Derivative of the Softmax Function

John Purcell

May 20, 2021

Abstract

1 Terms

a_i the activation (output) of the ith neuron of a layer of neurons

z_i the weight of the ith neuron in the layer

$ln()$ the natural logarithm function, $log_e()$

We will also use the symbol σ to indicate the softmax function.

2 Softmax

We apply the softmax function to the weighted sum of a neuron to find its activation (output). It's defined like this:

$$a_i = \sigma(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{1}$$

Here, j ranges over all possible values for the layer, iterating over all neurons in the layer.

We take the exponential function of the weighted sum of a neuron and then divide by the sum of the exponentials of all the weighted sums in the layer, so that the outputs of all the neurons in the layer sum to 1.

3 The Chain Rule and Logs of Functions

The *chain rule* enables us to differentiate functions of functions. We will use it here to differentiate $ln(f(x))$, where $f()$ is any differentiable function.

Let

$$y = ln(f(x))$$

and

$$u = f(x)$$

so that

$$y = ln(u)$$

then

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \frac{1}{u}\frac{d}{dx}f(x) = \frac{1}{f(x)}\frac{d}{dx}f(x)$$

Therefore

$$\boxed{\frac{d}{dx}ln(f(x)) = \frac{1}{f(x)}\frac{d}{dx}f(x)} \tag{2}$$

This is known as the *logarithmic derivative* of the function f.

4 Derivative of Softmmax

We will now use (2) to find the derivative of the softmax function. We can simplify by first taking the log of the activation and introducing a variable s.

$$s = ln(a_i) \tag{3}$$

$$\frac{\partial s}{\partial z_j} == \frac{1}{a_i}\frac{\partial a_i}{\partial z_j}$$

Rearranging gives:

$$a_i \frac{\partial s}{\partial z_j} = \frac{\partial a_i}{\partial z_j} \tag{4}$$

The term on the right is what we are looking for; the derivative of the softmax function with respect to an arbitrary weighted sum in the same layer, z_j. We simply need to find $\frac{\partial s}{\partial z_j}$.

First we will actually take the log of the softmax function, keeping in mind equations (1) and (3).

$$ln\left(\frac{e^{z_i}}{\sum_n e^{z_n}}\right) = ln(e^{z_i}) - ln(\sum_n e^{z_n}) = z_i - ln(\sum_n e^{z_n})$$

Here we have used the following property of logarithms.

$$log\left(\frac{x}{y}\right) = log(x) - log(y)$$

Now we can differentiate this expression.

$$\frac{\partial s}{\partial z_j} = \frac{\partial z_i}{\partial z_j} - \frac{\partial}{\partial z_j} ln(\sum_n e^{z_n})$$

The term $\frac{\partial z_i}{\partial z_j}$ evaluates to 1 if $i = j$, otherwise it's zero.

Here we can introduce the *Kronecker delta*, δ_{ij}. The Kronecker delta is defined like this:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

So

$$\frac{\partial s}{\partial z_j} = \delta_{ij} - \frac{\partial}{\partial z_j} ln(\sum_n e^{z_n}) = \delta_{ij} - \frac{1}{\sum_n e^{z_n}} \frac{\partial}{\partial z_j} (\sum_n e^{z_n})$$

The derivative of the sum in the above equation expands to a series of terms, all of which are zero except for the one containing e^{z_j}, and

$$\frac{\partial}{\partial z_j}(e^{z_j}) = e^{z_j}$$

So

$$\frac{\partial s}{\partial z_j} = \delta_{ij} - \frac{e^{z_j}}{\sum_n e^{z_n}} = \delta_{ij} - \sigma(z_j)$$

Plugging this into (4)

$$\frac{\partial}{\partial z_j}\sigma(z_i) = \sigma(z_i)(\delta_{ij} - \sigma(z_j)) \tag{5}$$