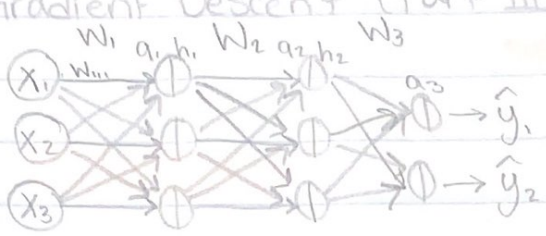
Gradient Descent (Part III)

04/18/2023



$$\frac{dL}{dW_{111}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{da_3} \cdot \frac{da_3}{dh_2} \cdot \frac{dh_2}{da_2} \cdot \frac{da_2}{dh_1} \cdot \frac{dh_1}{da_1} \cdot \frac{da_1}{dW_{111}}$$

1) $\nabla_{\hat{y}} L = -\frac{1}{y_\ell} e(\ell)$

ex.) $e(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$e(1) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

2) $\nabla_{a_3} L = -(e(\ell) - \hat{y})$

3) $\nabla_{h_2} L = W_3^T \nabla_{a_3} L$

$8\times1 \quad 3\times2 \quad 2\times1$

4) $\nabla_{a_2} L = (W_3^T \nabla_{a_3} L) \odot g'(a_2) \quad \{ h_2 = g(a_2) \}$

Side note:

$g(a) = \frac{1}{1+e^{-a}}$ $\quad g'(a) = \frac{-1}{(1+e^{-a})^2}(e^{-a})(-1) = \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}}$

$= \frac{1}{1+e^{-a}}\left(1 - \frac{1}{1+e^{-a}}\right) = g(a)(1-g(a))$

5) $\nabla_{h_1} L = W_2^T \nabla_{a_2} L$

6) $\nabla_{a_1} L = (W_2^T \cdot \nabla_{a_2} L) \odot g'(a_1)$

7) $\nabla_{W_1} L = \nabla_{a_1} L \cdot X^T$
$3\times 3 \quad 3\times 1 \quad 1\times 3$

8) $\nabla_{b_1} L = \nabla_{a_1} L$
$3\times 1 \quad 3\times 1$

9) $\nabla_{W_2} L = \nabla_{a_2} L \cdot h_1^T$

10) $\nabla_{b_2} L = \nabla_{a_2} L$

11) $\nabla_{W_3} L = \nabla_{a_3} L \cdot h_2^T$

12) $\nabla_{b_3} L = \nabla_{a_3} L$

Gradient descent algorithm

$t \leftarrow 0$

max_iter = epochs

$\theta_0 = [W_1, b_1, W_2, b_2, W_3, b_3]$ random initialization

While $t <$ max_iter:

$a_1, h_1, a_2, h_2, a_3, \hat{y} \leftarrow$ forward_prop(θ_t, X)

$\nabla_{W_1} L, \nabla_{b_1} L, \nabla_{W_2} L, \nabla_{b_2} L, \nabla_{W_3} L, \nabla_{b_3} L \leftarrow$ back_prop$(a_1, h_1, a_2, h_2, a_3, \hat{y})$

$\theta_{t+1} \leftarrow \theta_t - \alpha \nabla_{\theta_t} L$ } 6 equations (e.g.) $W_1^{t+1} \leftarrow W_1^t - \alpha \nabla_{W_1}^t L$

*only 1 and 2 change for regression

for regression:

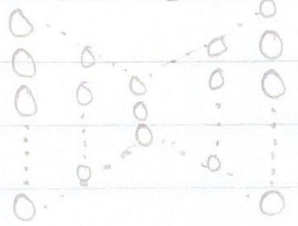
(for 1 sample)

$$L = (\hat{y} - y)^2$$

$$\frac{dL}{d\hat{y}} = 2(\hat{y} - y) \qquad a_3 = \hat{y}$$

$$\nabla_{a_3} L = 2(\hat{y} - y)$$

- autoencoder from scratch



$$\hat{y} = \hat{x}$$

what changes?

$$\hat{y} = \hat{x}; \quad \nabla_{\hat{y}} L; \quad \nabla_{a_3} L$$

$$L = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$\frac{dL}{d\hat{y}_i} = \frac{2}{m} (y_i - \hat{y}_i)(-1)$$

$$\frac{dL}{d\hat{y}} = \begin{bmatrix} \frac{2}{m}(\hat{y}_1 - y_1) \\ \frac{2}{m}(\hat{y}_2 - y_2) \\ \vdots \end{bmatrix} = \frac{2}{m}(\hat{y} - y) = \frac{dL}{da_3} \quad \text{b/c } \hat{y} = a_3$$