ANN Backpropagation: Weight updates for hidden nodes

Kiri Wagstaff

February 1, 2008

First, recall how a multilayer perceptron (or artificial neural network, ANN) predicts a value for a new input, x. Assume that there is a single hidden layer. Each hidden node z_h in this layer produces an intermediate output based on a weighted sum of the inputs:

$$z_h = \text{sigmoid}(w_h^T x)$$

where w^T indicates taking the transpose of vector w. We'll get to the sigmoid function in a minute.

Next, the final output $\hat{y}$ of the ANN is a weighted sum of the hidden node outputs (including v_0, which is the weight associated with a node that always has the value +1, just like w_0).

$$\hat{y} = \sum_{h=1}^{H} v_h z_h + v_0 \tag{1}$$

where H is the number of nodes in the hidden layer.

The error associated with this prediction is

$$E(W, v|x) = \frac{1}{2}(y - \hat{y})^2$$

where W and v are the learned weights (W is an H-by-d matrix because there are $d + 1$ weights (one for each input feature plus w_0) for each of the H hidden nodes, and v is a vector of H weights, one for each hidden node). The output of the network, $\hat{y}$, is an approximation to the true (desired) output, y. If you're wondering why there is a factor of $\frac{1}{2}$ in there, you'll see the reason shortly.

To do backpropagation, we need to: 1) update the weights v and 2) update the weights W. Here I will show how to derive the updates if there is only a single training example, x, with an associated label y. This result can be easily extended to handle a data set X containing n items, each with their own output y_i (and this is what you see in the book, p. 246).

Step 1: Update the weights v

To figure out how to update each the weights in v, we compute the partial derivative of E with respect to v_h (for the weight that connects the output $\hat{y}$ to hidden node h) and multiply it by the learning factor η. Actually, we use $-\eta$ to indicate that we want to reverse the error that $\hat{y}$ made:

$$\begin{aligned}
\Delta v_h &= -\eta \frac{\delta E}{\delta v_h} \\
&= -\eta \frac{\delta E}{\delta \hat{y}} \frac{\delta \hat{y}}{\delta v_h} \\
&= -\eta \frac{\delta \frac{1}{2}(y - \hat{y})^2}{\delta \hat{y}} \frac{\delta \hat{y}}{\delta v_h} \\
&= -\eta (y - \hat{y})(-1) \frac{\delta \hat{y}}{\delta v_h} \\
&= \eta (y - \hat{y}) \frac{\delta \hat{y}}{\delta v_h}
\end{aligned}$$

$$= \eta(y - \hat{y})z_h$$

The $\frac{1}{2}$ factor is canceled when we take the partial derivative, and we include a multiplicative factor of -1 for the derivative with respect to $\hat{y}$ "inside" $(y - \hat{y})$. From Equation 1, we know that $\frac{\delta \hat{y}}{\delta v_h}$ is z_h.

Step 2: Update the weights W

To figure out how to update each of the weights in W, we compute the partial derivative of E with respect to w_{hj} (for the weight that connects hidden node h with input j) and multiply it by the learning factor η. Let's break down the partial derivative:

$$\begin{aligned}
\Delta w_{hj} &= -\eta \frac{\delta E}{\delta w_{hj}} \\
&= -\eta \frac{\delta E}{\delta \hat{y}} \frac{\delta \hat{y}}{\delta z_h} \frac{\delta z_h}{\delta w_{hj}}
\end{aligned}$$

We know that $\frac{\delta E}{\delta \hat{y}}$ is $-(y - \hat{y})$ from the previous step. The partial derivative $\frac{\delta \hat{y}}{\delta z_h}$ is also simple; from Equation 1 we see it is just v_h.

The interesting part is finding $\frac{\delta z_h}{\delta w_{hj}}$. This can be re-written as $\frac{\delta z_h}{\delta w_h^T x} \frac{\delta w_h^T x}{\delta w_{hj}}$. The first part, $\frac{\delta z_h}{\delta w_h^T x}$, is where the sigmoid comes in. The sigmoid equation, in general, is

$$\text{sigmoid}(a) = \frac{1}{1 + e^{-a}}.$$

I am using a here to represent any argument given to the sigmoid function. For z_h, $a = w_h^T x$. Now the partial derivative of the sigmoid with respect to its argument, a, is:

$$\begin{aligned}
\frac{\delta \text{sigmoid}(a)}{\delta a} &= \frac{\delta \frac{1}{1+e^{-a}}}{\delta a} \\
&= -\frac{1}{(1 + e^{-a})^2} \, e^{-a} \, (-1) \\
&= \frac{1}{(1 + e^{-a})^2} \, e^{-a} \\
&= \frac{1}{1 + e^{-a})} \, \frac{e^{-a}}{1 + e^{-a}} \\
&= \frac{1}{1 + e^{-a})} \, \frac{(1 - 1) + e^{-a}}{1 + e^{-a}} \\
&= \frac{1}{1 + e^{-a})} \, \frac{(1 + e^{-a}) - 1}{1 + e^{-a}} \\
&= \frac{1}{1 + e^{-a})} \left(1 - \frac{1}{1 + e^{-a}}\right) \\
&= \text{sigmoid(a)}(1 - \text{sigmoid(a)})
\end{aligned}$$

So for z_h, we have $\frac{\delta z_h}{\delta w_h^T x} = z_h(1 - z_h)$. Isn't that neat?

But don't forget the second half, $\frac{\delta w_h^T x}{\delta w_{hj}}$. Luckily, this is straightforward: it is just x_j.

Thus, overall we have

$$\begin{aligned}
\Delta w_{hj} &= -\eta \frac{\delta E}{\delta w_{hj}} \\
&= -\eta \frac{\delta E}{\delta \hat{y}} \frac{\delta \hat{y}}{\delta z_h} \frac{\delta z_h}{\delta w_{hj}} \\
&= -\eta(-(y-\hat{y}))\, v_h\, (z_h(1-z_h)x_j) \\
&= \eta(y-\hat{y})\, v_h\, z_h(1-z_h)x_j
\end{aligned}$$

That's it! Let me know if you have any questions.