Backpropagation Implementation Helper Formulas

Notations

- b^l_j for the bias of the j^{th} neuron in the l^{th} layer
- w^l_{jk} for the weight for the connection from the k^{th} neuron in the $(l-1)^{th}$ layer to the j^{th} neuron in the l^{th} layer
- z^l_j for the weighted input to the activation function for neuron j in layer l
- a^l_j for the activation of the j^{th} neuron in the l^{th} layer
- σ for the sigmoid function

Formulas

1. the activation a^l_j of the j^{th} neuron in the l^{th} layer is related to the activations in the $(l-1)^{th}$ layer

$$a^l_j = \sigma\left(\sum_k w^l_{jk} a^{l-1}_k + b^l_j\right)$$

   which is equivalent to

$$a^l = \sigma(w^l a^{l-1} + b^l)$$

2. the cost function

$$C = \frac{1}{2}\|y - a^L\|^2 = \frac{1}{2}\sum_j (y_j - a^L_j)^2$$

3. the gradient δ^l_j of neuron j in layer l

$$\delta^l_j = \frac{\partial C}{\partial z^l_j}$$

4. the gradient in the output layer, δ^L

   $

   \begin{eqnarray}
   \delta^L_j = \frac{\partial C}{\partial z^L_j}= \frac{\partial C}{\partial a^L_j} \sigma'(z^L_j) = (a_j^L-y_j)\sigma'(z^L_j)
   \end{eqnarray}

   $

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L) = (a_j^L - y_j)\sigma'(z_j^L)$$

which is equivalent to

$$\delta^L = (a^L - y) \odot \sigma'(z^L)$$

5. the gradient δ^l in terms of the gradient in the next layer, δ^{l+1}, where $l < L$

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

Proof

$$\begin{aligned}
\delta_j^l &= \frac{\partial C}{\partial z_j^l} \\
&= \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} \\
&= \sum_k \frac{\partial z_k^{l+1}}{\partial z_j^l} \delta_k^{l+1}
\end{aligned}$$

where

$$z_k^{l+1} = \sum_j w_{kj}^{l+1} a_j^l + b_k^{l+1} = \sum_j w_{kj}^{l+1} \sigma(z_j^l) + b_k^{l+1}$$

so

$$\frac{\partial z_k^{l+1}}{\partial z_j^l} = w_{kj}^{l+1} \sigma'(z_j^l)$$

then we have

$$\delta_j^l = \sum_k w_{kj}^{l+1} \delta_k^{l+1} \sigma'(z_j^l)$$

$\square$

6. the gradient of the bias b^l in layer l

$$\frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \frac{\partial z_j^l}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \frac{\partial (w^l a^{l-1} + b_j^l)}{\partial b_j^l} = \delta_j^l$$

which is equivalent to

$$\frac{\partial C}{\partial b^l} = \delta^l$$

7. the gradient of the weight w^l in layer l

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

which is equivalent to

$$\frac{\partial C}{\partial w^l} = a^{l-1} \delta^l$$