Understanding the derivative of the softmax loss function

Tang Xin

October 8, 2018

1 Qustion

There are n inputs, $\{o_1, o_2, ..., o_n\}$ for a node, the ith output is $\hat{y}_i$ and the Equation is as follow:

$$\hat{y}_i = \frac{e^{o_i}}{\sum_{k=1}^{n} e^{o_k}} = \frac{e^{o_i}}{\sum_k e^{o_k}}, i = 1, 2, ..., n \tag{1}$$

However, the true output is $(y_1, y_2, ..., y_n)$. Then, we use the cross-entropy loss function to measure the error, so we have:

$$l = -\sum_j y_j \log \hat{y}_j \tag{2}$$

How to compute the derivative of the error l with respect to the inputs $\{o_1, o_2, ..., o_n\}$.

First, we compute the derivative of $\hat{y}_j$ with respect to o_i.

If $j \neq i$, then we have,

$$\begin{aligned}\frac{\partial \hat{y}_j}{\partial o_i} &= \partial \frac{\frac{e^{o_j}}{\sum_{k \neq i} e^{o_k} + e^{o_i}}}{\partial o_i} \\ &= -\frac{e^{o_j} e^{o_i}}{\sum_{k \neq i} e^{o_k} + e^{o_i}{}^2} \\ &= -\frac{e^{o_j}}{\sum_k e^{o_k}} \frac{e^{o_i}}{\sum_k e^{o_k}} \\ &= -\hat{y}_j \hat{y}_i\end{aligned} \tag{3}$$

If $j = i$, we have

$$\begin{aligned}\frac{\partial \hat{y}_i}{\partial o_i} &= \partial \frac{\frac{e^{o_i}}{\sum_{k \neq i} e^{o_k} + e^{o_i}}}{\partial o_i} \\ &= \frac{e^{o_i} \sum_k e^{o_k} - e^{o_i} e^{o_i}}{\sum_k e^{o_k}{}^2} \\ &= \frac{e^{o_i}}{\sum_k e^{o_k}} \frac{\sum_k e^{o_k} - e^{o_i}}{\sum_k e^{o_k}} \\ &= \hat{y}_i (1 - \hat{y}_i)\end{aligned} \tag{4}$$

On the other hand, we should compute the derivative of the error l with respect to $\hat{y}_j$, so we get,

$$
\begin{aligned}
\frac{\partial l}{\partial o_i} &= \frac{-\sum_j y_j \log \hat{y}_j}{o_i} = -\sum_j y_j \frac{\partial \log \hat{y}_j}{\partial o_i} \\
&= -y_i \frac{\partial \log \hat{y}_i}{\partial o_i} - \sum_{j \neq i} y_j \frac{\partial log \hat{y}_j}{o_i} \\
&= -y_i (\frac{1}{\hat{y}_i})(\hat{y}_i(1-\hat{y}_i)) - \sum_{j \neq i} y_j (\frac{1}{\hat{y}_j})(-\hat{y}_i \hat{y}_j)) \\
&= -y_i(1-\hat{y}_i) + \sum_{j \neq i} y_j \hat{y}_i \\
&= -y_i + y_i \hat{y}_i + \sum_{j \neq i} y_j \hat{y}_i \\
&= \sum_j y_j \hat{y}_i - y_i
\end{aligned} \tag{5}
$$

As we know, $\sum_j y_j = 1$, so we get

$$
\frac{\partial l}{\partial o_i} = \hat{y}_i - y_i \tag{6}
$$

Hence, let $o = (o_1, o_2, ..., o_n)$, we have

$$
\begin{aligned}
\frac{\partial l}{\partial o} &= (\hat{y}_1 - y_1, \hat{y}_1 = 2 - y_2, ..., \hat{y}_n - y_n) \\
&= (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n) - (y_1, y_2, ..., y_n) \\
&= \hat{y} - y
\end{aligned} \tag{7}
$$

We should note that, Equation 6 is just for one sample. If we have many samples, we should sum them up. It means, for N samples, we have

$$
\frac{\partial l}{\partial o} = \sum_{i=1}^{N} \{\hat{y}^i - y^i\} \tag{8}
$$

While, $\hat{y}^i$ and y^i are vectors.

References