Lecture Note - 07: Neural Network

Dihui Lai

March 12, 2020

Contents

1 A Single Neuron

An artificial neuron is the basic computing unit in an artificial neural network. There are different way to define a neuron. The most common one is shown in Figure 1.

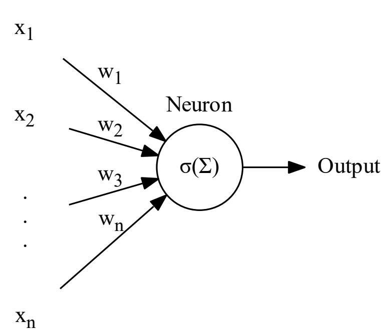

Figure 1: Schema: a single neuron with activation function σ

Input: A neuron receives multiple inputs x_1, x_2, ..., x_n. The signals are summed up after modulated by a set of weights w_1, w_2, ... w_n. Let us denote the weighted sum z

$$z = \sum_{i=1}^{n} w_i x_i \tag{1}$$

Usually a biased term w_0 is added to the summation and we have

$$z = \sum_{i=1}^{n} w_i x_i + w_0 \tag{2}$$

Output: The weighted sum is further transferred via an activation function σ and becomes the final output of the neuron

$$a = \sigma(z) = \sigma(\sum_{i=1}^{n} w_i x_i + w_0) \tag{3}$$

Activation: The activation function can of different types. Below is a list of common activation functions. Almost all activation functions have an S-shape except for the ReLu function.

Name	Definition
Step Function	$\sigma(z) = \begin{cases} 0 & \text{for } z < 0 \\ 1 & \text{for } z \geq 0 \end{cases}$
Logistic or sigmoid	$\sigma(z) = \frac{1}{1+e^{-z}}$
hyperbolic tangent	$\sigma(z) = \frac{(e^z - e^{-z})}{(e^z + e^{-z})}$
ReLU	$\sigma(z) = \begin{cases} 0 & \text{for } z \leq 0 \\ z & \text{for } z > 0 \end{cases}$

2 Neural Network

Each neuron at layer l receives inputs from all neuron from the previous layer $l-1$,

$$z_k^l = \sum_j w_{kj}^{l-1} a_j^{l-1} \tag{4}$$

Here, a_j^{l-1} is the input from jth neuron from $l-1$ layer. The neurons transfer the input signal z_k^l via a transfer function σ and send as input to to the next layer

$$a_k^l = \sigma(z_k^l) \tag{5}$$

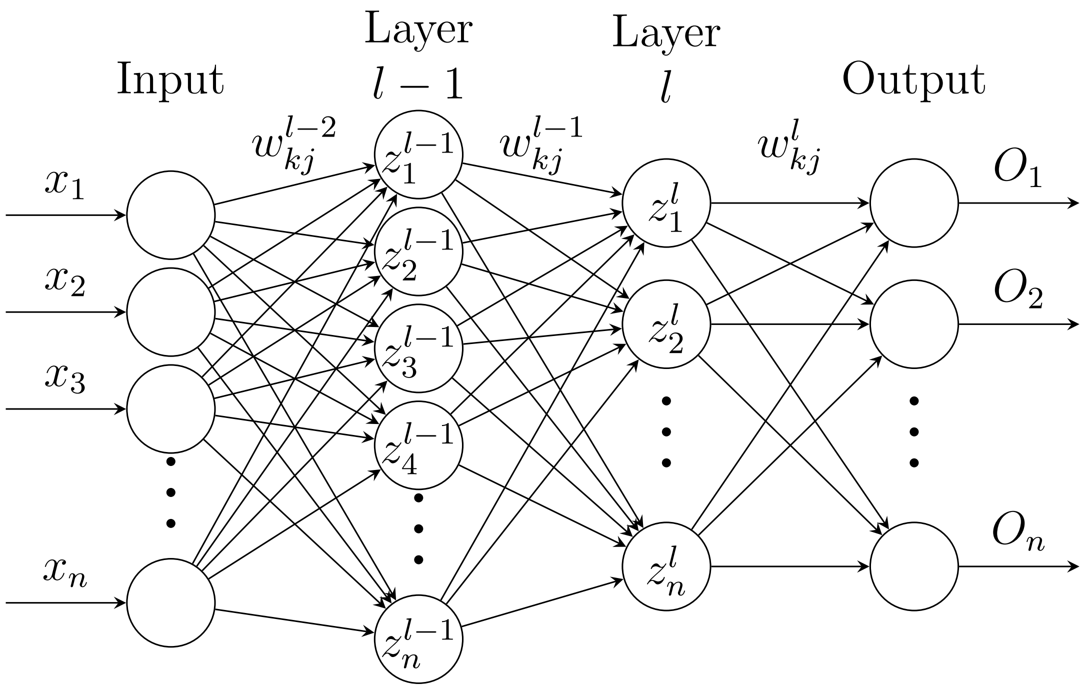

Figure 2: A neural network of multiple layer structures. The hidden layers before layer $l-1$ and after layer l are not shown.

Inserting equation (4) to (5), we have

$$z_k^l = \sum_j w_{kj}^{l-1}\sigma(z_j^{l-1}) = \sum_j w_{kj}^{l-1}a_j^{l-1} \tag{6}$$

3 SGD and Backpropagation

Consider neural network that has N layers, the cost function is dependent on all the zs of neurons in all layers

$$C\left(\vec{z}^N(\vec{z}^{N-1}(...\vec{z}^l(\vec{z}^{l-1})...)...\vec{z}^1)\right) \tag{7}$$

1. Update the weights by changing it along the gradient to reduce the cost function
2. do one data point at a time

$$w_{kj}^L \leftarrow w_{kj}^L - \eta\frac{\partial C}{\partial w_{kj}^L}, L = 1, 2, ...l-1, l, ...N \tag{8}$$

Now let us figure out what $\frac{\partial C}{\partial w_{kj}^L}$ is. Without losing generality, let us consider how the derivative looks like when we consider how signals passes from layer $l-1$ to layer l. By using the chain rule, we have

$$\frac{\partial C(...(z_k^l(w_{kj}^{l-1}, ...)))}{\partial w_{kj}^{l-1}} = \frac{\partial C}{\partial z_k^l}\frac{\partial z_k^l}{\partial w_{kj}^{l-1}} \tag{9}$$

There are two terms that we need to understand from the R.H.S. of equation (8).

- The second term is simply the output from jth neuron in layer $l-1$ since we have from equation (6)

$$\frac{\partial z_k^l}{\partial w_{kj}^{l-1}} = \sigma(z_j^{l-1}) = a_j^{l-1} \tag{10}$$

- The first term after applying two sets of chain rules, we have

$$\begin{aligned}
\delta_k^l = \frac{\partial C}{\partial z_k^l} &= \sum_m \frac{\partial C}{\partial z_m^{l+1}}\frac{\partial z_m^{l+1}}{\partial z_k^l} \text{ (chain rule w.r.t. the composite } z_m^{l+1}(z_m^l)) \\
&= \left(\sum_m \frac{\partial C}{\partial z_m^{l+1}}\frac{\partial z_m^{l+1}}{\partial a_k^l}\right)\frac{da_k^l}{dz_k^l} \text{ (chain rule w.r.t. the composite } z_m^{l+1}(a_m^l)) \\
&= \sum_m \delta_m^{l+1} w_{mk}^l \sigma'(z_k^l) \text{ (noting } \frac{\partial z_k^{l+1}}{\partial a_j^l} = w_{kj}^l \text{ and } \frac{da_k^l}{dz_k^l} = \sigma'(z_k^l))
\end{aligned}$$

  Use notation $\delta_k^l = \frac{\partial C}{\partial z_k^l}$ for short, we have

$$\delta_k^l = \sum_m \delta_m^{l+1} w_{mk}^l \sigma'(z_k^l) \tag{11}$$

Finally, we get our iteration methods for calculating the gradient of the cost function i.e.

$$\begin{cases} \dfrac{\partial C}{\partial w_{kj}^{l-1}} = \delta_k^l a_j^{l-1} \\ \delta_k^l = \displaystyle\sum_m \delta_m^{l+1} w_{mk}^l \sigma'(z_k^l) \end{cases} \tag{12}$$