NN Loss Exploration

2023-01-24

1 Cross Entropy

$$\ell_{CE} = -\mathbf{y}_n^T \mathbf{s}_n + \log \sum_k e^{s_{n,k}}$$

if we take $\frac{\partial \ell_{CE}}{\partial s_{n,i}}$ we get the following

$$\frac{\partial \ell_{CE}}{\partial s_{n,i}} = -y_{n,i} + \frac{e^{s_{n,i}}}{\sum_k e^{s_{n,k}}}$$

$$\nabla_{s_n} \ell_{CE} = \begin{bmatrix} \frac{\partial \ell_{CE}}{\partial s_{n,1}} \\ \frac{\partial \ell_{CE}}{\partial s_{n,2}} \\ \dots \\ \frac{\partial \ell_{CE}}{\partial s_{n,k}} \end{bmatrix} = \begin{bmatrix} -y_{n,1} + \frac{e^{s_{n,1}}}{\sum_k e^{s_{n,k}}} \\ -y_{n,2} + \frac{e^{s_{n,2}}}{\sum_k e^{s_{n,k}}} \\ \dots \\ -y_{n,k} + \frac{e^{s_{n,k}}}{\sum_k e^{s_{n,k}}} \end{bmatrix} = \begin{bmatrix} -y_{n,1} \\ -y_{n,2} \\ \dots \\ -y_{n,k} \end{bmatrix} + \begin{bmatrix} \frac{e^{s_{n,1}}}{\sum_k e^{s_{n,k}}} \\ \frac{e^{s_{n,2}}}{\sum_k e^{s_{n,k}}} \\ \dots \\ \frac{e^{s_{n,k}}}{\sum_k e^{s_{n,k}}} \end{bmatrix} = \begin{bmatrix} -y_{n,1} \\ -y_{n,2} \\ \dots \\ -y_{n,k} \end{bmatrix} + \frac{1}{\sum_k e^{s_{n,k}}} \begin{bmatrix} e^{s_{n,1}} \\ e^{s_{n,2}} \\ \dots \\ e^{s_{n,k}} \end{bmatrix} = -\mathbf{y}_n + \frac{e^{\mathbf{s}_n}}{\sum_k e^{s_{n,k}}}$$

Therefore,

$$\nabla_{s_n} \ell_{CE} = -\mathbf{y}_n + \frac{e^{\mathbf{s}_n}}{\sum_k e^{s_{n,k}}} = -\mathbf{y}_n + \mathbf{softmax}(\mathbf{s}_n)$$

2 MSE

$$\ell_{MSE} = \frac{1}{K} ||\mathbf{y}_n - \mathbf{s}_n||_2^2 = \frac{1}{K} \left[\sum_{i=1}^{K} (y_{n,i} - s_{n,i})^2 \right]$$

$$\frac{\partial \ell_{MSE}}{\partial s_{n,j}} = \frac{1}{K} \left[\sum_{i=1}^{K} \frac{\partial (y_{n,i} - s_{n,i})^2}{\partial s_{n,j}} \right]$$

when $i \neq j$ then $y_{n,i} - s_{n,i}$ is not a function of $s_{n,j}$. Therefore, $\frac{\partial (y_{n,i} - s_{n,i})^2}{\partial s_{n,j}} = 0$ for $i \neq j$

when $i = j$ then $\frac{\partial (y_{n,i} - s_{n,i})^2}{\partial s_{n,j}} = 2(y_{n,j} - s_{n,j})(-1)$

Therefore,

$$\frac{\partial \ell_{MSE}}{\partial s_{n,j}} = \frac{1}{K}(2(y_{n,j} - s_{n,j})(-1)) = \frac{-2}{K}(y_{n,j} - s_{n,j})$$

Therefore,

$$\nabla_{s_n} \ell_{MSE} = \frac{-2}{K}(\mathbf{y}_n - \mathbf{s}_n)$$

3 MSE Variant

$$\ell_{softmse} = \frac{1}{K}||\mathbf{y}_n - \mathbf{softmax}(\mathbf{s}_n)||_2^2 = \frac{1}{K}\left[\sum_{i=1}^{K}\left(y_{n,i} - \frac{e^{s_{n,i}}}{\sum_k e^{s_{n,k}}}\right)^2\right]$$

let $f_i = \left(\frac{e^{s_{n,i}}}{\sum_k e^{s_{n,k}}}\right)$

then by the chain rule:

$$\frac{\partial \ell_{softmse}}{\partial s_{n,j}} = \frac{1}{K}\left[\sum_{i=1}^{K}\left(y_{n,i} - \frac{e^{s_{n,i}}}{\sum_k e^{s_{n,k}}}\right)(-2)\left(\frac{\partial f_i}{\partial s_{n,j}}\right)\right]$$

However, $\frac{\partial f_i}{\partial s_{n,j}}$ takes different forms depending on if $i = j$

if $i = j$ we use the quotient rule to find the derivative:

$$\frac{\partial f_i}{\partial s_{n,j}} = \frac{e^{s_{n,i}}\sum_k e^{s_{n,k}} - e^{s_{n,i}}e^{s_{n,j}}}{(\sum_k e^{s_{n,k}})^2} = \frac{e^{s_{n,i}}(\sum_k e^{s_{n,k}} - e^{s_{n,j}})}{(\sum_k e^{s_{n,k}})(\sum_k e^{s_{n,k}})} = softmax(s_{n,i})(1 - softmax(s_{n,j}))$$

if $i \neq j$ then $e^{s_{n,i}}$ is not a function of $s_{n,j}$ we don't use the quotient rule and get the following

$$\frac{\partial f_i}{\partial s_{n,j}} = -\frac{e^{s_{n,i}}}{(\sum_k e^{s_{n,k}})^2}e^{s_{n,j}} = \frac{-e^{s_{n,i}}e^{s_{n,j}}}{(\sum_k e^{s_{n,k}})(\sum_k e^{s_{n,k}})} = -softmax(s_{n,i})softmax(s_{n,j}) = softmax(s_{n,i})(0 - softmax(s_{n,j}))$$

Therefore we can rewrite $\frac{\partial \ell}{\partial s_{n,j}}$ as

$$\begin{aligned}\frac{\partial \ell_{softmse}}{\partial s_{n,j}} &= \frac{-2}{K}\left[\sum_{i=1}^{K}\left(y_{n,i} - \frac{e^{s_{n,i}}}{\sum_k e^{s_{n,k}}}\right)(softmax(s_{n,i})(1_{i=j} - softmax(s_{n,j})))\right] \\ &= \frac{-2}{K}\left[\sum_{i=1}^{K}(y_{n,i} - softmax(s_{n,i}))\,(softmax(s_{n,i})(1_{i=j} - softmax(s_{n,j})))\right]\end{aligned}$$

Therefore,

$$\nabla_{s_n} \ell_{softmse} = \frac{-2}{K}\left[\sum_{i=1}^{K}(y_{n,i} - softmax(s_{n,i}))\,(softmax(s_{n,i})(1_{i=j} - \mathbf{softmax}(\mathbf{s}_n)))\right]$$

4 Analyzing gradients

If we take the partial derivative of MSE with respect to $s_{n,k}$, we then update $s_{n,k}$ in the direction of the sign of the derivative. If k is not the correct label for s_n but we get a negative derivative we will update s_n by increasing it for an incorrect label which results in an incorrect update of the parameter which will hinder learning the true correct weights. We only want to increase s_n when it is the true label. During gradient descent the model will be moving in the opposite direction of the actual solution.

My proposed MSE variant loss function is the following:

$$\ell_{MSEVar} = \frac{1}{K}||\mathbf{y}_n - e^{\mathbf{s}_n}||_2^2$$

This solves the problem of negative gradients for incorrect classification

$$\frac{1}{K}||\mathbf{y}_n - e^{\mathbf{s}_n}||_2^2 = \frac{1}{K}\left[\sum_{i=1}^{K}(y_{n,i} - e^{s_{n,i}})^2\right]$$

$$\frac{\partial \ell_{MSEVar}}{\partial s_{n,j}} = \frac{1}{K}\left[\sum_{i=1}^{K}\frac{\partial\,(y_{n,i} - e^{s_{n,i}})^2}{\partial s_{n,j}}\right]$$

And just as with the gradient of the MSE, if $i \neq j$ then $\frac{\partial (y_{n,i} - e^{s_{n,i}})^2}{\partial s_{n,j}} = 0$. Therefore,

$$\frac{\partial \ell_{MSEVar}}{\partial s_{n,j}} = \frac{-2}{K}(y_{n,j} - e^{s_{n,j}})(e^{s_{n,j}})$$

and for an incorrect class $y_{j,n} = 0$, therefore

$$= \frac{-2}{K}(0 - e^{s_{n,j}})(e^{s_{n,j}}) = \frac{-2}{K}(-e^{s_{n,j}})(e^{s_{n,j}}) = \frac{2}{K}(e^{2s_{n,j}}) > 0$$

Thus,

$$\nabla_{s_n}\ell_{MSEVar} = \frac{-2}{K}(\mathbf{y}_n - e^{\mathbf{s}_n}) \odot (e^{\mathbf{s}_n})$$