Note on backprop with softmax

1 Intro

I will give some of the important steps of the backpropagation process using a softmax activation. Note that it might be useful if you try to see how precicely the results I give here are obtained.

2 The steps

Using the chain rule, the gradient of the loss function with respect to the weights can be written as

$$\frac{\partial L}{\partial w_{ij}} = \sum_k \frac{\partial L}{\partial z_k} \frac{\partial z_k}{\partial w_{ij}} \tag{1}$$

The elements in this equation are defined as follows:

$$y_k = \frac{e^{z_k}}{S}, S = \sum_i e^{z_i} \tag{2}$$

$$z_k = \sum_i w_{ik} x_i + b_k \tag{3}$$

$$L = -\sum_k t_k \log y_k \quad (t_k \text{ are the targets}). \tag{4}$$

We first calculate the derivative of the loss with respect to an arbitrary component of z. Using that the derivative of the logarithm of the softmax with respect to an arbitrary component of z is

$$\frac{\partial \log y_k}{\partial z_i} = \delta_{ki} - \frac{1}{S} \frac{\partial S}{\partial z_i}, \tag{5}$$

where δ_{ki} is the Kronecker delta, we find

$$\frac{\partial L}{\partial z_i} = y_i - t_i. \tag{6}$$

We assumed here that the targets are one-hot encoded. So, since the derivative of some z-component with respect to the weights is just

$$\frac{\partial z_k}{\partial w_{ij}} = \delta_{kj} x_i \tag{7}$$

the gradient of the loss function becomes

$$\frac{\partial L}{\partial w_{ij}} = x_i (y_j - t_j). \tag{8}$$