

## Objectives

Training Networks

Chain Rule

Backpropagation

---

## 1 Training Networks Review

1. Forward feed
2. Calculate cost (Loss, Error)
3. Perform backpropagation
4. Repeat starting at 1 until the cost is reduced to a satisfactory value

## 2 Chain Rule

A core component of backpropagation is the chain rule. The following example looks for  $dc$  with respect to  $dh$ ; however, to arrive at that point we need to chain together a series of derivatives.

$$\frac{dc}{dh} = \frac{dc}{dl} \times \frac{dl}{db} \times \frac{db}{dh}$$

On the right side of the equation

$$\overline{dl} \times \frac{dl}{db}$$

and

$$\overline{db} \times \frac{db}{dh}$$

will ultimately cancel out, leaving

$$\frac{dc}{dl} \times \overline{dh}$$

or

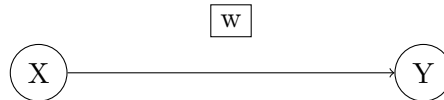
$$\frac{dc}{dh}$$

, which is what we are looking for.

The above is a general example of how the chain rule works. For our specific problems, we are looking for  $dc$  with respect to  $dw$ , where  $dc$  is the rate of change of cost and  $dw$  is the rate of change of weights.

$$\frac{dc}{dw} = ?$$

### 3 Backpropagation



The above example is an extremely simplified network where there is one input value,  $X$ , one weight,  $w$ , the activation function  $a = wx$ , and the final output  $Y$ . In this example:

$$X = 1.5$$

$$Y = 0.5$$

and  $w$  can be found,

$$1.5w = 0.5$$

$$w = \frac{0.5}{1.5}$$

$$w = \frac{1}{3}$$

#### Sigmoid Activation Function

$$\frac{1}{1 + e^{-x}}$$

Continuing with the previous simple example, we can change the activation function to the sigmoid function (above) and obtain an entirely new  $w$ .

$$a = \frac{1}{1 + e^{-z}} = Y$$

$$z = wx$$

$$a = \frac{1}{1 + e^{-1.5w}} = 0.5$$

We eventually end up with

$$e^{-1.5w} = 1$$

This gives us,

$$w = 0$$

Using the sigmoid activation function has caused  $w$  to go to 0.

Returning to the example with the  $a = wx$  activation function, we have the equation,

$$w = w - \eta \nabla C(w)$$

where  $\eta$  is the learning rate,  $C$  is our cost function, and

$$\nabla C = \frac{dc}{dw} = \frac{dc}{da} \times \frac{da}{dw}$$

Our cost function is

$$C = \frac{1}{2}(a - y)^2$$

### Example with Multiple Data Points

Training data = (1.5, 0.5), (6, 2.1) = X

Our cost is the average of the forward feed of all the data points:

$$C = \frac{1}{2}((1.2 - 0.5) + (4.8 - 2.1))^2$$

In general:

$$C = \frac{1}{n}(c_1 + c_2 + \dots + c_n)^2$$

$$C = \frac{1}{n} \sum_{i=1}^n (a - y)^2$$

### Activation Functions

$\sigma$  represents an activation function, for example:

$$\sigma = \frac{1}{1 + e^{-x}}$$

$$\sigma = f(x) = x$$

These functions being sigmoid and identity respectively.

### Hidden Layers

In an example with hidden layers, the process goes as follows:

1.  $w$  values are randomly set
2.  $w_1 x_1 \rightarrow a_1 w_2 \rightarrow a_2$
3.  $C = (a_2 - y_j)^2$
4. Back propagate to adjust the weights.

$$\nabla C = \begin{bmatrix} \frac{dc}{dw_1} \\ \frac{dc}{dw_2} \end{bmatrix}$$

where

$$\frac{dc}{dw_2} = \frac{dc}{da_2} \times \frac{da_2}{dw_2}$$

$$\frac{dc}{dw_1} = \frac{dc}{da_2} \times \frac{da_2}{da_1} \times \frac{da_1}{dw_1}$$

## References