Gradient of the Softmax

Victor BUSA victor.busa@gmail.com

April 12, 2017

I already created an explanation on how to compute the gradient of the svm hinge loss in a previous paper. I will detail how to compute the gradient of the softmax function here. This paper will help us practice math and also show us how to use the chain rule.

The Problem Before we delve into the calculation of the gradient, we will set the problem. In this case we want to compute:

$$\frac{\partial L_i(f(w_k))}{\partial w_k}$$

where:

$$L_i = -log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

and:

$$f_j = w_j x_i$$

Chain Rule Before we compute the gradient of this function let's recall the chain rule. The chain rule states that:

$$\frac{\partial(g(f(x)))}{dx} = \frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x}$$
$$where\ u = f(x)$$

in this case we will use the chain rule because what we want to compute is:

$$\frac{\partial L_i}{\partial w_j}$$

but we will compute:

$$\frac{\partial L_i(f(w_k))}{\partial w_k} = \frac{\partial L_i(f_k)}{\partial f_k}\frac{\partial f_k}{\partial w_k} \tag{1}$$
$$where\ f_k = f(w_k) = w_k x_i$$

Analytic gradient We will firstly compute the quantity $\frac{\partial L_i(f_k)}{\partial f_k}$:

$$\begin{aligned}
\frac{\partial L_i(f_k)}{\partial f_k} &= \frac{\partial}{\partial f_k}\left(-log\left(\frac{e^{f_{y_i}}}{\sum\limits_j e^{f_j}}\right)\right) \\
&= -\left[\frac{\frac{\partial}{\partial f_k}\left(\frac{e^{f_{y_i}}}{\sum\limits_j e^{f_j}}\right)}{\frac{e^{f_{y_i}}}{\sum\limits_j e^{f_j}}}\right] \\
&= -\left[\frac{e^{f_{y_i}}\frac{\partial}{\partial f_k}\left(\frac{1}{\sum\limits_j e^{f_j}}\right) + \frac{\partial}{\partial f_k}\left(e^{f_{y_i}}\right)\frac{1}{\sum\limits_j e^{f_j}}}{\left(\frac{e^{f_{y_i}}}{\sum\limits_j e^{f_j}}\right)}\right] \\
&= -\left[\frac{-e^{f_{y_i}}\frac{\sum\limits_j \frac{\partial}{\partial f_k} e^{f_j}}{(\sum\limits_j e^{f_j})^2} + 1(k = y_i)\frac{e^{f_{y_i}}}{\sum\limits_j e^{f_j}}}{\left(\frac{e^{f_{y_i}}}{\sum\limits_j e^{f_j}}\right)}\right] \\
&= \frac{\frac{e^{f_{y_i}}e^{f_k}}{\sum\limits_j e^{f_j}} - 1(k = y_i)e^{f_{y_i}}}{e^{f_{y_i}}} = (p_k - 1(k = y_i))
\end{aligned} \tag{2}$$

where we used the fact that $p_k = \frac{e^{f_k}}{\sum\limits_j e^{f_j}}$ Now the other quantity to compute is straightforward:

$$\frac{\partial f_k}{\partial w_k} = \frac{\partial(w_k x_i)}{dw_k} = x_i \tag{3}$$

Finally, using relations (1), (2), (3) we have:

$$\frac{\partial L_i(f(w_k))}{\partial w_k} = \frac{\partial L_i(f_k)}{\partial f_k}\frac{\partial f_k}{\partial w_k} = (p_k - 1(k = y_i))x_i$$

Conclusion We saw how to compute the gradient of the hinge loss function. it wasn't difficult. We've just used derivative relations like: $\frac{d}{dx}\left(log(u)\right) = \frac{\frac{du}{dx}}{u}$, or $\frac{d}{dx}\left(u.v\right) = \frac{du}{dx}v + u\frac{dv}{dx}$. Then we apply the chain rule to obtain the gradient w.r.t the variables we want.