Stats 315B Homework

Rachael Caelie (Rocky) Aikens, Daniel Sosa, Christine Tataru

Problem 1

First, some notation. Let $K^{(l)}$ denote the number of nodes in the l^{th} layer of a neural net, and let $l = 0...L$ index the layers of the neural net, with $l = 0$ denoting the input layer and $l = L$ denoting the output layer. Additionally, we define:

$$\delta_j^{(l)} = \frac{\partial q}{\partial a_j^{(l)}},$$

Where $a_j^{(l)} = \sum_{j=0}^{K^{(l)}} w_{ij} o_i^{(l-1)}$.

Here, I am assuming that we are applying stochastic gradient descent and I have dropped the subscript (t) denoting the observation we are using, for the purposes of this derivation. I also assume that $o_0^{(l)} = 1$ for each layer, denoting the intercept inputs to each layer

To define the update rule, we need to calculate $G(w_{ij}^{(l)}) = \frac{\partial q}{\partial w_{ij}^{(l)}}$. Applying the chain rule, this is:

$$\begin{aligned}
\frac{\partial q}{\partial w_{ij}^{(l)}} &= \frac{\partial q}{\partial a_j^{(l)}} \frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} \\
&= \delta_j^{(l)} o_i^{(l-1)}
\end{aligned}$$

Moreover, we already know that $\delta^{(L)} = y - o^{(L)}$, the error for this example from forward propagation. Applying the chain rule, we can recursively calculate $\delta_j^{(l)}$ backwards through the network as:

$$\delta_j^{(l)} = S'(a_j^{(l)}) \sum_{k=1}^{K^{(l+1)}} \delta_k^{(l+1)} w_{jk}.$$

For the sigmoid activation function, this is:

$$\delta_j^{(l)} = o_j^{(l)}(1 - o_j^{(l)}) \sum_{k=1}^{K^{(l+1)}} \delta_k^{(l+1)} w_{jk}.$$

Now we can write an algorithm for backpropagation (let e denote the error from forward propagation, O the outputs from each node in forward propagation, x the example we are using, and W the weights for the whole network).

Backprop(e, O, W, x):

$\delta^{(l)} = e$

$g(w_i^{(l)}) = \delta^{(l)} o_i^{(l)}$

For $l = L - 1$ to 1:

$\{\delta_j^{(l)} = S'(a_j^{(l)}) \sum_{k=1}^{K^{(l+1)}} \delta_k^{(l+1)} w_{jk}$ for each $j = 1...K^{(l)}\}$

$\{g(w_{ij}^{(l)}) = \delta_j^{(l)} o_i^{(l-1)}$ for each $j = 1...K^{(l)}, i = 0...K^{(l-1)}\}$

$\{w_{ij}^{(l)} \leftarrow w_{ij} - \eta g(w_{ij}^{(l)})$ for each $j = 1...K^{(l)}, i = 0...K^{(l-1)}, l = 1...L\}$