MATHEMATICS

Derivative of Softmax loss function

Asked 5 years, 8 months ago Active 3 years, 4 months ago Viewed 141k times

150

116

I am trying to wrap my head around back-propagation in a neural network with a Softmax classifier, which uses the Softmax function:

$$p_j = \frac{e^{o_j}}{\sum_k e^{o_k}}$$

This is used in a loss function of the form

$$L = -\sum_j y_j \log p_j,$$

where o is a vector. I need the derivative of L with respect to o. Now if my derivatives are right,

$$\frac{\partial p_j}{\partial o_i} = p_i(1 - p_i), \quad i = j$$

and

$$\frac{\partial p_j}{\partial o_i} = -p_i p_j, \quad i \neq j.$$

Using this result we obtain

$$\frac{\partial L}{}$$

$$
\begin{aligned}
\frac{}{\partial o_i} &= -\left(y_i(1-p_i) + \sum_{k\neq i} -p_k y_k\right) \\
&= p_i y_i - y_i + \sum_{k\neq i} p_k y_k \\
&= \left(\sum_i p_i y_i\right) - y_i
\end{aligned}
$$

According to slides I'm using, however, the result should be

$$\frac{\partial L}{\partial o_i} = p_i - y_i.$$

Can someone please tell me where I'm going wrong?

linear-algebra derivatives machine-learning

edited Oct 25 '16 at 1:15
QuantumFool
837 1 6 21

asked Sep 25 '14 at 16:43
Moos Hueting
1,697 3 9 10

10 For others who end up here, this thread is about computing the derivative of the cross-entropy function, which is the cost function often used with a softmax layer (though the derivative of the cross-entropy function uses the derivative of the softmax, -p_k * y_k, in the equation above). Eli Bendersky has an awesome derivation of the softmax and its associated cost function here: eli.thegreenplace.net/2016/... – duhaime Jan 1 '18 at 17:52

1 Answer

Active Oldest Votes

144

Your derivatives $\frac{\partial p_j}{\partial o_i}$ are indeed correct, however there is an error when you differentiate the loss function L with respect to o_i.

We have the following (where I have highlighted in *red* where you have gone wrong)

$$
\begin{aligned}
\frac{\partial L}{\partial o_i} &= -\sum_k y_k \frac{\partial \log p_k}{\partial o_i} = -\sum_k y_k \frac{1}{p_k}\frac{\partial p_k}{\partial o_i} \\
&= -y_i(1-p_i) - \sum_{k\neq i} y_k \frac{1}{p_k}(-p_k p_i) \\
&= -y_i(1-p_i) + \sum_{k\neq i} y_k (p_i) \\
&= -y_i + y_i p_i + \sum_{k\neq i} y_k (p_i) \\
&= p_i\left(\sum_k y_k\right) - y_i = p_i - y_i
\end{aligned}
$$

given that $\sum_k y_k = 1$ from the slides (as y is a vector with only one non-zero element, which is

1).

edited Sep 19 '15 at 9:09

answered Sep 25 '14 at 17:27

Alijah Ahmed
10.5k 2 14 17

1 Ah, yes, I see. And I'm not even tired - no one to blame but me! Thanks for your help, Alijah. – Moos Hueting Sep 25 '14 at 17:30

1 Moos, you are most welcome. Glad to be of help. – Alijah Ahmed Sep 25 '14 at 17:54

9 @FatalMojo I have added an extra line between the last and the penultimate lines, and highlighted some terms in blue. – Alijah Ahmed Sep 19 '15 at 9:10

5 @aceminer For the first line, the y_k do not depend on o_j, so they are constants. This leads to $\frac{\partial L}{\partial o_i} = -\sum_k y_k \frac{\partial \log p_k}{\partial o_i}$. Then, you use the differential identity $\frac{\partial \log f(x)}{\partial x} = \frac{1}{f(x)} \frac{\partial f(x)}{\partial x}$, leading to result $-\sum_k y_k \frac{1}{p_k} \frac{\partial p_k}{\partial o_i}$, at the end of the first line. For the second line, we use result $\frac{\partial p_j}{\partial o_i} = p_i(1-p_i), \quad i=j$ and $\frac{\partial p_j}{\partial o_i} = -p_i p_j, \quad i \neq j$. – Alijah Ahmed Jun 4 '17 at 11:51

2 Awesome question-awesome answer now i feel calmness inside , thanks – MIRMIX Nov 14 '17 at 22:33