Softmax derivative

Neural networks for people who get confused easily

James McCammon

June 11, 2024

Abstract

Your abstract.

1 Introduction

The derivative of $\frac{\partial f_i}{\partial z_j}$ is given by:

$$f_i = \frac{e^{z_j}}{\sum_{k=1}^{p} e^{z_k}}$$

Define $u = e^{z_j}$ and $v = \sum_{k=1}^{p} e^{z_k}$. By substitution, we have $f_i = \frac{u}{v}$, and by the law of total derivatives, we have:

$$\frac{\partial f_i}{\partial z_j} = \frac{\partial f_i}{\partial u}\frac{\partial u}{\partial z_j} + \frac{\partial f_i}{\partial v}\frac{\partial v}{\partial z_j}$$

Now, calculating each partial:

$$\frac{\partial f_i}{\partial u} = v^{-1}$$

$$\frac{\partial f_i}{\partial v} = -uv^{-2}$$

Case 1: $i \neq j$

$$\frac{\partial u}{\partial z_j} = e^{z_j}$$

$$\frac{\partial v}{\partial z_j} = e^{z_j} \quad \text{(since all other terms in the summation are treated as constants)}$$

Case 2: $i = j$

$$\frac{\partial u}{\partial z_j} = 0$$

$$\frac{\partial v}{\partial z_j} = e^{z_i} \quad \text{(since all other terms in the summation are treated as constants)}$$

Plugging back in, we have for Case 1:

$$\begin{aligned}
\frac{\partial f_i}{\partial z_j} &= v^{-1}e^{z_j} - uv^{-2}e^{z_j} && (1)\\
&= (1 - uv^{-1})v^{-1}e^{z_j} \quad \text{(Factoring out an } v^{-1}e^{z_j}\text{)} && (2)\\
&= (1 - f_i)f_j \quad \text{Since both } uv^{-1} \text{ and } v^{-1}e^{z_j} \text{ equal } f_i \text{ by construction.} && (3)
\end{aligned}$$

And for Case 2:

$$\begin{aligned}
\frac{\partial f_i}{\partial z_j} &= v^{-1} \cdot 0 - uv^{-2}e^{z_i} && (4)\\
&= uv^{-1}v^{-1}e^{z_i} \quad \text{(Factoring out an } v^{-1}e^{z_i}\text{)} && (5)\\
&= (0 - f_i)f_j && (6)
\end{aligned}$$

These results can be combined using the Kronecker delta, which is defined as:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

And thus our equation becomes:
δ_{ij}, which simplifies to:

$$f_i(\delta_{ij} - f_j)$$