# Gradient of Softmax Followed by Cross Entropy Loss Function

Pramook Khungurn

September 27, 2018

Let there be $k$ classes in a classification problem. Let the logit of the classses be denoted by the vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)^T$. The softmax function define the probabilities $\mathbf{q} = (q_1, q_2, \ldots, q_k)^T$ where:

$$q_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)}.$$

Suppose the correct class is $c$. We have that the cross entropy loss of the probabilities $\mathbf{q}$ is:

$$L = -\ln q_c = -\ln\left(\frac{\exp(x_c)}{\sum \exp(x_j)}\right) = \ln\left(\frac{\sum \exp(x_j)}{\exp(x_c)}\right).$$

Let $\mathcal{L}$ be the overall loss function. We have that:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_i} &= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\partial L}{\partial x_i} \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\partial}{\partial x_j}\left\{\ln\left(\frac{\sum \exp(x_j)}{\exp(x_c)}\right)\right\} \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\exp(x_c)}{\sum \exp(x_j)} \cdot \frac{\partial}{\partial x_i}\left\{\frac{\sum \exp(x_j)}{\exp(x_c)}\right\}.
\end{aligned}$$

Now, there are two cases. If $i \neq c$, we have that:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_i} &= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\exp(x_c)}{\sum \exp(x_j)} \cdot \frac{1}{\exp(x_c)} \cdot \frac{\partial}{\partial x_i} \sum \exp(x_j) \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\exp(x_c)}{\sum \exp(x_j)} \cdot \frac{1}{\exp(x_c)} \cdot \exp(x_i) = \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\exp(x_i)}{\sum \exp(x_j)} \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot q_i.
\end{aligned}$$

If $i = c$, we have that:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_i} &= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\exp(x_c)}{\sum \exp(x_j)} \cdot \frac{1}{\exp(x_c)^2}\left[\frac{\partial \sum \exp(x_j)}{\partial x_c} \exp(x_c) - \left(\sum \exp(x_j)\right) \frac{\partial \exp(x_c)}{\partial x_c}\right] \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{\exp(x_c)}{\sum \exp(x_j)} \cdot \frac{1}{\exp(x_c)^2}\left[\exp(x_c)^2 - \left(\sum \exp(x_j)\right) \exp(x_c)\right] \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot \frac{1}{\sum \exp(x_j)}\left[\exp(x_c) - \left(\sum \exp(x_j)\right)\right] \\
&= \frac{\partial \mathcal{L}}{\partial L} \cdot (q_c - 1)
\end{aligned}$$