_2}{\partial x_2} = W^{h2}_{x2}\end{aligned}$$

$$\delta^h 1 = W^h j \delta^0 f'(h) \tag{6}$$

The Output Layer

Assume that during forward propagation j is an input to h. During back propagation, this order of inputs reverses. Recall that the derivative of the sum is the sum of the derivatives. Bearing this in mind, we see that the partial derivative of the error term E with respect to node j - the function which outputs the value for j that minimizes E - is as follows:

$$\frac{\delta E}{\delta g_j} = \sum_i \sigma'(h_i) v_{ij} \frac{\delta E}{\delta h_i} \tag{7}$$

Back Propagation

Let there by two nodes - one hidden node i and one output node j. Let y_j represent the final output of the neural net. y_i will represent the output of the hidden node i. Let z_j represent the total input received by the output unit, j.

We want an expression that shows the extent to which the error term E changes as a function of total input received by the output unit. For this we use the partial derivative of the error term with respect to z_j.

$$\frac{\partial E}{\partial z_j}$$

Because the loss function that derives E relies on inputs from another function - the j node, we must invoke the chain rule. If function $F(x)$ is itself a function of $g(x)$, then the derivative of $F(x)$ is equal to the derivative of the parent function times the child function.

$$\text{If } F(x) = f(g(x))$$

$$\text{then } F'(x) = f'(g(x))g'(x)$$

...or in our case:

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j}$$

The Sum of Squared Errors

The derivative of the error term with respect to the input node j is the derivative of the sum of squares. In this instance, there is only a single output, so we then drop the summation operator. Given that we are interested in minimizing the loss function, we change the sign on the resulting function to effectively maximize negative error.

$$\begin{aligned} \frac{\delta E}{\delta y_j} &= \frac{\delta}{\delta y_j}\left(1/2\sum(y^\mu - \hat{y}^\mu)^2\right)\\ &= 2(1/2)\sum(y^\mu - \hat{y}^\mu)^1\\ &= \sum(y^\mu - \hat{y}^\mu) \qquad \rightarrow \qquad y^\mu - \hat{y}^\mu \end{aligned} \tag{8}$$

$$\downarrow$$

$$\frac{\delta E}{\delta y_j} = -(y^\mu - \hat{y}^\mu)$$