1 Gradient Descent

1.1 Finite difference

using finite difference to get derivative of cost function

$$C'(w) = \lim_{\epsilon \to 0} \frac{C(w+\epsilon) - C(w)}{\epsilon} \tag{1}$$

1.2 Linear Model

$$input \longrightarrow w \longrightarrow output$$

$$y = w * x \tag{2}$$

1.2.1 Cost

$$C(w) = \frac{1}{n} \sum_{i=1}^{n} (x_i w - y_i)^2 \tag{3}$$

$$C'(w) = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i w - y_i)^2 \right)' \tag{4}$$

$$= \frac{1}{n} \left(\sum_{i=1}^{n} (x_i w - y_i)^2 \right)' \tag{5}$$

$$= \frac{1}{n} \left((x_1 w - y_1)^2 + \ldots + (x_n w - y_n)^2 \right)' \tag{6}$$

$$= \frac{1}{n} \left((x_1 w - y_1) + \ldots + (x_n w - y_n) \right)' \tag{7}$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2x_i (x_i w - y_i) \tag{8}$$

$$= \frac{2}{n} \sum_{i=1}^{n} x_i (x_i w - y_i) \tag{9}$$

1.3 One Neuron Model with 2 inputs

$$\begin{tikzcd}
i_1 \arrow[dr, "\text{w1}"] & & \\
& \sigma(i_n * w_n + b) \arrow[r] & output \\
i_2 \arrow[ur, "\text{w2}"] & &
\end{tikzcd}$$

$$
\begin{aligned}
output = y &= \sigma(i_1 * w1 + i_2 * w2 + b) && (10)\\
\sigma(x) &= \frac{1}{1+e^{-x}} && (11)\\
\sigma'(x) &= \sigma(x)(1-\sigma(x)) && (12)
\end{aligned}
$$

1.3.1 Cost

$$
\begin{aligned}
a_i &= \sigma(i_1 * w1 + i_2 * w2 + b) && (13)\\
\partial_{w_1} a_i &= \partial_{w_1}(\sigma(i_1 * w1 + i_2 * w2 + b)) && (14)\\
&= a_i(1-a_i)\partial_{w_1}(i_1 w_1 + i_2 w_2 + b) && (15)\\
&= a_i(1-a_i)i_1 && (16)\\
\partial_{w_2} a_i &= \partial_{w_2}(\sigma(i_1 * w1 + i_2 * w2 + b)) && (17)\\
&= a_i(1-a_i)\partial_{w_2}(i_1 w_1 + i_2 w_2 + b) && (18)\\
&= a_i(1-a_i)i_2 && (19)\\
\partial_b a_i &= \partial_b(\sigma(i_1 w_1 + i_2 w_2 + b)) && (20)\\
&= a_i(1-a_i) && (21)\\
&(z_i : \text{expected output}) && (22)\\
C &= \frac{1}{n}\sum_{i=1}^{n}(a_i - z_i)^2 && (23)\\
\partial_{w_1} C &= \frac{1}{n}\sum_{i=1}^{n}\partial_{w_1}\left((a_i - z_i)^2\right) = && (24)\\
&= \frac{1}{n}\sum_{i=1}^{n}2(a_i - z_i)\partial_{w_1} a_i = && (25)\\
&= \frac{1}{n}\sum_{i=1}^{n}2(a_i - z_i)a_i(1-a_i)i_1 && (26)\\
\partial_{w_2} C &= \frac{1}{n}\sum_{i=1}^{n}2(a_i - z_i)a_i(1-a_i)i_2 && (27)\\
\partial_b C &= \frac{1}{n}\sum_{i=1}^{n}2(a_i - z_i)a_i(1-a_i) && (28)
\end{aligned}
$$

1.4 Two Neurons Model with 1 input

$$
x \xrightarrow{w^{(1)}} \sigma, b^{(1)} \xrightarrow{w^{(2)}} \sigma, b^{(2)} \longrightarrow y
$$

$$a^{(1)} = \sigma(xw^{(1)} + b^{(1)}) \tag{29}$$

$$y = \sigma(a^{(1)}w^{(2)} + b^{(2)}) \tag{30}$$

The superscript in parenthesis denotes the current layer. For example $a_i^{(l)}$ denotes the activation from the l-th layer on i-th sample.

1.4.1 Feed-Forward

$$a_i^{(1)} = \sigma(x_i w^{(1)} + b^{(1)}) \tag{31}$$

$$\partial_{w^{(1)}} a_i^{(1)} = a_i^{(1)}(1 - a_i^{(1)})x_i \tag{32}$$

$$\partial_{b^1} a_i^{(1)} = a_i^{(1)}(1 - a_i^{(1)}) \tag{33}$$

$$a_i^{(2)} = \sigma(a_i^{(1)} w^{(2)} + b^{(2)}) \tag{34}$$

$$\partial_{w^{(2)}} a_i^{(2)} = a_i^{(2)}(1 - a_i^{(2)})a_i^{(1)} \tag{35}$$

$$\partial_{b^{(2)}} a_i^{(2)} = a_i^{(2)}(1 - a_i^{(2)}) \tag{36}$$

$$\partial_{a_i^{(1)}} a_i^{(2)} = a_i^{(2)}(1 - a_i^{(2)})w^{(2)} \tag{37}$$

1.4.2 Back-Propagation

$$C^{(2)} = \frac{1}{n}\sum_{i=1}^{n}(a_i^{(2)} - y_i)^2 \tag{38}$$

$$\partial_{w^{(2)}} C^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\partial_{w^{(2)}}((a_i^{(2)} - y_i)^2) \tag{39}$$

$$= \frac{1}{n}\sum_{i=1}^{n}2(a_i^{(2)} - y_i)\partial_{w^{(2)}} a_i^{(2)} \tag{40}$$

$$= \frac{1}{n}\sum_{i=1}^{n}2(a_i^{(2)} - y_i)a_i^{(2)}(1 - a_i^{(2)})a_i^{(1)} \tag{41}$$

$$\partial_{b^{(2)}} C^{(2)} = \frac{1}{n}\sum_{i=1}^{n}2(a_i^{(2)} - y_i)a_i^{(2)}(1 - a_i^{(2)}) \tag{42}$$

$$\partial_{a_i^{(1)}} C^{(2)} = \frac{1}{n}\sum_{i=1}^{n}2(a_i^{(2)} - y_i)a_i^{(2)}(1 - a_i^{(2)})w^{(2)} \tag{43}$$

$$e_i = a_i^{(1)} - \partial_{a_i^{(1)}} C^{(2)} \tag{44}$$

$$C^{(1)} = \frac{1}{n}\sum_{i=1}^{n}(a_i^{(1)} - e_i)^2 \tag{45}$$

$$\partial_{w^{(1)}} C^{(1)} = \partial_{w^{(1)}}\left(\frac{1}{n}\sum_{i=1}^{n}(a_i^{(1)} - e_i)^2\right) \tag{46}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\partial_{w^{(1)}}\left((a_i^{(1)} - e_i)^2\right) \tag{47}$$

$$= \frac{1}{n}\sum_{i=1}^{n}2(a_i^{(1)} - e_i)\partial_{w^{(1)}} a_i^{(1)} \tag{48}$$

$$= \frac{1}{n}\sum_{i=1}^{n}2(\partial_{a_i^{(1)}} C^{(2)})a_i^{(1)}(1 - a_i^{(1)})x_i \tag{49}$$

$$\partial_{b^1} C^{(1)} = \frac{1}{n}\sum_{i=1}^{n}2(\partial_{a_i^{(1)}} C^{(2)})a_i^{(1)}(1 - a_i^{(1)}) \tag{50}$$

1.5 Arbitrary Neurons Model with 1 input

Let's assume that we have m layers.

1.5.1 Feed-Forward

Let's assume that $a_i^{(0)}$ is x_i.

$$a_i^{(l)} = \sigma(a_i^{(l-1)} w^{(l)} + b^{(l)}) \tag{51}$$

$$\partial_{w^{(l)}} a_i^{(l)} = a_i^{(l)}(1 - a_i^{(l)}) a_i^{(l-1)} \tag{52}$$

$$\partial_{b^{(l)}} a_i^{(l)} = a_i^{(l)}(1 - a_i^{(l)}) \tag{53}$$

$$\partial_{a_i^{(l-1)}} a_i^{(l)} = a_i^{(l)}(1 - a_i^{(l)}) w^{(l)} \tag{54}$$

1.5.2 Back-Propagation

Let's denote $a_i^{(m)} - y_i$ as $\partial_{a_i^{(m)}} C^{(m+1)}$.

$$C^{(l)} = \frac{1}{n} \sum_{i=1}^{n} (\partial_{a_i^{(l)}} C^{(l+1)})^2 \tag{55}$$

$$\partial_{w^{(l)}} C^{(l)} = \frac{1}{n} \sum_{i=1}^{n} 2(\partial_{a_i^{(l)}} C^{(l+1)}) a_i^{(l)} (1 - a_i^{(l)}) a_i^{(l-1)} \tag{56}$$

$$\partial_{b^{(l)}} C^{(l)} = \frac{1}{n} \sum_{i=1}^{n} 2(\partial_{a_i^{(l)}} C^{(l+1)}) a_i^{(l)} (1 - a_i^{(l)}) \tag{57}$$

$$\partial_{a_i^{(l-1)}} C^{(l)} = \frac{1}{n} \sum_{i=1}^{n} 2(\partial_{a_i^{(l)}} C^{(l+1)}) a_i^{(l)} (1 - a_i^{(l)}) w^{(l)} \tag{58}$$