SimplePyML Softmax Layer

Vikram Rangarajan

January 2024

1 Definitions

This layer takes an input X and applies the softmax function to it to achieve the output, Y.
Formula:

$$Y_i = \frac{e^{X_i}}{\sum_{k=1}^{n} e^{X_k}} \text{ for all i}$$

Given $\frac{\partial L}{\partial Y}$, we must calculate $\frac{\partial L}{\partial X}$.
First, observe that every Y_i is dependent on every X_j, so we must use the multivariable chain rule to get

$$\frac{\partial L}{\partial X_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial Y_j} \frac{\partial Y_j}{\partial X_i}$$

We know $\frac{\partial L}{\partial Y_j}$ as it is given. We must derive $\frac{\partial Y_j}{\partial X_i}$.
When dealing with the softmax function, it is easier to differentiate with logarithmic differentiation.

$$\frac{\partial \log_b(Y_j)}{\partial X_i} = \frac{\partial \log_b(Y_j)}{\partial X_i} \frac{\partial Y_j}{\partial Y_j} = \frac{\partial Y_j}{\partial X_i} \frac{\partial \log_b(Y_j)}{\partial Y_j} = \frac{\partial Y_j}{\partial X_i} \frac{1}{Y_j \ln b}$$

Rearranging, we get

$$\frac{\partial Y_j}{\partial X_i} = \frac{\partial \log_b(Y_j)}{\partial X_i} Y_j \ln b$$

$$\log_b(Y_j) = \log_b\left(\frac{e^{X_j}}{\sum_{k=1}^{n} e^{X_k}}\right) = \log_b(e^{X_j}) - \log_b\left(\sum_{k=1}^{n} e^{X_k}\right) = \frac{X_j}{\ln(b)} - \log_b\left(\sum_{k=1}^{n} e^{X_k}\right)$$

$$\frac{\partial \log_b(Y_j)}{\partial X_i} = \frac{\partial}{\partial X_i}\left(\frac{X_j}{\ln(b)} - \log_b\left(\sum_{k=1}^{n} e^{X_k}\right)\right) = \frac{1}{\ln(b)}(i == j) - \frac{e^{X_i}}{\ln(b) \sum_{k=1}^{n} e^{X_k}} = \frac{1}{\ln(b)}(i == j) - \frac{Y_i}{\ln(b)}$$

Plugging this result into the previous equation,

$$\frac{\partial Y_j}{\partial X_i} = \frac{\partial \log_b(Y_j)}{\partial X_i} Y_j \ln b = Y_j\left[\frac{1}{\ln(b)}(i == j) - \frac{Y_i}{\ln(b)}\right] \ln(b) = Y_j((i == j) - Y_i)$$

Now, we can get $\frac{\partial L}{\partial X_i}$:

$$\frac{\partial L}{\partial X_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial Y_j} \frac{\partial Y_j}{\partial X_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial Y_j} Y_j((i == j) - Y_i)$$

We can see that $\frac{\partial L}{\partial X} = \begin{bmatrix} \frac{\partial L}{\partial Y_1} Y_1(1 - Y_1) + \frac{\partial L}{\partial Y_2} Y_2(-Y_1) + \frac{\partial L}{\partial Y_3} Y_3(-Y_1) + ... + \frac{\partial L}{\partial Y_n} Y_n(-Y_1) \\ \frac{\partial L}{\partial Y_1} Y_1(-Y_2) + \frac{\partial L}{\partial Y_2} Y_2(1 - Y_2) + \frac{\partial L}{\partial Y_3} Y_3(-Y_2) + ... + \frac{\partial L}{\partial Y_n} Y_n(-Y_2) \\ ... \\ \frac{\partial L}{\partial Y_1} Y_1(-Y_n) + \frac{\partial L}{\partial Y_2} Y_2(Y_n) + \frac{\partial L}{\partial Y_3} Y_3(-Y_n) + ... + \frac{\partial L}{\partial Y_n} Y_n(1 - Y_n) \end{bmatrix}$

$$\text{Let } g_k = \frac{\partial L}{\partial Y_k} Y_k \text{ for all k. Then,}$$

$$\frac{\partial L}{\partial X} = \begin{bmatrix} g_1(1-Y_1) + g_2(-Y_1) + g_3(-Y_1) + ... + g_n(-Y_1) \\ g_1(-Y_2) + g_2(1-Y_2) + g_3(-Y_2) + ... + g_n(-Y_2) \\ ... \\ g_1(-Y_n) + g_2(Y_n) + g_3(-Y_n) + ... + g_n(1-Y_n) \end{bmatrix} = \begin{bmatrix} g_1 - g_1Y_1 - g_2Y_1 - g_3Y_1 - ... - g_nY_1 \\ g_2 - g_1Y_2 - g_2Y_2 - g_3Y_2 - ... - g_nY_2 \\ ... \\ g_n - g_1Y_n - g_2Y_n - g_3Y_n - ... - g_nY_n \end{bmatrix}$$

$$= \begin{bmatrix} -g_1Y_1 - g_2Y_1 - g_3Y_1 - ... - g_nY_1 \\ -g_1Y_2 - g_2Y_2 - g_3Y_2 - ... - g_nY_2 \\ ... \\ -g_1Y_n - g_2Y_n - g_3Y_n - ... - g_nY_n \end{bmatrix} + \begin{bmatrix} g_1 \\ g_2 \\ ... \\ g_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} \odot \begin{bmatrix} -\sum_{k=1}^n g_k \\ -\sum_{k=1}^n g_k \\ ... \\ -\sum_{k=1}^n g_k \end{bmatrix} + \begin{bmatrix} g_1 \\ g_2 \\ ... \\ g_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} \cdot (-\textstyle\sum_{k=1}^n g_k) + \begin{bmatrix} g_1 \\ g_2 \\ ... \\ g_n \end{bmatrix}$$

Finally, we get these formulas:

$$g_k = \frac{\partial L}{\partial Y_k} Y_k \Rightarrow g = \frac{\partial L}{\partial Y} \odot Y$$

$$\frac{\partial L}{\partial X} = (-\sum_{k=1}^{n} g_k)Y + g$$