Question.3-09

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 error를 사용할 수 있다.

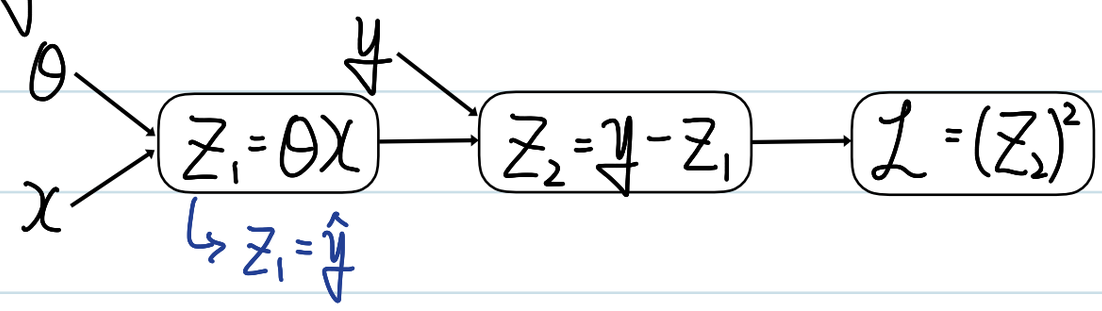
θ를 update하기 위해 하나의 data sample만 이용할때, 1번의 iteration에 대해 θ가 dataset을 잘 표현하는

θ로 update되는 과정을 설명하시오.

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

① model/loss setting

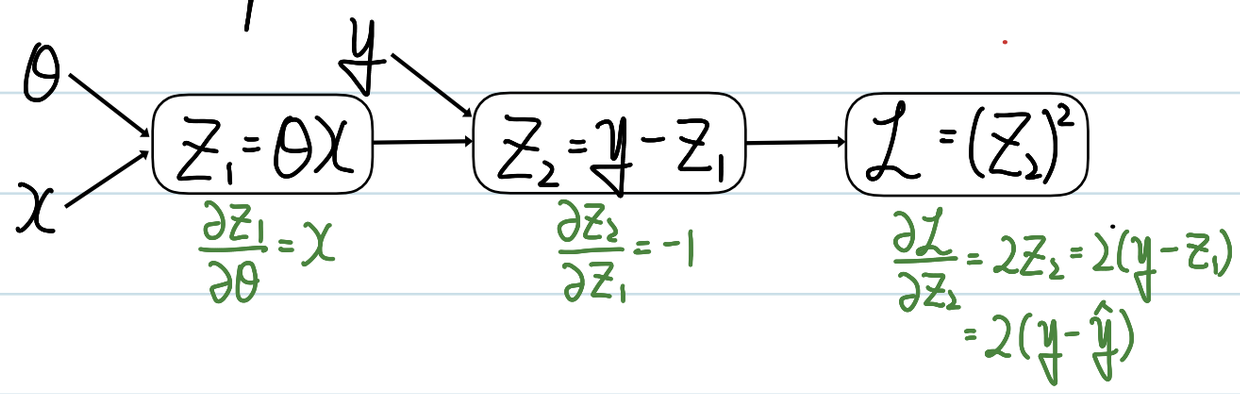
주어진 상황을 basic building node를 이용하여 나타내면 다음과 같다.



이때 Z_1은 θx이므로 $\hat{y}$가 된다.

② partial derivatives

위의 그림에서 θ를 update하기 위해 필요한 partial derivative를 나타내면 다음과 같다.



③ backpropagation

②에서 구한 partial derivative와 chain rule을 이용하여 $\frac{\partial L}{\partial \theta}$를 구하는 과정은 다음과 같다.

$$\frac{\partial L}{\partial Z_1} = \frac{\partial L}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial Z_1} = 2(y-\hat{y}) \cdot (-1) = -2(y-\hat{y})$$

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial Z_1} \cdot \frac{\partial Z_1}{\partial \theta} = -2(y-\hat{y}) \cdot x = -2x(y-\hat{y})$$

④ gradient descent method

$\frac{\partial L}{\partial \theta}$를 이용하여 θ를 update시키는 식은

$$\theta := \theta - \alpha \frac{\partial L}{\partial \theta}$$

식과 ③에서 구한 $\frac{\partial L}{\partial \theta}$를 대입하면 다음과 같다.

$$\theta := \theta + 2\alpha x(y-\hat{y})$$

④ parameter learning

③에서 구한 θ의 update 식

$$\theta := \theta + 2\alpha x(y-\hat{y})$$

을 통해 θ가 dataset을 잘 표현하는 θ로 학습되는 것을 보이기 위해 $x>0$, $x<0$와 $y-\hat{y}>0$, $y-\hat{y}$, 즉 $y>\hat{y}$, $y<\hat{y}$로 나누어서 생각해보면 다음과 같다.

i) $y>\hat{y}$	$x>0$	$x<0$	ii) $y<\hat{y}$	$x>0$	$x<0$
$2\alpha x(y-\hat{y})$	+	−	$2\alpha x(y-\hat{y})$	−	+
$\theta := \theta+2\alpha x(y-\hat{y})$	↑	↓	$\theta := \theta+2\alpha x(y-\hat{y})$	↓	↑
$y=\theta\cdot x$	↑	↑	$y=\theta\cdot x$	↓	↓
$\lvert y-\hat{y}\rvert$	↓	↓	$\lvert y-\hat{y}\rvert$	↓	↓

즉 모든 경우에 대해 θ는 update가 되고 난 뒤에 같은 (x,y)가 입력되었을 때 x가 충분히 작다면 y와 $\hat{y}$의 차이가 줄어들도록 학습된다. 따라서 충분한 iteration이 지나면 predictor는 y와 동일한 $\hat{y}$를 예측하게 된다.