Softmax / CCE Derivatives

Eckel, TJHSST AI2, Spring 2024

Background & Explanation

You've read the part of the Generative AI lab where I give you the formula for back propagating the CCE error across the last layer of the network:

$$\Delta^{N,F} = (I + -1 \cdot a^{N,F}) \cdot y$$

…where **I** represents an identity matrix (a square matrix of all zeroes except for 1s on the diagonal from upper left to lower right) that is of dimensions height(y) x height(y).

To get credit for this assignment, you need to derive this on your own paper. This is not a coding task.

Your Task

To simplify the notation, we won't refer to layer or step; we know this is the last layer and step. Instead, we'll do one specific example – a network with 4 outputs. We'll just use subscripts to identify these four specific outputs, subscripted 1 through 4. Specifically, call the outputs of the last layer, pre-activation and pre-error, dot_1, dot_2, dot_3, and dot_4. (These are values. Not vectors. The dot vector of the last layer would be made of these four values.)

That means that, for the softmax activation function, you would get for example this:

$$a_1 = \frac{e^{dot_1}}{e^{dot_1} + e^{dot_2} + e^{dot_3} + e^{dot_4}}$$

…and similarly for a_2, a_3, a_4.

We would also have y_1, y_2, y_3, and y_4 – the desired outputs.

Thus:

$$CCE = -(y_1 \ln(a_1) + y_2 \ln(a_2) + y_3 \ln(a_3) + y_4 \ln(a_4))$$

Your task comes in three parts.

1) Find $-\frac{\partial CCE}{\partial dot_1}$ and show your work meticulously. (Please note that all four terms in the formula for CCE are functions of dot_1 because of the softmax denominator!) When you're done, simplify as much as possible. You will find that this derivative is most simply expressed in terms of the a and y values, and won't contain dot_1 or any other dot values at all.
2) Analogously write $-\frac{\partial CCE}{\partial dot_2}$, $-\frac{\partial CCE}{\partial dot_3}$, and $-\frac{\partial CCE}{\partial dot_4}$. You don't need to show your work meticulously on these.
3) Meticulously show your work and demonstrate how the formula I gave above produces these four derivatives in the particular case of a four-output softmax layer. Don't forget that the addition broadcasts; refer back to Generative AI 1 if you don't remember what this means.
4) Briefly argue why this would generalize to any size final layer.

Show me when you're done, and talk me through it!