Backdrop 4

Proof by my own

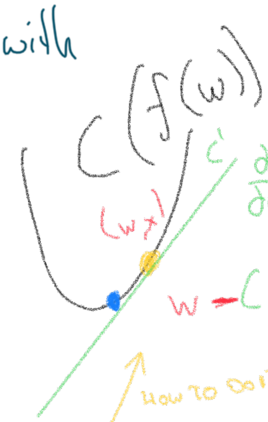
Vector = X, Y, b, a, z

Matrix = W

Scalar = x_m, y_n, w_{mn}, L, M, N

Function = f_{NN}, g, σ, C, len

Def: Neural Network is a composed function with minimizes a loss funtion C



$$\frac{dC}{dw} = \frac{dC}{df}\frac{df}{dw}$$

$$w - C'(w) = w_{update}$$

$$f_{NN}(x) = \underbrace{f^{L-1}(f^{L-1}(\dots f^{1}(f^{0}(x)))}_{\text{LAYERS } L} = \hat{X}$$

HOW TO DO IT

Length of D

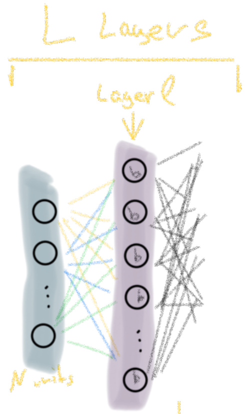
$$\text{Dataset} = D = \{(x^i, y^i)\}^{S}$$

$$\underbrace{\min_{C} C\left(f_{NN}(x^i), y^i\right)}_{\text{WHAT WE WANT}}$$

$a^{-1} = x$ $\quad a^{L-1} = \hat{x}$ $\quad C = (\hat{x}_n - y_n)^2$

$g^{L-1}(z) = z$

$g^{l} = g$



$$f^{l}(a^{l-1}) = a^{l} = \overset{\text{activation}}{g^{l}}\Big(\overbrace{\underset{M\times N}{W^{l}} \cdot \underset{N\times 1}{a^{l-1}} + \underset{M\times 1}{b^{l}}}^{z^{l}}\Big)$$

N = len(l-1) $\quad$ M = len(l)

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_M \end{bmatrix} = z^{l} = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1N} \\ W_{21} & W_{22} & \cdots & W_{2N} \\ \vdots & \vdots & & \vdots \\ W_{M1} & W_{M2} & \cdots & W_{MN} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix}$$

ANY

$$z_m^{l} = \left(W_{m1}^{l} \cdot a_1^{l-1} + W_{m2}^{l} \cdot a_2^{l-1} + \cdots W_{mN}^{l} \cdot a_N^{l-1} + b_m^{l}\right)$$

$$
\frac{\partial C}{\partial a_n^{\ell-1}} = \begin{cases} \text{if } \ell = L-1 \\ \quad \dfrac{\partial C}{\partial \hat{x}} = 2(\hat{x}_n - y_n)^2 \\ \text{else} \\ \quad \displaystyle\sum_{m=0}^{M-1} \frac{\partial C}{\partial a_m^\ell} \frac{\partial a_m^\ell}{\partial z_m^\ell} \frac{\partial z_m^\ell}{\partial a_n^{\ell-1}} = \sum_{m=0}^{M-1} \frac{\partial C}{\partial a_m^\ell} g'^{\ell}(z_m^\ell) \cdot W_{mn}^\ell \end{cases}
$$

HEIGHT OF LAYER ℓ

$$
\frac{\partial C}{\partial W_{mn}^\ell} = \frac{\partial C}{\partial a_m^\ell} \frac{\partial a_m^\ell}{\partial z_m^\ell} \frac{\partial z_m^\ell}{\partial W_{mn}} = \frac{\partial C}{\partial a_m^\ell} g'^{\ell}(z_m^\ell) \cdot a_n^{\ell-1}
$$

$$
\frac{\partial C}{\partial b_m^\ell} = \frac{\partial C}{\partial a_m^\ell} \frac{\partial a_m^\ell}{\partial z_m^\ell} \frac{\partial z_m^\ell}{\partial b_m^\ell} = \frac{\partial C}{\partial a_m^\ell} g'^{\ell}(z_m^\ell)
$$

THESE EQUATIONS DESCRIBE GRADIENT DESCENT

We need to update the weights and biases once we have $\frac{\partial C}{\partial w_{mn}^\ell}$ and $\frac{\partial C}{\partial b_m^\ell}$ but this correction is computed by looking at a single datapoint (x_i, y_i). To get the true correction we must average it over all datapoints.

Error for datapoint d

$$
W_{mn}^\ell \leftarrow W_{mn}^\ell - \rho \cdot \frac{1}{D} \sum_{d=0}^{D-1} \frac{\partial C_d}{\partial w_{mn}^\ell}
$$

Updated

learning rate

$$
b_m^\ell \leftarrow b_m^\ell - \rho \cdot \frac{1}{D} \sum_{d=0}^{D-1} \frac{\partial C_d}{\partial b_m^\ell}
$$