SOFTMAX FUNCTION DERIVATIVE

Eren Gultepe

The following is a detailed explanation of the derivative for the softmax function, which is used in the ipynb **Classification Tutorial** notebook under the Gradient Descent for Multiclass Logisitc Regression subsection.

The goal is to show how the following derivative, of the softmax function, is achieved:

$$\frac{\partial y_l}{\partial z_k} = y_k(I_{k,l} - y_l)$$

We first start with the the softmax function as:

$$s_i = \frac{e^{z_i}}{\sum_{l=1}^{n} e^{z_l}}, \qquad \forall i = 1, \dots, n$$

Where each output of the softmax function depends on all the input values (due to the denominator)

Since outputs of the softmax function are strictly positive values, we can make the following derivation super short, by applying the following trick: instead of taking the partial derivative of the output, we take the partial derivative of the log of the output (also called “logarithmic derivative”):

$$\frac{\partial}{\partial z_j} log(s_i) = \frac{1}{s_i} \cdot \frac{\partial s_i}{\partial z_j}$$

we rearrange the upper formula and obtain:

$$\frac{\partial s_i}{\partial z_j} = s_i \cdot \frac{\partial}{\partial z_j} log(s_i)$$

The left-hand side is exactly the partial derivative we’re looking for.

As we will shortly see, the right-hand side simplifies the computation of the derivative, such that we don’t require the quotient rule of derivatives. We must first take the logarithm of s:

$$log\, s_i = log\left(\frac{e^{z_i}}{\sum_{l=1}^{n} e^{z_l}}\right) = z_i - log\left(\sum_{l=1}^{n} e^{z_l}\right)$$

The partial derivative of the resulting expression is:

$$\frac{\partial}{\partial z_j} log\, s_i = \frac{\partial z_i}{\partial z_j} - \frac{\partial}{\partial z_j} log \left(\sum_{l=1}^{n} e^{z_l} \right)$$

Let's have a look at the first term on the right-hand side:

$$\frac{\partial z_i}{\partial z_j} = \begin{cases} 1, & if\ i = j \\ 0, & otherwise \end{cases}$$

which can be concisely written using the indicator function $1\{\cdot\}$ or as I. The indicator function takes on a value of 1 if its argument is true, and 0 otherwise.

The second term on the right-hand side can be evaluated by applying the chain rule:

$$\frac{\partial}{\partial z_j} log\, s_i = 1\{i = j\} - \frac{1}{\sum_{l=1}^{n} e^{z_l}} \cdot \left(\frac{\partial}{\partial z_j} \sum_{l=1}^{n} e^{z_l} \right)$$

In the step above we used the derivative of the natural logarithm:

$$\frac{d}{dx} log(x) = \frac{1}{x}$$

Obtaining the partial derivative of the sum is trivial:

$$\frac{\partial}{\partial z_j} \sum_{l=1}^{n} e^{z_l} = \frac{\partial}{\partial z_j} \left[e^{z_1} + e^{z_2} + \cdots + e^{z_j} + \cdots + e^{z_n} \right] = \frac{\partial}{\partial z_j} \left[e^{z_j} \right] = e^{z_j}$$

Plugging the result into the formula yields:

$$\frac{\partial}{\partial z_j} log\, s_i = 1\{i = j\} - \frac{e^{z_j}}{\sum_{l=1}^{n} e^{z_l}} = 1\{i = j\} - s_j$$

Finally, we have to multiply the upper expression with s, as shown at the beginning of this section:

$$\frac{\partial s_i}{\partial z_j} = s_i \cdot \frac{\partial}{\partial z_j} log(s_i) = s_i \cdot \big(1\{i = j\} - s_j\big)$$

This concludes our derivation, which the same thing as:

$$\frac{\partial y_l}{\partial z_k} = y_k(I_{k,l} - y_l)$$

Adapted from Thomas Kurbiel.