CENG499 Homework 1: Part 1

Ozan Kamalı

November 2, 2024

Part 1

These are the equations which we will derive the update rules for:

1. Squared Error Loss for Weight Update:
$$w_{ij}^{\text{new}} = w_{ij} - \alpha \frac{\partial SE(y, O_0)}{\partial w_{ij}}$$
2. Squared Error Loss for Gamma Update:
$$\gamma_{k0}^{\text{new}} = \gamma_{k0} - \alpha \frac{\partial SE(y, O_0)}{\partial \gamma_{k0}}$$
3. Cross-Entropy Loss for Weight Update:
$$w_{ij}^{\text{new}} = w_{ij} - \alpha \frac{\partial CE([l_0, l_1, l_2], O = [O_0, O_1, O_2])}{\partial w_{ij}}$$
4. Cross-Entropy Loss for Gamma Update:
$$\gamma_{jk}^{\text{new}} = \gamma_{jk} - \alpha \frac{\partial CE([l_0, l_1, l_2], O = [O_0, O_1, O_2])}{\partial \gamma_{jk}}$$

Part 1: Regression

0.1 Δw_{ij}

The Δw_{ij} represents the update needed in the weights between the input layer and the hidden layer for each data sample, where α represents the learning rate:

$$\Delta w_{ij} = -\alpha * \frac{\partial SE(y, O_0)}{\partial w_{ij}}$$

Using the chain rule derivative rule we can write this as:

$$= -\alpha * \frac{\partial SE(y, O_0)}{\partial O_0} \cdot \frac{\partial O_0}{\partial H_j} \cdot \frac{\partial H_j}{\partial w_{ij}}$$

where:

$$SE(y, O_0) = (y - O_0)^2, O_0 = \left(\sum_{j=0}^{3} H_j \gamma_{j0} \right), H_j = \sigma \left(\sum_{i=0}^{2} X_i w_{ij} \right)$$

Using these given values, the chain rule equation becomes:

$$= -\alpha * \left(\frac{\partial (y - O_0)^2}{\partial O_0} \right) \left(\frac{\partial \left(\sum_{j=0}^{3} H_j \gamma_{j0} \right)}{\partial H_j} \right) \left(\frac{\partial \sigma \left(\sum_{i=0}^{2} X_i w_{ij} \right)}{\partial w_{ij}} \right)$$

The first term's derivative is simple, we just take the derivative w.r.t O_0:

$$\frac{\partial (y - O_0)^2}{\partial O_0} = -2(y - O_0)$$

The second term's derivative contains the sigma sum, but knowing that differentiation can be distributed across the terms of the sum, provided the functions being summed are differentiable w.r.t the variable H_j. Thus we can write:

$$\frac{\partial \left(\sum_{j=0}^{3} H_j \gamma_{j0} \right)}{\partial H_j} = \sum_{j=0}^{3} \frac{\partial (H_j \gamma_{j0})}{\partial H_j}$$

Now, differentiating each term individually the final result becomes:

$$= \sum_{j=0}^{3} \gamma_{j0} = \gamma_{j0}$$

The same logic can be applied to the third term. However we need to consider the sigmoid function first.
Here is what happens when we take the derivative of the sigmoid function when there is another function to be taken the derivative of:

$$\frac{d}{dx} \sigma(f(x)) = \sigma(f(x))(1 - \sigma(f(x)))f'(x)$$

Now, if we take the derivative with respect to w_{ij}, we get:

$$\frac{d(H_j)}{d(w_{ij})} = (H_j)(1 - (H_j)) \frac{d(\sum_{i=0}^{2} X_i w_{ij})}{d(w_{ij})}$$

Which can be written as using the sigma sum and derivative rule we have used in the second step:

$$= (H_j)(1 - (H_j))(X_i)$$

Combining these, the result becomes:

$$= -\alpha * (-2(y - O_0)) (\gamma_{j0}) (H_j (1 - H_j) (X_i))$$

Therefore, the final expression for updating w_{ij} is:

$$\Delta w_{ij} = -\alpha * (-2(y - O_0)) (\gamma_{j0}) (H_j (1 - H_j) (X_i))$$

0.2 $\Delta\gamma_{j0}$

The $\Delta\gamma_{j0}$ represents the update needed in the weights between the hidden layer and the output for each data sample, where α represents the learning rate:

$$\Delta\gamma_{j0} = -\alpha * \frac{\partial SE(y, O_0)}{\partial\gamma_{j0}}$$

Using the chain rule, we can rewrite this as:

$$= -\alpha * \frac{\partial SE(y, O_0)}{\partial O_0} \cdot \frac{\partial O_0}{\partial\gamma_{j0}}$$

where:

$$SE(y, O_0) = (y - O_0)^2, \quad O_0 = \left(\sum_{j=0}^{3} H_j\gamma_{j0}\right)$$

Using these definitions, the chain rule expression becomes:

$$= -\alpha * \left(\frac{\partial(y - O_0)^2}{\partial O_0}\right)\left(\frac{\partial\left(\sum_{j=0}^{3} H_j\gamma_{j0}\right)}{\partial\gamma_{j0}}\right)$$

The first term's derivative with respect to O_0 is simple again:

$$\frac{\partial(y - O_0)^2}{\partial O_0} = -2(y - O_0)$$

The second term's derivative with respect to γ_{j0} contains a summation over H_j. Since differentiation can be distributed across the terms of the sum, provided the functions being summed are differentiable with respect to γ_{j0}, we can write:

$$\frac{\partial\left(\sum_{j=0}^{3} H_j\gamma_{j0}\right)}{\partial\gamma_{j0}} = \sum_{j=0}^{3} \frac{\partial(H_j\gamma_{j0})}{\partial\gamma_{j0}}$$

Differentiating each term individually, we find:

$$= \sum_{j=0}^{3} H_j = H_j$$

Combining these, the result becomes:

$$= -\alpha * (-2(y - O_0))\,(H_j)$$

Therefore, the final expression for updating γ_{j0} is:

$$\Delta\gamma_{j0} = -\alpha * (-2(y - O_0)H_j)$$

0.3 $\Delta w_{\textbf{bias}}$

The Δw_{bias} represents the update needed for the bias term between the input layer and the hidden layer for each data sample, where α represents the learning rate:

$$\Delta w_{\text{bias}} = -\alpha * \frac{\partial SE(y, O_0)}{\partial w_{\text{bias}}}$$

Using the chain rule derivative rule, we can write this as:

$$= -\alpha * \frac{\partial SE(y, O_0)}{\partial O_0} \cdot \frac{\partial O_0}{\partial H_j} \cdot \frac{\partial H_j}{\partial w_{\text{bias}}}$$

We can use the previously derived values for the first two terms:

$$= -\alpha * (-2(y - O_0)) (\gamma_{j0}) \left(\frac{\partial H_j}{\partial w_{\text{bias}}} \right)$$

The sigmoid derivative for the third term will stay the same but this time the derivative inside the sigmoid will change because we are taking the derivative w.r.t w_{bias}, we can express this as:

$$H_j = \sigma \left(\sum_{i=0}^{2} X_i w_{ij} + w_{\text{bias}} \right)$$

Taking the derivative of H_j with respect to w_{bias}:

$$\frac{d(H_j)}{d(w_{\text{bias}})} = \sigma'(f(x)) = H_j(1 - H_j) * 1$$

Thus, the final expression for the update rule becomes:

$$\Delta w_{\text{bias}} = -\alpha * (-2(y - O_0)) (\gamma_{j0}) (H_j(1 - H_j)) * 1$$

0.4 $\Delta \gamma_{\textbf{bias}}$

The $\Delta \gamma_{\text{bias}}$ represents the update needed for the bias term between the hidden layer and the output for each data sample, where α represents the learning rate:

$$\Delta \gamma_{\text{bias}} = -\alpha * \frac{\partial SE(y, O_0)}{\partial \gamma_{\text{bias}}}$$

Using the chain rule, we can rewrite this as:

$$= -\alpha * \frac{\partial SE(y, O_0)}{\partial O_0} \cdot \frac{\partial O_0}{\partial \gamma_{\text{bias}}}$$

We can again use the previously derived value for the first term:

$$= -\alpha * (-2(y - O_0)) \left(\frac{\partial O_0}{\partial \gamma_{\text{bias}}} \right)$$

The second term requires the derivative with respect to γ_{bias}:

$$O_0 = \left(\sum_{j=0}^{3} H_j \gamma_{j0} + \gamma_{\text{bias}} \right)$$

Taking the derivative of O_0 with respect to γ_{bias}:

$$\frac{\partial O_0}{\partial \gamma_{\text{bias}}} = 1$$

Thus, combining these results, the final expression for the update rule becomes:

$$\Delta \gamma_{\text{bias}} = -\alpha * (-2(y - O_0)) * 1$$

Part 1: Classification

0.5 $\Delta \gamma_{jk}$

The $\Delta \gamma_{jk}$ represents the update needed in the weights between the hidden layer and the output for each data sample, where α represents the learning rate:

$$\Delta \gamma_{jk} = -\alpha * \frac{\partial CE(l, O)}{\partial \gamma_{jk}}$$

Using the chain rule, we can rewrite this as:

$$= -\alpha * \frac{\partial CE(l, O)}{\partial X_k} \cdot \frac{\partial X_k}{\partial \gamma_{jk}}$$

where:

$$CE(l, O) = -\sum_{i=0} l_i \log(O_i), X_k = \sum_{j=0}^{3} H_j \gamma_{jk}, O_k = softmax(X_k, X)$$

Using these definitions, the chain rule expression becomes:

$$= -\alpha * \left(\frac{\partial CE(l, O)}{\partial X_k} \right) \left(\frac{\partial X_k}{\partial \gamma_{jk}} \right)$$

Calculating the first term's derivative with respect to X_k:

$$\frac{\partial CE(l, O)}{\partial X_k} = -\sum_{i=0}^{2} l_i \left(\frac{d \log(O_i)}{d O_i} \frac{d O_i}{d X_k} \right)$$

To calculate this we need to first consider the softmax function derivative. It is defined as:

$$O_i(1 - O_i), i = k$$

$$-O_i O_k, i \neq k$$

Also considering the derivative of the log function:

$$\frac{d}{dx}\log(x) = \frac{1}{x}$$

Deconstructing the sum into two parts where $i = k$, and $i \neq k$ we get the equation:

$$-\sum_{i \neq k}^{2} \left(\frac{l_i}{O_i}\right). - O_i O_k + \left(\frac{l_k}{O_k}\right) O_k(1 - O_k)$$

Simplifying it further we get:

$$-\sum_{i \neq k}^{2} -l_i.O_k + l_k(1 - O_k)$$

Since O_k is independent of the sum where $i \neq k$ we can move it outside of the sum:

$$-(-O_k \sum_{i \neq k}^{2} l_i + l_k(1 - O_k))$$

Another rule we have to remember here is that the l_i values are one-hot encoded. Meaning:

$$\sum_{i=0}^{2} l_i = 1$$

This gives us:

$$\sum_{i \neq k}^{2} l_i = 1 - l_k$$

Putting this back in the equation we get:

$$= -(l_k - O_k)$$

$$= O_k - l_k$$

So the derivative for the first term becomes:

$$\frac{\partial CE(l, O)}{\partial X_k} = O_k - l_k$$

Using the previous values and sum properties of the derivative, the second term becomes:

$$\frac{\partial X_k}{\partial \gamma_{ij}} = \sum_{j=0}^{3} H_j = H_j$$

Thus, the final expression for updating γ_{jk} is:

$$\Delta\gamma_{jk} = -\alpha * (O_k - l_k) H_j$$

0.6 Δw_{ij}

The weight update Δw_{ij} represents the adjustment needed for the weights between the input layer and the hidden layer for each data sample, where α represents the learning rate. Starting from the chain rule, we can express this as:

$$\Delta w_{ij} = -\alpha * \frac{\partial CE(l, O)}{\partial w_{ij}}$$

Using the chain rule, we rewrite this as:

$$= -\alpha * \frac{\partial CE(l, O)}{\partial X_k} \cdot \frac{\partial X_k}{\partial H_j} \cdot \frac{\partial H_j}{\partial w_{ij}}$$

Calculating the first term, we already have:

$$\frac{\partial CE(l, O)}{\partial X_k} = O_k - l_k$$

Now, we need to compute the second term where:

$$X_k = \sum_{j=0}^{3} H_j \gamma_{jk}$$

$$\frac{\partial X_k}{\partial H_j} = \gamma_{jk}$$

Next, we compute the third term where the hidden layer uses the sigmoid activation function again:

$$\frac{\partial H_j}{\partial w_{ij}} = H_j(1 - H_j) \frac{d(\sum_{i=0}^{4} X_i w_{ij})}{dw_{ij}}$$

$$\frac{\partial H_j}{\partial w_{ij}} = H_j(1 - H_j) X_i$$

Putting it all together:

$$= (O_k - l_k) \gamma_{jk} X_i$$

Thus, the final expression for updating w_{ij} is:

$$\Delta w_{ij} = -\alpha * (O_k - l_k) \gamma_{jk} X_i$$

0.7 $\Delta \gamma_{\textbf{bias}}$

The weight bias update $\Delta \gamma_{\text{bias}}$ represents the bias of the output layer for each data sample, where α represents the learning rate. Starting from the chain rule, we can express this as:

$$\Delta \gamma_{\text{bias}} = -\alpha * \frac{\partial CE(l, O)}{\partial \gamma_{\text{bias}}}$$

Using the chain rule, we rewrite this as:

$$= -\alpha * \frac{\partial CE(l,O)}{\partial X_k} \cdot \frac{\partial X_k}{\partial \gamma_{\text{bias}}}$$

Calculating the first term, we already have:

$$\frac{\partial CE(l,O)}{\partial X_k} = O_k - l_k$$

Now, we need to compute the second term where:

$$X_k = \sum_{j=0}^{3} H_j \gamma_{jk} + \gamma_{\text{bias}}$$

We can see this equals to:

$$\frac{\partial X_k}{\partial \gamma_{\text{bias}}} = 1$$

Putting it all together:

$$= (O_k - l_k) * 1$$

Thus, the final expression for updating γ_{bias} is:

$$\Delta \gamma_{\text{bias}} = -\alpha * (O_k - l_k) * 1$$

0.8 Δw_{bias}

The weight bias update Δw_{bias} represents the bias of the hidden layer for each data sample, where α represents the learning rate. Starting from the chain rule, we can express this as:

$$\Delta w_{\text{bias}} = -\alpha * \frac{\partial CE(l,O)}{\partial w_{\text{bias}}}$$

Using the chain rule, we rewrite this as:

$$= -\alpha * \frac{\partial CE(l,O)}{\partial X_k} \cdot \frac{\partial X_k}{\partial H_j} \cdot \frac{\partial H_j}{\partial w_{\text{bias}}}$$

Calculating the first term, we already have:

$$\frac{\partial CE(l,O)}{\partial X_k} = O_k - l_k$$

For the second term, we already have:

$$X_k = \sum_{j=0}^{3} H_j \gamma_{jk}$$

$$\frac{\partial X_k}{\partial H_j} = \gamma_{jk}$$

Next, we compute the third term. The only difference is we are calculating the derivative w.r.t w_{bias} inside the sigmoid activation function:

$$\frac{\partial H_j}{\partial w_{\text{bias}}} = H_j(1 - H_j)\frac{d(H_j)}{d(w_{bias})}$$

$$H_j = \sigma(\sum_{j=0}^{4} X_i w_{ij} + w_{bias})$$

We can observe this derivative equals to 1:

$$\frac{\partial H_j}{\partial w_{\text{bias}}} = H_j(1 - H_j) * 1$$

Putting it all together:

$$= (O_k - l_k)\gamma_{jk} * 1$$

Thus, the final expression for updating w_{bias} is:

$$\Delta w_{\text{bias}} = -\alpha * (O_k - l_k)\gamma_{jk} * 1$$