499-hw1-part1

Batuhan Karaca 2310191

October 2022

1 Derivation

Note that for weights, we will use ω instead of a as used in the assignment text. Our calculations are based around a specific weight ω_{ij}^{ℓ} such that the superscript $\ell \in \{0, 1, ..., L\}$ denotes the layer, whereas the the subscripts $i \in \{0, 1, ..., n_\ell - 1\}$ and $j \in \{0, 1, ..., n_{\ell+1} - 1\}$ tell that the weight is between the nodes O_i^{ℓ} and $O_j^{\ell+1}$, as given in the assignment text. Note that L is the last layer, and n_ℓ denotes the number of nodes in a layer ℓ.

The update rule for both classification and regression problems (and commonly used in many networks) is as follows

$$\omega_{ij}^{\ell} = \omega_{ij}^{\ell} - \alpha \frac{\delta E(x)}{\delta \omega_{ij}^{\ell}}$$

α is the step size and $E(x)$ is the function to be minimized (loss function). If we find $\frac{\delta E(x)}{\delta \omega_{ij}^{\ell}}$, we find ω_{ij}^{ℓ} for both problems. We know that in feed-forward networks, every output O_i^{ℓ} is a function of outputs in the layers $\{0, 1, ..., \ell - 1\}$. Therefore, we can expand the derivative using Chain rule as

$$\frac{\delta E(x)}{\delta \omega_{i_j i_n}^{L-n}} = \sum_{\text{all paths to } O_{i_n}^{L-n+1}} \frac{\delta E(x)}{\delta O_{i_1}^{L}} \frac{\delta O_{i_1}^{L}}{\delta O_{i_2}^{L-1}} ... \frac{\delta O_{i_n}^{L-n+1}}{\delta \omega_{i_j i_n}^{L-n}}$$

We have sum of combinations of nodes. Using the superposition property, we can rearrange the terms and get

$$\frac{\delta E(x)}{\delta \omega_{i_j i_n}^{L-n}} = (\sum_{i_{n-1}=0}^{n_{L-n+2}-1} (\sum_{i_{n-2}=0}^{n_{L-n+3}-1} (...(\sum_{i_1=0}^{n_L-1} \frac{\delta E(x)}{\delta O_{i_1}^{L}} \frac{\delta O_{i_1}^{L}}{\delta O_{i_2}^{L-1}})...) \frac{\delta O_{i_{n-2}}^{L-n+3}}{\delta O_{i_{n-1}}^{L-n+2}}) \frac{\delta O_{i_{n-1}}^{L-n+2}}{\delta O_{i_n}^{L-n+1}}) \frac{\delta O_{i_n}^{L-n+1}}{\delta \omega_{i_j i_n}^{L-n}}$$

Let us find the partial derivatives for the hidden layers

$$O_i^{\ell} = \sigma(s_i^{\ell-1} = \sum_{j=0}^{n_{\ell-1}-1} O_j^{\ell-1} \omega_{ji}^{\ell-1})$$

σ is the activation function. Then

$$\frac{\delta O_i^\ell}{\delta \omega_{ji}^{\ell-1}} = \sigma'(s_i^{\ell-1})O_j^{\ell-1} = \sigma_i^{\ell-1}O_j^{\ell-1}$$
$$\frac{\delta O_i^\ell}{\delta O_j^{\ell-1}} = \sigma'(s_i^{\ell-1})\omega_{ji}^{\ell-1} = \sigma_i^{\ell-1}\omega_{ji}^{\ell-1}$$

Substituting gives

$$\frac{\delta E(x)}{\delta \omega_{i_j i_n}^{L-n}} = (\sum_{i_{n-1}=0}^{n_{L-n+2}-1}(\sum_{i_{n-2}=0}^{n_{L-n+3}-1}(...(\sum_{i_1=0}^{n_L-1}\frac{\delta E(x)}{\delta O_{i_1}^L}\frac{\delta O_{i_1}^L}{\delta O_{i_2}^{L-1}})...)\sigma_{i_{n-2}}^{L-n+2}\omega_{i_{n-1}i_{n-2}}^{L-n+2})\sigma_{i_{n-1}}^{L-n+1}\omega_{i_n i_{n-1}}^{L-n+1})\sigma_{i_n}^{L-n}O_{i_j}^{L-n}$$

We can deduce a recursive expression such that

$$\frac{\delta E(x)}{\delta \omega_{ij}^{L-n}} = O_i^{L-n}\Delta_j^n$$
$$\Delta_i^\ell = \sigma_i^{L-\ell}\sum_{j=0}^{n_{L-\ell+2}-1}\omega_{ij}^{L-\ell+1}\Delta_j^{\ell-1}$$
$$\Delta_i^2 = \sigma_i^{L-2}\sum_{j=0}^{n_L-1}\frac{\delta E(x)}{\delta O_j^L}\frac{\delta O_j^L}{\delta O_i^{L-1}}$$

For sigmoid function (and many other activation functions such as ReLU), we can find $\sigma_i^{\ell-1}$ in terms of O_i^ℓ. For the sigmoid function, the derivative $\sigma_i^{\ell-1}$ is simply $\sigma(s_i^{\ell-1})(1-\sigma(s_i^{\ell-1}))$, which is $O_i^\ell(1-O_i^\ell)$ from the definition above.

For the regression problem, as stated in the assignment text, we know that $O_i^L = s_i^{L-1}$ without the activation function. Then $\frac{\delta O_j^L}{\delta O_i^{L-1}}$ is simply ω_{ij}^{L-1}. However, for the classification problem derivation is a little bit trickier. We know

$$O_i^L = \frac{e^{s_i^{L-1}}}{\sum_{j=0}^{n_L-1}e^{s_j^{L-1}}}$$

Similar to the first expression we obtained, we write the partial derivative as sum of derivatives

$$\frac{\delta O_i^L}{\delta O_j^{L-1}} = \sum_{k=0}^{n_L-1}\frac{\delta O_i^L}{\delta s_k^{L-1}}\frac{\delta s_k^{L-1}}{\delta O_j^{L-1}}$$

Similar to the regression problem, $\frac{\delta s_k^{L-1}}{\delta O_j^{L-1}}$ part is simply ω_{jk}^{L-1}. From [1], $\frac{\delta O_i^L}{\delta s_k^{L-1}}$ becomes $O_i^L(1\{i=k\}-O_k^L)$. $x\{c\}$ is the indicator function that evaluates to x when the condition c is met, otherwise zero (0). I did not want to repeat the same steps; therefore, if you are interested please refer to the blogpost. Notice that $\frac{\delta O_i^L}{\delta s_k^{L-1}}$ evaluates to the derivative of the sigmoid $O_i^L = \sigma(s_i^{L-1})$ when $i=k$.

For the loss functions, we have

$$\begin{aligned}
E(x) &= SE(y, O_i^L) = (y - O_i^L)^2 \\
\frac{\delta E(x)}{\delta O_i^L} &= -2(y - O_i^L) \\
\Delta_i^2 &= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} \frac{\delta E(x)}{\delta O_j^L} \frac{\delta O_j^L}{\delta O_i^{L-1}} \\
&= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} \omega_{ij}^{L-1} \Delta_j^1 \\
\Delta_i^1 &= -2(y - O_i^L)
\end{aligned}$$

for the regression problem, where y is the ground truth as given in the text

$$\begin{aligned}
E(x) &= CE(l, O_i^L) = -\sum_{i=0}^{n_L-1} l_i log(O_i^L) \\
\frac{\delta E(x)}{\delta O_i^L} &= -\frac{l_i}{O_i^L} \\
\Delta_i^2 &= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} \frac{\delta E(x)}{\delta O_j^L} \frac{\delta O_j^L}{\delta O_i^{L-1}} \\
&= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} \frac{\delta E(x)}{\delta O_j^L} (\sum_{k=0}^{n_L-1} \frac{\delta O_j^L}{\delta s_k^{L-1}} \frac{\delta s_k^{L-1}}{\delta O_i^{L-1}}) \\
&= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} -\frac{l_j}{O_j^L} (\sum_{k=0}^{n_L-1} \omega_{ik}^{L-1} O_j^L (1\{j=k\} - O_k^L) \\
&= \sigma_i^{L-2} \sum_{k=0}^{n_L-1} \omega_{ik}^{L-1} (-\sum_{j=0}^{n_L-1} l_j (1\{j=k\} - O_k^L) \\
&= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} \omega_{ij}^{L-1} (-\sum_{k=0}^{n_L-1} l_k (1\{j=k\} - O_j^L) \\
&= \sigma_i^{L-2} \sum_{j=0}^{n_L-1} \omega_{ij}^{L-1} \Delta_j^1 \\
\Delta_i^1 &= -\sum_{j=0}^{n_L-1} l_j (1\{i=j\} - O_i^L)
\end{aligned}$$

for the classification problem, where l is the ground truth as given in the text (assuming *log* is the natural logarithm *ln*) concluding our derivation. To summarize

$$\omega_{ij}^{\ell} = \omega_{ij}^{\ell} - \alpha \frac{\delta E(x)}{\delta \omega_{ij}^{\ell}}$$
$$\frac{\delta E(x)}{\delta \omega_{ij}^{L-n}} = O_i^{L-n} \Delta_j^n$$
$$\Delta_i^{\ell} = \sigma_i^{L-\ell} \sum_{j=0}^{n_{L-\ell+2}-1} \omega_{ij}^{L-\ell+1} \Delta_j^{\ell-1}$$
$$\Delta_i^1 = -2(y - O_i^L) \quad \text{for regression}$$
$$\Delta_i^1 = -\sum_{j=0}^{n_L-1} l_j(1\{i=j\} - O_i^L) \quad \text{for classification}$$

For biases, $\frac{\delta E(x)}{\delta \omega_{ij}^{L-n}} = \frac{\delta E(x)}{\delta b_j^{L-n}}$. Note that the biases are indexed with a single parameter, the index of the node in the next layer that the biased weight is connected to. We have $\frac{\delta O_i^{\ell}}{\delta b_i^{\ell-1}} = \sigma_i^{\ell-1}$. Notice $O_j^{\ell-1}$ term is gone as it is now constant. Thus, $\frac{\delta E(x)}{\delta b_j^{L-n}} = \Delta_j^n$ and other terms are the same as they are not dependent on the bias terms. For batches, we need to divide the $\frac{\delta E(x)}{\delta O_i^L}$ term by number of samples in a batch if we are taking average. If we are taking sum, then no modification needed.

2 Sources

References

[1] Kurbiel, T. (2021, April 22). Derivative of the Softmax function and the categorical cross-entropy loss. Derivative of the Softmax Function and the Categorical Cross-Entropy Loss. Retrieved October 29, 2022, from `https://towardsdatascience.com/derivative-of-the-softmax-function-and-the-categorical-cross-entropy-loss-ffceefc081d1`