Softmax function

Kristian Wichmann

May 2, 2017

1 Definition

Given an n-dimensional input vector z, the *softmax function* (also known as the *normalized exponential function*), has the output:

$$\sigma_i(z) = \frac{e^{z_i}}{\sum_{k=1}^{n} e^{z_k}} \tag{1.1}$$

This means, that the outputs can be interpreted as an discrete probability distribution, since they will always sum to 1.

It will be convenient to give a shorthand for the normalization "constant", so we set:

$$N(z) = \sum_{k=1}^{n} e^{z_k} \tag{1.2}$$

1.1 Example

Figure 1 shows the softmax function applied to the set $\{1, 2, 3, \cdots, 8\}$. As is evident, comparatively small values are given much less overall weight than higher ones.

2 Derivative

We might now want to differentiate with respect to the component z_j, which is done by applying the quotient rule:

$$\frac{\partial \sigma_i(z)}{\partial z_j} = \frac{\partial}{\partial z_j} \frac{e^{z_i}}{N(z)} = \frac{\left(\frac{\partial}{\partial z_j} e^{z_i}\right) N(z) - e^{z_i} \left(\frac{\partial}{\partial z_j} N(z)\right)}{(N(z))^2} \tag{2.1}$$

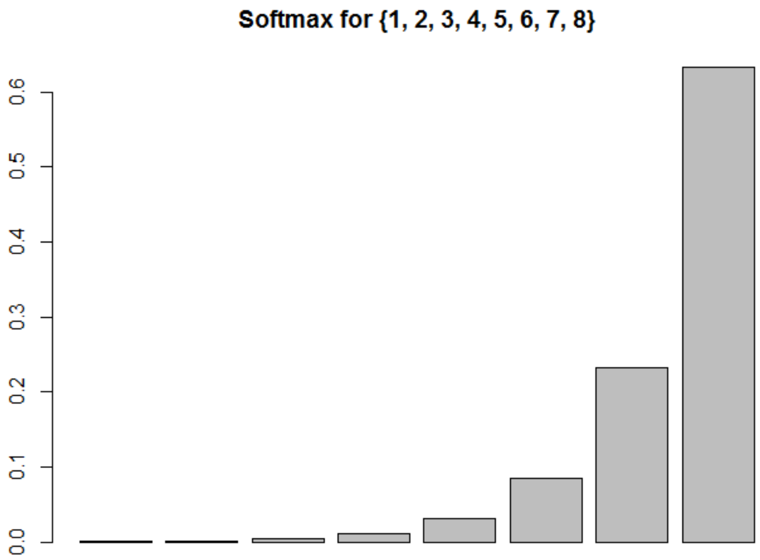

Figure 1: Softmax example

The two derivatives are:

$$\frac{\partial}{\partial z_j} e^{z_i} = \delta_{ij} e^{z_i}, \quad \frac{\partial}{\partial z_j} N(z) = \sum_{k=1}^{n} \delta_{jk} e^{z_k} = e^{z_j} \tag{2.2}$$

Inserting into equation 2.1 this yields:

$$\frac{\delta_{ij} e^{z_j} N(z) - e^{z_i + z_j}}{(N(z))^2} \tag{2.3}$$

The numerator can be rewritten:

$$e^{z_i} \left(\delta_{ij} N(z) - e^{z_j} \right) \tag{2.4}$$

Now divide by $N(z)$ twice, once "outside the parenthesis" and once "inside" to get:

$$\frac{\partial \sigma_i(z)}{\partial z_j} = \frac{e^{z_i}}{N(z)} \left(\delta_{ij} - \frac{e^{z_j}}{N(z)} \right) = \sigma_i(z) \left(\delta_{ij} - \sigma_j(z) \right) \tag{2.5}$$

The likeness to the derivative of the logistic function should be evident.

3 Cross-entropy error function

The softmax is often combined with a cross-entropy error function for classification:

$$J(z) = -\sum_{i=1}^{n} t_i \log \sigma_i(z) = -\sum_{i=1}^{n} t_i \log y_i \tag{3.1}$$

Here t_i represents the label for the data and $y_i = \sigma_i(z)$ is the softmax output. For classification, this is simply $t_i = \delta_{ic}$, where c is the correct label. In this case, the error function is simply:

$$J(z) = -\sum_{i=1}^{n} \delta_{ic} \log \sigma_i(z) = -\log \sigma_c(z) = -y_c \tag{3.2}$$

The derivative with respect to z_i is found through the chain rule:

$$\frac{\partial J}{\partial z_i} = \frac{\partial J}{\partial y_c} \frac{\partial y_c}{\partial z_i} = -\frac{\partial}{\partial z_i} \log \sigma_c(z) \frac{\partial \sigma_c}{\partial z_i} = -\frac{1}{\sigma_c(z)} \sigma_c(z) \left(\delta_{ci} - \sigma_i(z)\right) = y_i - \delta_{ic} \tag{3.3}$$

But this is exactly the difference between the real value t_i and the output y_i, also known as the *error* $\delta_i = y_i - \delta_{ic}$:

$$\frac{\partial J}{\partial z_i} = \delta_i \tag{3.4}$$