Back-propagation derivation

——518021910971 裴奕博

With back propagation:

The gradient is

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = \delta_j z_i$$

According to backprop formula,

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$$

So

$$\frac{\partial E}{\partial w_{ji}} = h'(a_j) z_i \sum_k w_{kj}\delta_k$$

Without back propagation：

The gradient can be expressed with:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = h'(a) z_i \frac{\partial E}{\partial z_j}$$

And

$$\frac{\partial E}{\partial z_j} = \sum_k \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial z_j} = \sum_k \frac{\partial E}{\partial y_k} w_{kj}$$

So

$$\frac{\partial E}{\partial w_{ji}} = h'(a) z_i \frac{\partial E}{\partial z_j} = h'(a) z_i \sum_k \frac{\partial E}{\partial y_k} w_{kj} = h'(a_j) z_i \sum_k w_{kj}\delta_k$$

Two results are the same.

Example

Given a 3 layers of feed-forward neural network where the input layer has 3 elements, the hidden layer has 4 elements and the output layer has 2 elements. The activation function for the hidden layer is sigmoid function and the activation function for output is identity. The loss is L2 loss.

At this time,

$$
\begin{aligned}
\frac{\partial E}{\partial z_j} &= \frac{\partial[(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2]}{\partial z_j} \\
&= 2(y_1-\hat{y}_1)\frac{\partial y_1}{\partial z_j} + 2(y_2-\hat{y}_2)\frac{\partial y_2}{\partial z_j} \\
&= 2(y_1-\hat{y}_1)w_{1j} + 2(y_2-\hat{y}_2)w_{2j} \\
&= \sum_{k=1}^{2} w_{kj}\delta_k
\end{aligned}
$$

The result has the same form with that used backprop method, and can easily write out $\partial E/\partial w_{ji}$ by adding $\frac{\partial z_j}{\partial a_j} = h'(a)$ and $\frac{\partial a_j}{\partial w_{ji}} = z_i$.