20170209 Neural Network

— Previous Notes

- Gradient Descent

$$\theta_i \leftarrow \theta_i - \eta \frac{\partial}{\partial \theta_i} J(\theta)$$

$$\theta \leftarrow \theta - \eta \nabla J(\theta)$$

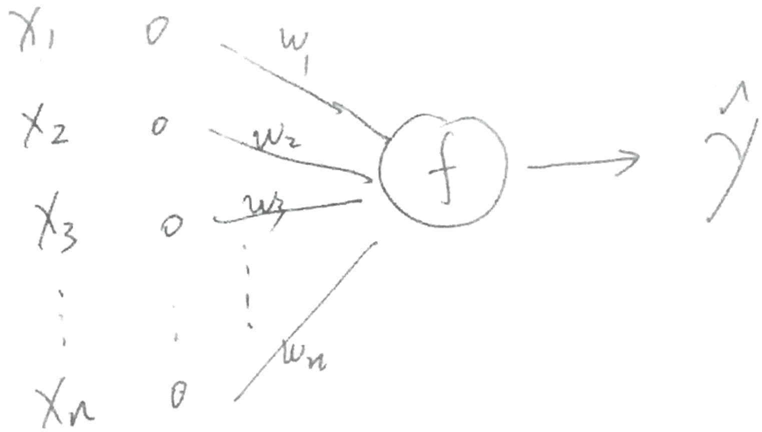
where $\nabla J(\theta) = \left(\frac{\partial}{\partial \theta_1} J(\theta), \frac{\partial}{\partial \theta_2} J(\theta), \cdots \frac{\partial}{\partial \theta_n} J(\theta) \right)$

- Mean Square Error

$$E = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)} \right)^2.$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - f(x^{(i)}) \right)^2.$$

- Neuron



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

$$\hat{y} = f(z)$$

$$\hat{y} = f\left(\sum_{i=1}^{n} x_i w_i\right)$$

when the activation function is sigmoid $\left(x \to \frac{1}{1+e^{-x}}\right)$

$$f'(x) = f(x)\,(1 - f(x))$$

because $f'(x) = \frac{d}{dx}(1+e^{-x})^{-1}$

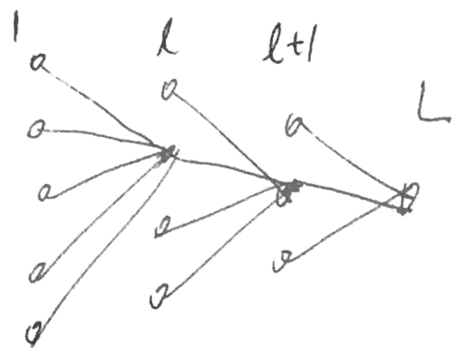
$$= (-1)(1+e^{-x})^{-2}\frac{d}{dx}(1+e^{-x})$$

$$= (1+e^{-x})^{-2}(e^{-x})$$

$$= \frac{e^{-x}}{1+e^{-x}}\,\frac{1}{1+e^{-x}} = \frac{1}{1+e^{-x}}\cdot\left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= f(x)\,(1 - f(x))$$

- Neural Network



- L: number of layers
- n_l: number of nodes in layer l.
- a_i^l: output of ith node in layer l.

$$a^l = \begin{pmatrix} a_1^l \\ a_2^l \\ \vdots \\ a_{n_l}^l \end{pmatrix} \quad \text{shape } n_l \times 1$$

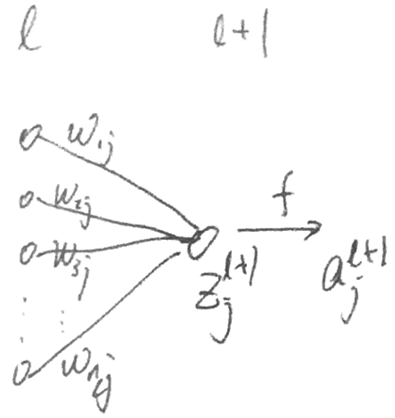
- z_i^l: output of ith node in layer l

$$z^l = \begin{pmatrix} z_1^l \\ z_2^l \\ \vdots \\ z_{n_l}^l \end{pmatrix} \quad \text{shape } n_l \times 1$$

- w_{ji}^l: ~~out~~ weights connecting ith node of layer l and jth node of layer $l+1$

$$W^l = \begin{pmatrix} w_{11}^l & w_{12}^l & \cdots & w_{1n_l}^l \\ w_{21}^l & w_{22}^l & \cdots & w_{2n_l}^l \\ \vdots & & & \vdots \\ w_{n_{l+1}1}^l & w_{n_{l+1}2}^l & \cdots & w_{n_{l+1}n_l}^l \end{pmatrix} \quad \text{shape } n_{l+1} \times n_l.$$

Forward Propagation



Consider the jth node of layer $l+1$

$$z_j^{l+1} = \sum_{i=1}^{n_l} w_{ij}^l \, a_i^l$$

Therefore. $z^{l+1} = w^l \cdot a^l$

$$a_j^{l+1} = f(z_j^{l+1})$$

we donate $f\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}\right) = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}$

Therefore $a^{l+1} = f(z^{l+1})$

we initialize $a^1 = X$

then apply $z^{l+1} = w^l \cdot a^l$

$a^{l+1} = f(z^{l+1})$

finally we have $\hat{y} = a^L$

Backward Propagation.

$$E(w^1, w^2 \dots w^L; x) = \frac{1}{2}\|y - \hat{y}\|^2.$$

we compute $\nabla_{w^1} E, \nabla_{w^2} E, \cdots \nabla_{w^L} E$

i.e. $\frac{\partial}{\partial w_{ji}^l} E$ for every i, j, l.

$$\frac{\partial}{\partial w_{ji}^l} E = \frac{\partial E}{\partial z_j^{l+1}} \frac{\partial z_j^{l+1}}{\partial w_{ji}} \quad \text{(chain rule)}$$

Define $\delta_j^l = \frac{\partial E}{\partial z_j^l}$

$$\delta^l = \begin{pmatrix} \frac{\partial E}{\partial z_1^l} \\ \frac{\partial E}{\partial z_2^l} \\ \vdots \\ \frac{\partial E}{\partial z_{n_l}^l} \end{pmatrix}$$

while $\frac{\partial}{\partial w_{ji}} z_j^{l+1} = \frac{\partial}{\partial w_{ji}} \sum_{k=1}^{n} w_{jk}^l a_k^l = a_i^l$

$$\frac{\partial}{\partial w_{ji}^l} E = \delta_j^{l+1} \cdot a_i^l$$

Then we have in matrix form.

$$\nabla_{w^l} E = \delta^{l+1} \cdot (a^l)^T$$

We only need to compute δ^l now

if $L = l$, i.e. the output layer

$$\delta_j^L = \frac{\partial}{\partial z_j^L} E = \frac{\partial}{\partial z_j^L} \frac{1}{2} \| y - \hat{y} \|^2$$

$$= \frac{\partial}{\partial z_j^L} \frac{1}{2} \sum_{k=1}^{n_L} \left(y_k - f(z_k^L) \right)^2$$

$$= \frac{\partial}{\partial z_j^L} \frac{1}{2} \left(y_j - f(z_j^L) \right)^2$$

$$= -\left(y_j - f(z_j^L) \right) f'(z_j^L)$$

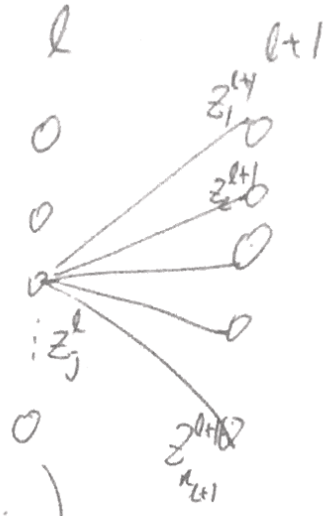
Therefore.

$$\delta^L = -\left(y - f(z^L) \right) * f'(z^L)$$

$$= -\left(y - a^L \right) * f'(z^L)$$

we denote $\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} * \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \\ \vdots \\ a_n b_n \end{pmatrix}$

if $l < L$ i.e. the ~~input~~ hidden layer



$$\delta_j^l = \frac{\partial}{\partial z_j^l} E$$

$$= \sum_{i=1}^{n_{l+1}} \frac{\partial E}{\partial z_i^{l+1}} \frac{\partial z_i^{l+1}}{\partial z_j^l} \quad \text{(multi variable chain rule.)}$$

$$= \sum_{i=1}^{n_{l+1}} \delta_i^{l+1} \frac{\partial}{\partial z_j^l} \sum_{k=1}^{n_l} w_{ik}^l a_k^l$$

$$= \sum_{i=1}^{n_{l+1}} \delta_i^{l+1} \frac{\partial}{\partial z_j^l} w_{ij}^l a_j^l$$

$$= \sum_{i=1}^{n_{l+1}} \delta_i^{l+1} w_{ij}^l \, f'(z_j^l)$$

In matrix form

$$\delta^l = (W^l)^T \delta^{l+1} * f'(z^l)$$

cheat sheet (all column vectors)

<u>Initialization:</u>

$$a^l = \begin{pmatrix} a_1^l \\ a_2^l \\ \vdots \\ a_{n_l}^l \end{pmatrix} \quad z^l = \begin{pmatrix} z_1^l \\ z_2^l \\ \vdots \\ z_{n_l}^l \end{pmatrix} \quad \delta^l = \begin{pmatrix} \delta_1^l \\ \delta_2^l \\ \vdots \\ \delta_{n_l}^l \end{pmatrix}$$

shape $n_l \times 1$

$$W^l = \begin{pmatrix} w_{11}^l & w_{12}^l & \cdots & w_{1n_l}^l \\ w_{21}^l & w_{22}^l & \cdots & w_{2n_l}^l \\ \vdots & \vdots & & \vdots \\ w_{n_{l+1}1}^l & w_{n_{l+1}2}^l & \cdots & w_{n_{l+1}n_l}^l \end{pmatrix}$$

shape $n_{l+1} \times n_l$

<u>Forward</u>

$$a^1 = X \qquad z^{l+1} = W^l a^l \qquad \hat{y} = a^L$$

$$a^{l+1} = f(z^{l+1})$$

<u>Backward</u>

$$\delta^L = -(y - a^L) * f'(z^L)$$

$$\delta^l = (W^l)^T \delta^{l+1} * f'(z^l)$$

$$\nabla_{W^l} E(x) = \delta^{l+1} (a^l)^T$$

$$f'(z^l) = a^l(1 - a^l) \quad \text{if } f \text{ is sigmoid}$$

$$f'(z^l) = 1 \quad \text{if } f(x) = x$$

<u>Update</u>

$$W^l \leftarrow W^l - \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_{W^l} E(x^{(i)})$$

Cheat sheet (all row vectors)

Initialization

$$a^l = (a_1^l \ a_2^l \cdots a_{n_l}^l) \qquad z^l = (z_1^l \ z_2^l \cdots z_{n_l}^l)$$

$$\delta^l = (\delta_1^l, \delta_2^l \cdots \delta_{n_l}^l) \qquad \text{shape } 1 \times n_l.$$

$$w^l = \begin{pmatrix} w_{11}^l & w_{12}^l & & w_{1n_{l+1}}^l \\ w_{21}^l & w_{22}^l & & w_{2n_{l+1}}^l \\ \vdots & & & \\ w_{n_l 1}^l & w_{n_l 2}^l & \cdots & w_{n_l n_{l+1}}^l \end{pmatrix} \qquad \text{shape } n_l \times n_{l+1}$$

Forward

$$a^1 = X \qquad z^{l+1} = a^l w^l \qquad \hat{y} = a^L$$

$$a^{l+1} = f(z^{l+1})$$

Backward

$$\delta^L = -(y - a^L) * f'(z^L)$$

$$\delta^l = \delta^{l+1} (w^l)^T * f'(z^l)$$

$$\nabla_{w^l} E(x) = (a^l)^T \delta^{l+1}$$

$$f'(z^l) = a^l (1 - a^l) \quad \text{if } f \text{ is sigmoid}$$

$$f'(z^l) = 1 \quad \text{if } f(x) = X$$

Update

$$w^l \leftarrow w^l - \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_{w^l} E(X^{(i)})$$

Question:

1. if we add bias term b^l

$$z^{l+1} = w^l a^l + b^l$$

compute $\nabla_{b^l} E$

2. if we add regulization term in the cost function.

$$E = \frac{1}{2N} \sum_{i=1}^{N} \|\hat{y}(x^{(i)}) - y\|^2 + \frac{\lambda}{2} \sum_{l=1}^{L} \sum_{i=1}^{n_l} \sum_{j=1}^{n_{l+1}} (w_{ji}^{l})^2$$

compute $\nabla_{w^l} E$ $\nabla_{b^l} E$

3. compute $f'(z^l)$ when f is ReLU

$$f(x) = \max(0, x)$$