Question.3-16

Linear regression을 위한 dataset이 다음과 같이 주어졌다.

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ \dots \ , (x^{(n)}, y^{(n)})\}$$

이때, dataset은 $y = ax$에서부터 만들어졌다.

따라서 linear regression을 통해 predictor를 학습시킬때,

model은 $\hat{y} = \theta x$, loss는 square error, cost는 MSE를 사용할 수 있다.

θ를 update하기 위해 n개의 data sample를 이용할때, 1번의 iteration에 대해 θ가 dataset을 잘 표현하는

θ로 update되는 과정을 Vector notation을 이용하여 설명하시오.

단, forward/backward propagation을 설명하기 위해 각 연산은 basic building node들을 이용하시오.

① model setting

주어진 상황을 vector의 input, output을 가지는 basic building node들로 표현하면 다음과 같다.

$$\begin{tikzcd}
\theta \arrow[dr] & & \vec{y} \arrow[d] & & \\
\vec{x} \arrow[r] & \boxed{\vec{z}_1 = \theta \cdot \vec{x}} \arrow[r] & \boxed{\vec{z}_2 = \vec{y} - \vec{z}_1} \arrow[r] & \boxed{\vec{\mathcal{L}} = \vec{z}_2 \circ \vec{z}_2} \arrow[r] & \boxed{\mathcal{J} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{(i)}}
\end{tikzcd}$$

↳ $\vec{z}_1 = \vec{\hat{y}}$

이때 $\vec{z}_1$는 model의 prediction인 $\vec{\hat{y}}$가 되고 n개의 data sample이 사용되므로 $\vec{x}, \vec{y}, \vec{z}_1, \vec{z}_2, \vec{\mathcal{L}}$는 모두 n차원 vector가 된다.

② Jacobians

위의 basic building node에서 θ를 학습시키는데 사용되는 Jacobian들을 구하면 다음과 같다.

$$\frac{\partial \mathcal{J}}{\partial \vec{\mathcal{L}}} = \begin{pmatrix} \frac{\partial \mathcal{J}}{\partial \mathcal{L}^{(1)}} & \frac{\partial \mathcal{J}}{\partial \mathcal{L}^{(2)}} & \cdots & \frac{\partial \mathcal{J}}{\partial \mathcal{L}^{(n)}} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \mathcal{L}^{(1)}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{(i)}\right] & \frac{\partial}{\partial \mathcal{L}^{(2)}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{(i)}\right] & \cdots & \frac{\partial}{\partial \mathcal{L}^{(n)}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{(i)}\right] \end{pmatrix} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

$$\frac{\partial \vec{\mathcal{L}}}{\partial \vec{z}_2} = \begin{pmatrix} \frac{\partial \mathcal{L}^{(1)}}{\partial z_2^{(1)}} & \frac{\partial \mathcal{L}^{(1)}}{\partial z_2^{(2)}} & \cdots & \frac{\partial \mathcal{L}^{(1)}}{\partial z_2^{(n)}} \\ \frac{\partial \mathcal{L}^{(2)}}{\partial z_2^{(1)}} & \frac{\partial \mathcal{L}^{(2)}}{\partial z_2^{(2)}} & \cdots & \frac{\partial \mathcal{L}^{(2)}}{\partial z_2^{(n)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}^{(n)}}{\partial z_2^{(1)}} & \frac{\partial \mathcal{L}^{(n)}}{\partial z_2^{(2)}} & \cdots & \frac{\partial \mathcal{L}^{(n)}}{\partial z_2^{(n)}} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial z_2^{(1)}}[z_2^{(1)} \cdot z_2^{(1)}] & \frac{\partial}{\partial z_2^{(2)}}[z_2^{(1)} \cdot z_2^{(1)}] & \cdots & \frac{\partial}{\partial z_2^{(n)}}[z_2^{(1)} \cdot z_2^{(1)}] \\ \frac{\partial}{\partial z_2^{(1)}}[z_2^{(1)} \cdot z_2^{(1)}] & \frac{\partial}{\partial z_2^{(2)}}[z_2^{(1)} \cdot z_2^{(1)}] & \cdots & \frac{\partial}{\partial z_2^{(n)}}[z_2^{(1)} \cdot z_2^{(1)}] \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial z_2^{(1)}}[z_2^{(1)} \cdot z_2^{(1)}] & \frac{\partial}{\partial z_2^{(2)}}[z_2^{(1)} \cdot z_2^{(1)}] & \cdots & \frac{\partial}{\partial z_2^{(n)}}[z_2^{(1)} \cdot z_2^{(1)}] \end{pmatrix}$$

$$= \begin{pmatrix} 2z_2^{(1)} & 0 & \cdots & 0 \\ 0 & 2z_2^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2z_2^{(n)} \end{pmatrix}$$

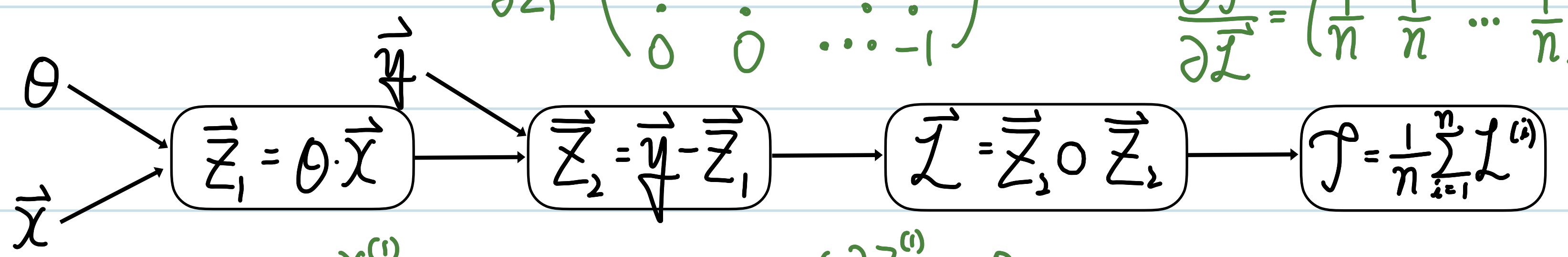
$$\frac{\partial \vec{z}_2}{\partial \vec{z}_1} = \begin{pmatrix} \frac{\partial z_2^{(1)}}{\partial z_1^{(1)}} & \frac{\partial z_2^{(1)}}{\partial z_1^{(2)}} & \cdots & \frac{\partial z_2^{(1)}}{\partial z_1^{(n)}} \\ \frac{\partial z_2^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_2^{(2)}}{\partial z_1^{(2)}} & \cdots & \frac{\partial z_2^{(2)}}{\partial z_1^{(n)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_2^{(n)}}{\partial z_1^{(1)}} & \frac{\partial z_2^{(n)}}{\partial z_1^{(2)}} & \cdots & \frac{\partial z_2^{(n)}}{\partial z_1^{(n)}} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial z_1^{(1)}}[y^{(1)} - z_1^{(1)}] & \frac{\partial}{\partial z_1^{(2)}}[y^{(1)} - z_1^{(1)}] & \cdots & \frac{\partial}{\partial z_1^{(n)}}[y^{(1)} - z_1^{(1)}] \\ \frac{\partial}{\partial z_1^{(1)}}[y^{(1)} - z_1^{(1)}] & \frac{\partial}{\partial z_1^{(2)}}[y^{(1)} - z_1^{(1)}] & \cdots & \frac{\partial}{\partial z_1^{(n)}}[y^{(1)} - z_1^{(1)}] \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial z_1^{(1)}}[y^{(1)} - z_1^{(1)}] & \frac{\partial}{\partial z_1^{(2)}}[y^{(1)} - z_1^{(1)}] & \cdots & \frac{\partial}{\partial z_1^{(n)}}[y^{(1)} - z_1^{(1)}] \end{pmatrix}$$

$$= \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{pmatrix}$$

$$\frac{\partial \vec{z}_1}{\partial \theta} = \begin{pmatrix} \frac{\partial z_1^{(1)}}{\partial \theta} \\ \frac{\partial z_1^{(2)}}{\partial \theta} \\ \vdots \\ \frac{\partial z_1^{(n)}}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \theta}[\theta \cdot x^{(1)}] \\ \frac{\partial}{\partial \theta}[\theta \cdot x^{(1)}] \\ \vdots \\ \frac{\partial}{\partial \theta}[\theta \cdot x^{(1)}] \end{pmatrix} = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{pmatrix}$$

따라서 위의 Jacobian들을 node들 위에 정리하면 다음과 같다.

$$\frac{\partial \vec{z}_2}{\partial \vec{z}_1} = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{pmatrix} \qquad \frac{\partial \mathcal{J}}{\partial \vec{\mathcal{L}}} = \left(\frac{1}{n} \ \ \frac{1}{n} \ \cdots \ \frac{1}{n}\right)$$



$$\frac{\partial \vec{z}_1}{\partial \theta} = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{pmatrix} \qquad \frac{\partial \vec{\mathcal{L}}}{\partial \vec{z}_2} = \begin{pmatrix} 2z_2^{(1)} & 0 & \cdots & 0 \\ 0 & 2z_2^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2z_2^{(n)} \end{pmatrix}$$

③ Backpropagation

②에서 구한 Jacobian들과 chain rule을 이용하여 $\frac{\partial \mathcal{J}}{\partial \theta}$를 구하면 다음과 같다.

$$\frac{\partial \mathcal{J}}{\partial \vec{z}_2} = \frac{\partial \mathcal{J}}{\partial \vec{\mathcal{L}}} \cdot \frac{\partial \vec{\mathcal{L}}}{\partial \vec{z}_2} = \left(\frac{1}{n} \ \ \frac{1}{n} \ \cdots \ \frac{1}{n}\right) \begin{pmatrix} 2z_2^{(1)} & 0 & \cdots & 0 \\ 0 & 2z_2^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2z_2^{(n)} \end{pmatrix} = \left(\frac{1}{n} \cdot 2z_2^{(1)} \ \ \frac{1}{n} \cdot 2z_2^{(2)} \ \cdots \ \frac{1}{n} \cdot 2z_2^{(n)}\right)$$

$$= \left(\frac{1}{n} \cdot 2(y^{(1)} - z_1^{(1)}) \ \ \frac{1}{n} \cdot 2(y^{(2)} - z_1^{(2)}) \ \cdots \ \frac{1}{n} \cdot 2(y^{(n)} - z_1^{(n)})\right) = \left(\frac{1}{n} \cdot 2(y^{(1)} - \hat{y}^{(1)}) \ \ \frac{1}{n} \cdot 2(y^{(2)} - \hat{y}^{(2)}) \ \cdots \ \frac{1}{n} \cdot 2(y^{(n)} - \hat{y}^{(n)})\right)$$

$$\frac{\partial \mathcal{J}}{\partial \vec{z}_1} = \frac{\partial \mathcal{J}}{\partial \vec{z}_2} \cdot \frac{\partial \vec{z}_2}{\partial \vec{z}_1} = \left(\frac{1}{n} \cdot 2(y^{(1)} - \hat{y}^{(1)}) \ \ \frac{1}{n} \cdot 2(y^{(2)} - \hat{y}^{(2)}) \ \cdots \ \frac{1}{n} \cdot 2(y^{(n)} - \hat{y}^{(n)})\right) \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{pmatrix}$$

$$= \left(\frac{1}{n} \cdot (-2(y^{(1)} - \hat{y}^{(1)})) \ \ \frac{1}{n} \cdot (-2(y^{(2)} - \hat{y}^{(2)})) \ \cdots \ \frac{1}{n} \cdot (-2(y^{(n)} - \hat{y}^{(n)}))\right)$$

$$\frac{\partial \mathcal{J}}{\partial \theta} = \frac{\partial \mathcal{J}}{\partial \vec{z}_1} \cdot \frac{\partial \vec{z}_1}{\partial \theta} = \left(\frac{1}{n} \cdot (-2(y^{(1)} - \hat{y}^{(1)})) \ \ \frac{1}{n} \cdot (-2(y^{(2)} - \hat{y}^{(2)})) \ \cdots \ \frac{1}{n} \cdot (-2(y^{(n)} - \hat{y}^{(n)}))\right) \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{pmatrix}$$

$$= \frac{1}{n} \cdot (-2x^{(1)}(y^{(1)} - \hat{y}^{(1)})) + \frac{1}{n} \cdot (-2x^{(2)}(y^{(2)} - \hat{y}^{(2)})) + \cdots + \frac{1}{n}(-2x^{(n)}(y^{(n)} - \hat{y}^{(n)}))$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[-2x^{(i)}(y^{(i)} - \hat{y}^{(i)})\right]$$

④ gradient descent method

먼저 J를 이용하여 θ를 update시키는 식은

$$\theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$$

이다. 이 식에 ③에서 구한 $\frac{\partial J}{\partial \theta}$를 대입하면

$$\theta := \theta - \alpha \frac{1}{n} \sum_{i=1}^{n} \left[-2x^{(i)} (y^{(i)} - \hat{y}^{(i)}) \right]$$

$$= \theta + \frac{\alpha}{n} \sum_{i=1}^{n} \left[2x^{(i)} (y^{(i)} - \hat{y}^{(i)}) \right]$$

이 된다. 그리고 이는

$$\theta := \theta - \alpha \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mathcal{L}^{(i)}}{\partial \theta} \right]$$

이므로 Question.3-1의 결과와 동일함을 알 수 있다.