

Question.3-01

Dataset이 다음과 같이 단 하나의 data sample로만 이루어졌다고 하자.

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

위의 dataset은 $y = 2x$ 에서부터 만들어졌기 때문에, model을 $\hat{y} = \theta_0 + \theta_1 x$ 로 설정할 때 다음 문제들에 답하시오.

1) θ 가 1, 1.5, 2일 때의 square error를 이용한 loss를 각각 구하고, θ 가 2에 가까워질 때 loss의 변화를 비교하시오.

(square error는 y 와 \hat{y} 를 이용하여 $(y - \hat{y})^2$ 로 구한다.)

2) 임의의 θ 에 대한 loss를 algebraic equation으로 표현하고 그래프를 그리시오.

$$L_{(1)} = (y_{(1)} - \hat{y}_{(1)})^2, \quad J = \frac{1}{N} \sum_i L_{(i)}, \quad \hat{y}_{(1)} = \theta_0 + \theta_1 x, \\ = \vec{\theta}^\top \vec{x}$$

1) $L_{(1)}$ when $\theta = 1$,

$$y = 2, \quad \hat{y} = 1 \cdot x = 1 \quad \therefore L_{(1)} = (2 - 1)^2 = 1$$

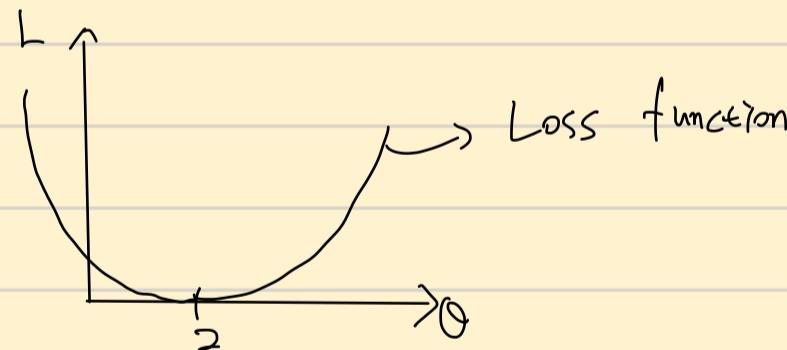
$L_{(1)}$ when $\theta = 1.5$

$$y = 2, \quad \hat{y} = 1.5 x = 1.5 \quad \therefore L_{(1)} = (2 - 1.5)^2 = 0.25$$

when $\theta = 2$

$$y = 2, \quad \hat{y} = 2x = 2 \quad \therefore L_{(1)} = (2 - 2)^2 = 0$$

2) $L_{(1)} = (y_{(1)} - \hat{y}_{(1)})^2 = (2 - \theta)^2 = \theta^2 - 4\theta + 4$



Question.3-02

Dataset이 다음과 같이 주어졌다고 하자.

$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\} = \{(1,2), (2,4), (3,6)\}$$

위의 dataset에 대해 다음의 물음에 답하시오.

1) model을 $\hat{y} = x$ 로 설정한다면, 각 data sample들에 대한 square loss를 구하고 서로 비교하시오.

2) square error loss를 임의의 data point $(x^{(i)}, y^{(i)})$ 에 대한 algebraic equation으로 표현하고,

$(x^{(i)}, y^{(i)})$ 에 대한 square error loss를 최소로 만드는 θ 를 구하시오.

단, θ 는 변수로 사용하여 model은 $\hat{y} = \theta x$ 로 설정하시오.

3) 2)에서 구한 algebraic equation이 θ 에 대한 몇차인지 구하고, convexity를 말하시오.

$$1) L_{(1)} = (2-1)^2, L_{(2)} = (4-2)^2, L_{(3)} = (6-3)^2 \\ = 1 \quad = 4 \quad = 9$$

$$\therefore L_{(3)} > L_{(2)} > L_{(1)}$$

$$2) L_{(i)} = (y^{(i)} - \theta x^{(i)})^2 = y^{(i)2} - 2\theta x^{(i)} y^{(i)} + \theta^2 x^{(i)2}$$

$$\arg \min_{\theta} L_{(i)} \Leftrightarrow \frac{\partial L_{(i)}}{\partial \theta} = 2\theta x^{(i)2} - 2x^{(i)}y^{(i)} = 0 \\ 2x^{(i)}(\theta x^{(i)} - y^{(i)}) = 0$$

$$\theta^* = \frac{y^{(i)}}{x^{(i)}}$$

즉, θ 가 $\frac{y^{(i)}}{x^{(i)}}$ 일 때, $L_{(i)}$ 는 최소가 된다.

3) algebraic equation은 θ 에 대한 2차이고,
 θ^2 의 계수가 양수 ($\because (x^{(i)})^2$) 이기 때문에,
 $L_{(i)}$ 는 θ 에 대해 convex한 그래프를 갖는다.

Question.3-03

문제에서의 상황이 Question.3-02와 동일할 때, 다음의 물음에 답하시오.

- 1) Question.3-02의 1) 상황에서 1)의 결과를 이용하여 MSE cost를 구하시오.
- 2) MSE cost를 algebraic equation으로 표현하고, MSE cost를 최소로 만드는 θ 를 구하시오.
- 3) 2)에서 구한 algebraic equation이 θ 에 대한 몇차인지 구하고, convexity를 말하시오.

$$1) Cost_{MSE} = \frac{1}{3}(1+4+9) = \frac{14}{3}$$

$$2) Cost_{MSE} = \frac{1}{3} \sum_{i=1}^3 (y^{(i)} - \theta x^{(i)})^2$$

$$= \frac{1}{3} \sum_{i=1}^3 (y^{(i)} - 2\theta x^{(i)} + \theta^2 x^{(i)2})$$

$$= \frac{1}{3} (y_1^2 - 2\theta x_1 y_1 + \theta^2 x_1^2 + y_2^2 - 2\theta x_2 y_2 + \theta^2 x_2^2 + y_3^2 - 2\theta x_3 y_3 + \theta^2 x_3^2)$$

$$= \frac{1}{3} (\theta^2(x_1^2 + x_2^2 + x_3^2) - 2\theta(y_1 x_1 + y_2 x_2 + y_3 x_3) + (y_1^2 + y_2^2 + y_3^2))$$

$$* Cost_{MSE} = \frac{1}{3} \sum_{i=1}^3 (y^{(i)} - \theta x^{(i)})^2 = \frac{1}{3} \sum_{i=1}^3 L(i)$$

$\therefore \theta$ 가 각 데이터 틀인트 i 에 대해, $\frac{y^{(i)}}{x^{(i)}}$ 일 때,
 $L(i)$ 가 최소가 되고, $Cost \leq$ 최소가 된다.

$$3) Cost_{MSE} = \frac{1}{3} (\theta^2(x_1^2 + x_2^2 + x_3^2) - 2\theta(y_1 x_1 + y_2 x_2 + y_3 x_3) + (y_1^2 + y_2^2 + y_3^2))$$

Cost function은 θ 에 대해 2차식이고,

θ^2 의 계수가 양수이므로 Convex한 그래프를 그린다.

Question.3-04

다음과 같이 하나의 data sample만 가지고 있는 Dataset이 주어졌다.

$$\mathcal{D} = \{(1, 2)\}$$

이때 다음 질문들에 답하시오. 단 prediction model은 $\hat{y} = \theta x$ 를 사용한다.

- 1) Square error를 loss로 사용할 때 loss에 대한 식을 구하고, $\frac{\partial \mathcal{L}(\theta)}{\partial \theta}$ 의 식을 구하시오.

그리고 gradient descent method를 이용하여 θ 를 update시키는 식을 구하시오.

- 2) initial θ 를 1로, learning rate을 0.1로 설정했을 때 3 iteration 동안 update되는 θ 들을 구하시오.

그리고 target θ 에 가까워지는지 확인하시오.

- 3) loss의 식을 graph로 표현하고 2)에서 구한 θ 가 update되는 위치들을 이 graph 위에 나타내시오.

$$\mathcal{D} = \{(1, 2)\}$$

$$1) \text{ loss} = (y_i - \theta x_i)^2 = \theta^2 x^2 - 2\theta x y + y^2 \Rightarrow \text{loss} = \theta^2 - 4\theta + 4$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 2x^2\theta - 2xy = 2x(x\theta - y) = \nabla_{\theta} \mathcal{L} = (\theta - 2)^2$$

$$\theta_2 = \theta_1 - \alpha \cdot \nabla_{\theta} \mathcal{L} = \theta_1 - \alpha (2x(x\theta - y))$$

$\mathcal{D} = \{(1, 2)\}$ 은 대체로,

$$\theta_2 = \theta_1 - \alpha (2(0 - 2))$$

$$2) \quad \theta_{\text{initial}} = 1, \quad L.R = 0.1, \quad 3 \text{ iteration.}$$

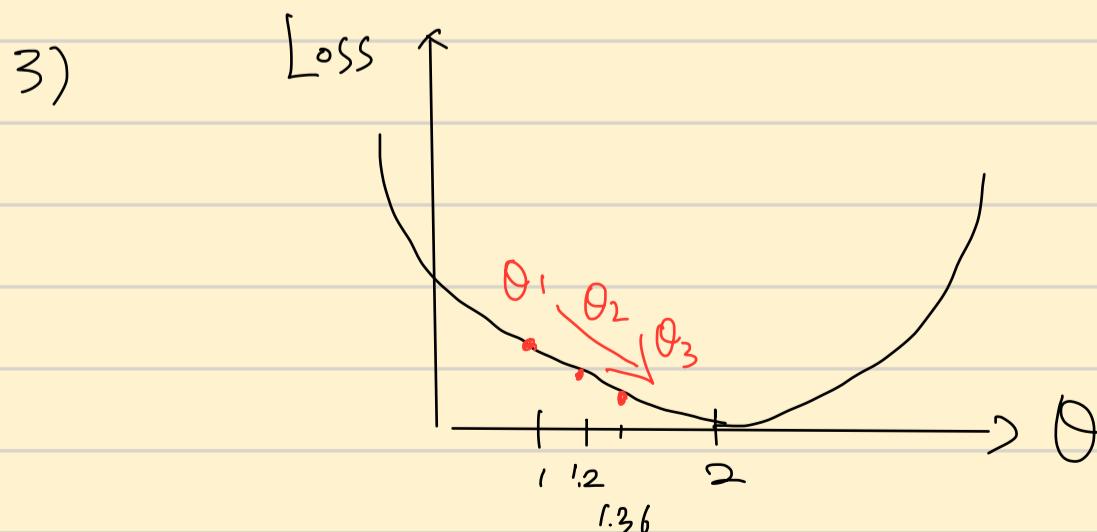
$$\theta_1 = 1, \quad \theta_2 = \theta_1 - 0.1 \cdot 2 \cdot (1 - 2) = 1 - 0.2 \cdot (-1) = 1 + 0.2 = 1.2$$

$$\theta_3 = \theta_2 - 0.1 \cdot 2 \cdot (1.2 - 2) = 1.2 - 0.2 \cdot (-0.8) = 1.2 + 0.16 = 1.36$$

$$\theta_1 \rightarrow \theta_2 \rightarrow \theta_3 \quad \theta_{\text{target}} = 2$$

1 1.2 1.36

θ_{target} 에 점점 가까워진다.



$$\theta_n = \theta_{n-1} - \alpha \nabla f(\theta_{n-1})$$

"Gradient descent"

Question.3-05

다음과 같이 하나의 data sample만 가지고 있는 Dataset이 주어졌다.

$$\mathcal{D} = \{(1, 2)\}$$

이때 다음 질문들에 답하시오.

단 Question.3-04와 마찬가지로 prediction model은 $\hat{y} = \theta_0 + \theta_1 x$ 를 사용하고, initial θ 는 1로 설정한다.

1) learning rate이 0.8일 때, 3 iteration 동안 update되는 θ 들을 구하시오.

2) learning rate이 1.1일 때, 3 iteration 동안 update되는 θ 들을 구하시오.

$$\theta_{n+1} = \theta_n - \alpha (\text{J}(\theta_n)) \quad \leftarrow \text{gradient descent update.}$$

1) LR = 0.8

$$\theta_1 = 1, \quad \theta_2 = 1 - 0.8 \cdot 2 \cdot (-1) = 1 + 1.6 = 2.6$$

$$\theta_3 = 2.6 - 0.8 \cdot 2 \cdot 0.6 = 2.6 - 0.96 = 1.64$$

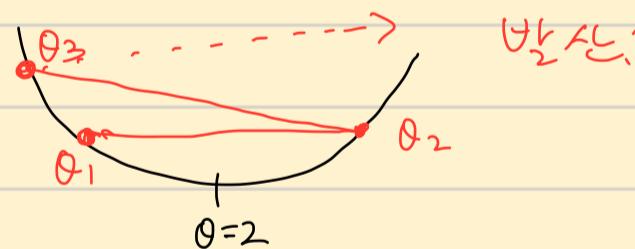
$$\begin{matrix} \theta_1 & \rightarrow & \theta_2 & \rightarrow & \theta_3 \\ / & & 2.6 & & 1.64 \end{matrix}$$



2) LR = 1.1

$$\theta_1 = 1, \quad \theta_2 = 1 - 1.1 \cdot 2 \cdot (-1) = 1 + 2.2 = 3.2$$

$$\theta_3 = 3.2 - 1.1 \cdot 2 \cdot 1.2 = 3.2 - 2.64 = 0.56$$



Question.3-06

다음과 같이 Dataset이 주어졌다.

$$\mathcal{D} = \{(1,2), (3,6), (4,8)\}$$

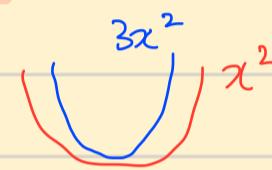
이때 다음 질문들에 답하시오.

단 Question.3-04, Question.3-05와 마찬가지로 prediction model은 $\hat{y} = \theta_0 + \theta_1 x$ 를 사용하고, initial θ 는 1로 설정한다.

- 1) learning rate이 0.1일 때, 각각 θ 의 loss에 대한 update equation을 구하시오.
- 2) 1)에 구한 update equation을 이용하여 3번의 iteration에 대해 각각 θ 의 변화를 구하고, Question.3-05 3)의 관점에서 data sample에 따른 학습의 불안정성을 설명하시오.
- 3) learning rate이 0.01일 때, 각각 θ 의 loss에 대한 update equation을 구하고, 3번의 iteration에 대해 각각 θ 의 변화를 구하시오. 추가로 2)에서의 학습의 불안정성이 해결되는지 설명하시오.

$L(c_i)$ 는 데이터 포인트의 input 값의 절대값 크기가 클수록 더

가파르게 블록해진다. ($\therefore L(c_i) = (x^{(i)})^2 \theta^2 + \dots$)



\therefore 아주 작은 α (learning rate) 라고,

$|input|$ 이 큰 경우,

학습이 불안정 혹은 발산할 수 있다.



$|input|$ 이 클수록, Learning Rate 를 더 작게

가져가야 안정적인 학습이 가능!

Question.3-07

다음과 같이 Dataset이 주어졌다.

$$\mathcal{D} = \{(1,2), (3,6), (4,8)\}$$

이때 다음 질문들에 답하시오.

단 Question.3-04, Question.3-05와 마찬가지로 prediction model은 $\hat{y} = \theta x$ 를 사용하고, initial θ 는 1로 설정한다.

1) 각 3개의 data point들의 loss를 이용하여 cost J를 구하고, $\frac{\partial J(\theta)}{\partial \theta}$ 를 구하시오.

2) $\frac{\partial J(\theta)}{\partial \theta}$ 와 gradient descent method를 이용한 θ 의 update equation을 구하고,

$\alpha = 0.1$ 일 때 4 iterations에 대한 θ 의 변화를 구하시오.

Cost function = $\frac{1}{N} \sum_n L_{(i)} = \text{Loss function 평균}$

Question.3-08

Question.3-08의 질문들은 Question.3-06과 Question.3-07을 바탕으로 해결하시오.

1) $\frac{\partial J(\theta)}{\partial \theta}$ 와 $\frac{\partial \mathcal{L}^{(1)}(\theta)}{\partial \theta}, \frac{\partial \mathcal{L}^{(2)}(\theta)}{\partial \theta}, \frac{\partial \mathcal{L}^{(3)}(\theta)}{\partial \theta}$ 의 관계를 설명하고, 이들을 이용한 gradient descent methods

$$\theta := \theta - \alpha \frac{\partial \mathcal{L}^{(i)}(\theta)}{\partial \theta} \quad \theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$$

의 관계를 설명하시오.

2) Question.3-06에서 (4,8)에 대한 loss를 이용하여 θ 를 학습시키면 학습이 되지 않았다.

하지만 Question.3-07에서는 cost를 이용하여 θ 를 학습시킬 때 (4,8)이 사용되었는데 올바르게 학습이 되었다.

두 과정의 차이점을 설명하시오.

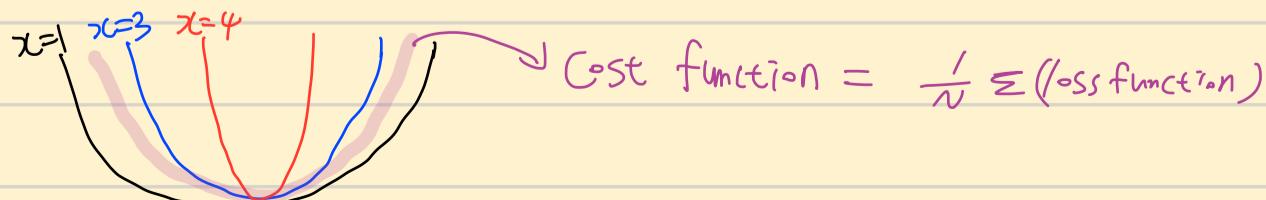
Cost function을 통한 gradient method는 θ 를 Loss를 통해 update 하는 것들의 평균 정도를 update 한다.



~~Cost~~ Cost를 이용해 θ update 하면 Outlier로 인한 학습의 불안정성을 막을 수 있다!

(Loss function을 이용하면 input이 커질수록
 $\frac{\partial^2 \mathcal{L}}{\partial \theta^2} - 2\theta x_i -$ 같은 learning rate와 불안정 학습가능!)

Cost function은 각 데이터 포인트의 영향을 중화시켜줌



$$\text{Cost function} = \frac{1}{N} \sum (\text{loss function})$$

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 학습시킬 때, model은 $\hat{y} = \theta x$, loss는 square error를 사용할 수 있다.

θ 를 update하기 위해 하나의 data sample만 이용할 때, 1번의 iteration에 대해 θ 가 dataset을 잘 표현하는

θ 로 update되는 과정을 설명하시오.

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

$$\begin{aligned} \frac{\partial z_1}{\partial \theta} &= x_1 & \frac{\partial z_2}{\partial z_1} &= -1 & \frac{\partial L}{\partial z_2} &= 2z_2 \\ \theta &\rightarrow \theta x_1 & \rightarrow \alpha x_1 - \theta x_1 & \rightarrow (\alpha x_1 - \theta x_1)^2 & = L \\ x_1 &= z_1 & = z_2 = y - z_1 & & \\ \text{Mul-node} & & \text{minus-node} & & \text{square-node} \end{aligned}$$

$$\begin{aligned} \therefore \frac{\partial L}{\partial \theta} &= \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial \theta} = 2z_2 \cdot (-1) \cdot x_1 = -2x_1(y - \theta x_1) \\ &= -2x_1(\alpha x_1 - \theta x_1) \\ &= \text{gradient } \nabla_{\theta} L \end{aligned}$$

$$\therefore \text{Updated } \theta = \theta - \alpha \cdot \nabla_{\theta} L = \theta - \alpha \cdot -2x_1(\alpha x_1 - \theta x_1)$$

$$= \theta + 2\alpha x_1(\alpha x_1 - \theta x_1)$$

Question.3-10

Linear regression을 위한 dataset이 다음과 같이 주어졌다.

$$D = \{(1,3), (3,9), (2,6), (-1, -3)\}$$

learning rate을 0.01로, initial θ 는 1로 설정하고, Question.3-09와 같은 방법으로 학습을 진행할 때 다음 질문들에 답하시오.

1) 1 epoch 동안 각 data sample들에 대해 loss와 θ 가 update되는 절댓값을 구하시오.

그리고 loss의 감소 graph의 fluctuation이 생기는 이유를 설명하시오.

2) 1)의 결과를 통하여 $x^{(i)}$ 의 크기가 γ 배 되었을 때, loss와 θ 가 update되는 양은 몇 배가 되는지 구하시오.

1) 1-epoch, Losser θ 가족

$$\begin{aligned} \frac{\partial L}{\partial \theta_0} &= x & \frac{\partial L}{\partial \theta_1} &= -1 & \frac{\partial L}{\partial \theta_2} &= 2x \\ \theta_0 & \rightarrow \theta_1 & x - \theta_0 & \rightarrow & (y - \theta_0)^2 &= L \\ x & \rightarrow z_1 & z_2 & \rightarrow & & \\ \theta_{initial} &= 1 & -1 & 4 & & \\ x &= 1 & \rightarrow & 3-1 & \rightarrow & \frac{2^2=4}{L} \quad , \quad \therefore L_{(1)} = 4 \end{aligned}$$

$$\begin{aligned} L &= (y - \theta x)^2 \\ \text{gradient} &= -2x(y - \theta x) \\ \theta_{i+1} &= \theta_i - \alpha \cdot \nabla L \\ &= \theta_i + 2\alpha x(y - \theta x) \end{aligned}$$

$$\begin{aligned} \theta_1 &= \theta_{initial} - \alpha \nabla L = 1 - 0.01 \cdot (-4) = 1.04 \\ \therefore L_{(1)} &= 4, \quad \theta_1 = 1.04 \end{aligned}$$

$$\begin{aligned} L_{(1)} &= (y_i - \hat{y}_i)^2 \\ &= (y_i - \theta x_i)^2 = \theta^2 x_i^2 - 2\theta x_i y_i + y_i^2 \quad \theta_{(1)} = \theta_{i-1} + 2\alpha x (y - \theta x) \\ &= \theta_{i-1} + 2\alpha x^2 (\alpha - \theta) \end{aligned}$$

$$\begin{aligned} L_{(1)} &= (3-1)^2 = 4 & \theta_1 &= 1 + 2(0.01)1(2) = 1.04 \\ L_{(2)} &= (9-3.04)^2 = 34.57 & \theta_2 &= 1.04 + 2(0.01)3(5.88) = 1.39 \\ L_{(3)} &= (6-1.39)^2 = 10.37 & \theta_3 &= 1.39 + 2(0.01)2(3.22) = 1.52 \\ L_{(4)} &= (-3-1.52)^2 = 2.19 & \theta_4 &= 1.52 + 2(0.01)(-1)(-1.48) = 1.55 \end{aligned}$$

2) $x_i \rightarrow r \cdot x_i$ 의 losser θ 의 update 규칙.
 $(x_i, \alpha x_i) \rightarrow (r x_i, \alpha r x_i)$

$$\begin{aligned} L_i &= (y - y_i)^2 = (r x_i - \theta r x_i)^2 = r^2 (\alpha x_i - \theta x_i)^2 \\ \Delta \theta_{(1)} &= 2\alpha r^2 (\alpha - \theta) \end{aligned}$$

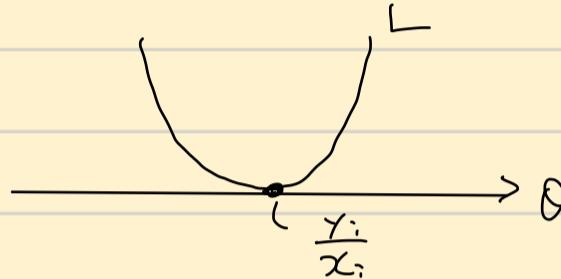
~~input이 r 의 증가하면, Losser $\Delta \theta$ 는 r^2 의 증가한다!~~

Question.3-11

Linear regression을 통해 predictor를 학습시킬 때, iteration이 지날 때마다 loss의 감소량과 θ 가 update되는 양이 줄어드는 이유를 설명하시오.

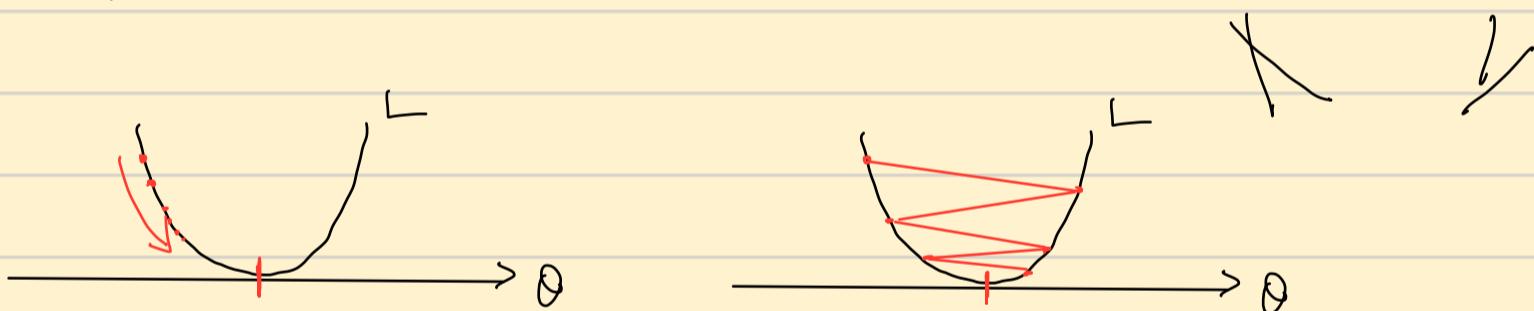
먼저 $\hat{y} = \theta x$ 일 때, Loss function은 다음과 같다.

$$Loss_{(i)} = (y_i - \hat{y}_i)^2 = (y_{(i)} - \theta x_{(i)})^2$$



그리고 learning rate가 충분히 작다면,

$\theta := \theta - \alpha \frac{\partial L}{\partial \theta}$ 를 $\theta_{initial}$ 이 어디이든 θ 는 optimal point로 이동한다. θ 의 update는 아래 2가지의 모듈 중 하나가 될 것이다. (learning rate가 충분히 작을 때)



∴ 양대 모두 $|\frac{\partial L}{\partial \theta}|$ 이 줄어든다.

$\therefore \Delta \theta$ 도 줄어들게 되어 Loss 값도 측면도 줄어들게 된다.

$$(\because Loss = (y_i - \underline{\theta} x_i)^2)$$

Question.3-12

대부분의 dataset에는 noise가 섞여 있다. $y = 2x$ 에서 이상적으로 만들어진 Dataset이 다음과 같을 때

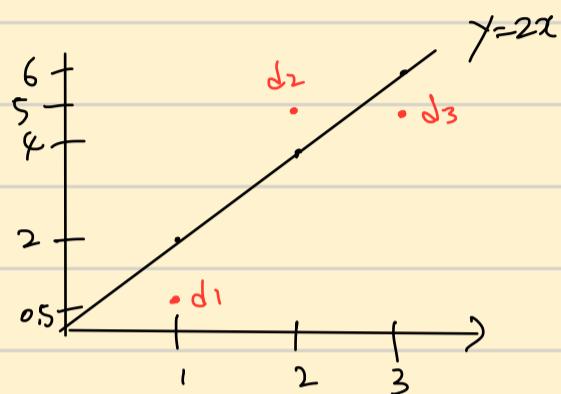
$$D = \{(1, 2), (2, 4), (3, 6)\}$$

noise에 의해 실제로 수집한 dataset이 다음과 같이 왜곡되었다고 가정하자.

$$D = \{(1, 0.5), (2, 5), (3, 5)\}$$

θ 가 2일었을 때, 즉 predictor가 target function 일 때 각 data sample에 의해 θ 가 update되는 값들을 구하고,

noise가 학습에 미치는 영향을 분석하시오.



$$\theta_{\text{init}} = 2$$

$$\theta_i = \theta_{i-1} + 2\alpha x_{i-1} (y_{i-1} - \theta_{i-1} x_{i-1})$$

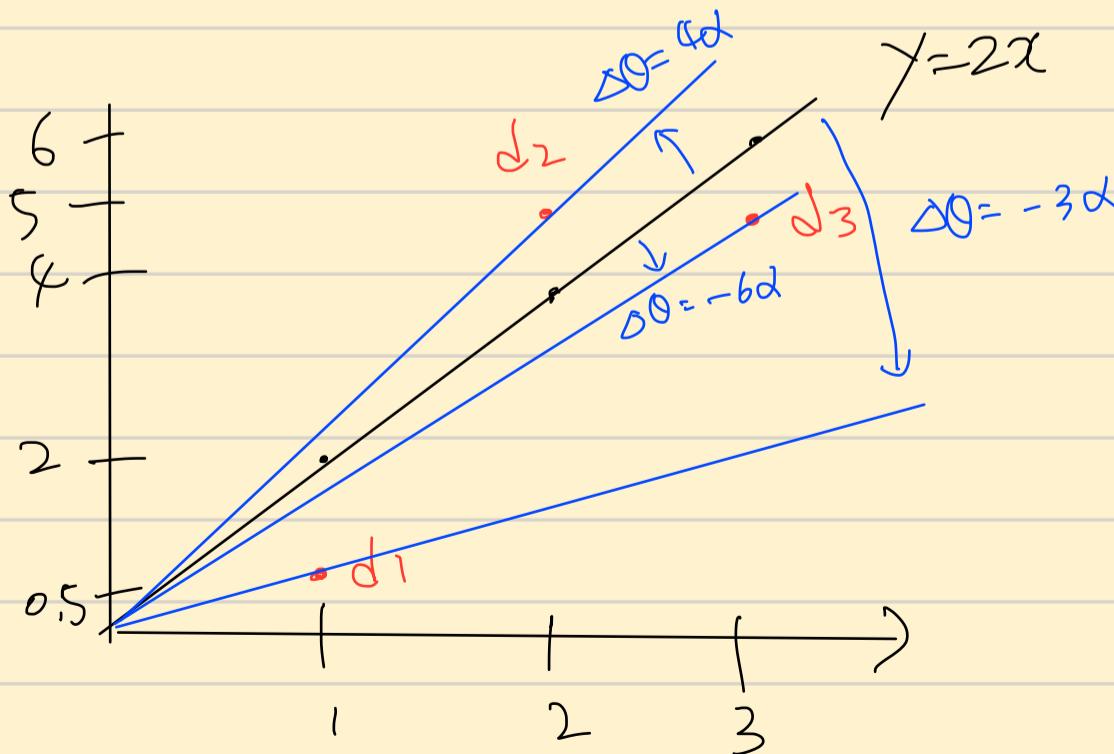
$$\text{Loss} = (y_i - \theta x_i)^2$$

Noise 있는 데이터에 대해 기존 Predictor ($\theta=2$)는 예상치는 성능을 보인다.
하지만 Noise로 인해 발생한 왜곡 때문에 Loss가 발생, θ 의 업데이트가
잘못된 상황이다. 각 noise의 흐트를 보기 위해, update는 시키지 않고
Loss와 θ 의 변화량을 살펴볼 것이다.

$$\text{Data point: } (1, 0.5) \quad L = (0.5 - 2)^2 = 2.25 \quad \Delta\theta = 2\alpha \cdot 1 (0.5 - 2) = -3\alpha$$

$$\text{Data point: } (2, 5) \quad L = (5 - 4)^2 = 1 \quad \Delta\theta = 2\alpha \cdot 2 (5 - 4) = 4\alpha$$

$$\text{Data point: } (3, 5) \quad L = (5 - 6)^2 = 1 \quad \Delta\theta = 2\alpha \cdot 3 (5 - 6) = -6\alpha$$



Noise
Data가 \downarrow 로 왜곡되면
Predictor도 \uparrow 로 왜곡되고,
Data가 \downarrow 로 왜곡되면,
Predictor도 \downarrow 로 왜곡된다.

Question.3-13

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 학습시킬 때,

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

θ 를 update하기 위해 2개의 data sample를 이용할 때, 1번의 iteration에 대해 θ 가 dataset을 잘 표현하는

θ 로 update되는 과정을 설명하시오.

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

본 모델인 $y = ax$ 를 학습하기 위해 각 구어진 데이터는 토대로 3가지로 학습해간다!

Data를 predictor에 넣어 발생하는 Loss를 줄이기 위해 gradient descent method를 활용해 θ (parameter)를 update 해 나간다.

$$\begin{aligned} \theta &= x^{(1)} & \frac{\partial L}{\partial \theta} &= -1 & \frac{\partial L}{\partial \theta} &= 2x_1 \\ x^{(1)} &\rightarrow \theta x^{(1)} & \leftarrow y^{(1)} - \theta x^{(1)} &\leftarrow (y^{(1)} - \theta x^{(1)})^2 & \frac{\partial J}{\partial \theta} &= \frac{1}{2} \\ \theta &= x^{(2)} & \frac{\partial L}{\partial \theta} &= 1 & J &= \frac{1}{2} \sum (y_i - \theta x_i)^2 = \frac{1}{2} \sum L_{(i)} \\ x^{(2)} &\rightarrow \theta x^{(2)} & \leftarrow y^{(2)} - \theta x^{(2)} &\leftarrow (y^{(2)} - \theta x^{(2)})^2 & \frac{\partial J}{\partial \theta} &= \frac{1}{2} \\ * \frac{\partial J}{\partial \theta} &= \frac{\partial J}{\partial L_{(1)}} \frac{\partial L_{(1)}}{\partial \theta} + \frac{\partial J}{\partial L_{(2)}} \frac{\partial L_{(2)}}{\partial \theta} = \frac{1}{2} \cdot 2x_1(-1) \cdot x^{(1)} = \frac{1}{2} \cdot -2x^{(1)} \cdot (y^{(1)} - \theta x^{(1)}) & L &= (y - \theta x)^2, \frac{\partial L}{\partial \theta} = -2x(y - \theta x) \\ &+ \frac{\partial J}{\partial L_{(2)}} \frac{\partial L_{(2)}}{\partial \theta} = \frac{1}{2} \cdot 2x_2(1) \cdot x^{(2)} = \frac{1}{2} \cdot 2x^{(2)}(-1) \cdot (y^{(2)} - \theta x^{(2)}) & J &= \frac{1}{n} \sum L_{(i)}, \frac{\partial J}{\partial \theta} = \frac{1}{n} \sum -2x_{(i)}(y_{(i)} - \theta x_{(i)}) \\ &= -x^{(1)}(y^{(1)} - \theta x^{(1)}) - x^{(2)}(y^{(2)} - \theta x^{(2)}) & \therefore \frac{\partial J}{\partial \theta} &= \frac{1}{n} (-2x_1(y_1 - \theta x_1) - 2x_2(y_2 - \theta x_2)) \end{aligned}$$

위와 같이 forward propagation을 통해 실제 발생한 Loss를 Cost를 계산한다 (SE, MSE)

이를 최소화하기 위해, θ 를 update 해 Loss, cost가 빠른 작은 θ model을 찾아 나간다. 이때, gradient descent 하는 방법을 활용한다. 과정은 아래와 같다.

(2개의 data sample, 1번의 iteration)
 "mini batch size = 2"

$$\theta_2 = \theta_1 + \frac{1}{n} \sum_{i=1}^2 2 \alpha x_{(i)} (y_{(i)} - \theta x_{(i)})$$

Question.3-14

Linear regression을 위한 dataset이 다음과 같이 주어졌고, 이 dataset을 이용하여 predictor를 학습시키려고 한다.

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

이때 loss를 사용하여 θ 를 update시키면 $(x^{(i)}, y^{(i)})$ 의 크기에 따라 θ 를 θ^* 에서 멀어지게 하는 $(x^{(i)}, y^{(i)})$ 가 존재할 수 있다.

cost를 사용하게 되면 이 문제에 대한 위험성을 낮출 수 있는데, 그 이유를 설명하시오.

Loss 를 사용해 gradient descent 방법으로 θ 를 업데이트 하면,

$$\theta_{n+1} = \theta_n - \alpha \cdot (-2) x_n (y_n - \theta x_n) \text{ 이다.}$$

Cost 를 활용할 경우,

$$\theta_{n+1} = \theta_n - \alpha \cdot \frac{1}{n} \sum_{i=1}^n (-2) x_i (y_i - \theta x_i) \text{ 이다.}$$

$$* \frac{\partial J}{\partial \theta} = \frac{1}{n} \in L_C(\cdot)$$

Cost를 활용할 때, $\Delta \theta$ 가 기존의 Loss 등의 평균으로 정의된다.

따라서, 개별 샘플 $(x^{(i)}, y^{(i)})$ 에 의해 $\Delta \theta$ 의 크기의 변동이 커질 가능성이 줄어든다.

Question.3-15

Linear regression을 위한 dataset이 다음과 같이 주어졌다.

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

Cost function update

Question.3-13와 같은 방법으로 학습을 진행할 때, n개의 data sample을 이용하여 θ 를 update한다면 loss의 감소에서 fluctuation은 어떻게 변하는지 설명하시오.

$$\theta := \theta + \frac{1}{n} \sum_{i=1}^n 2\alpha x_{(i)} (y_{(i)} - \theta x_{(i)})$$

" θ is defined ..."



$\frac{\partial L(i)}{\partial \theta}$ 는 평균적으로 반영!

($n = \text{mini batch size}$)

\therefore fluctuation은 감소한다!

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 학습시킬 때,

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

θ 를 update하기 위해 n 개의 data sample를 이용할 때, 1번의 iteration에 대해 θ 가 dataset을 잘 표현하는

θ 로 update되는 과정을 Vector notation을 이용하여 설명하시오.

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

$$\begin{aligned} \theta &\rightarrow \theta x_i \rightarrow y - \theta x_i \rightarrow (y - \theta x_i)^2 \rightarrow \frac{\partial L^{(i)}}{\partial \theta} = \frac{1}{n} \\ x_i &\rightarrow z_1^{(i)} \frac{\partial z_1^{(i)}}{\partial z_2^{(i)}} = -1 \quad z_2^{(i)} \frac{\partial z_2^{(i)}}{\partial L^{(i)}} = 2z_2^{(i)} \end{aligned}$$

$$\begin{aligned} \theta &\rightarrow \theta x_2 \rightarrow y - \theta x_2 \rightarrow (y - \theta x_2)^2 \rightarrow J = \frac{1}{n} \sum_{i=1}^n L^{(i)} \\ x_2 &\rightarrow z_1^{(2)} \quad z_2^{(2)} \end{aligned}$$

$$\vdots$$

$$\begin{aligned} \theta &\rightarrow \theta x_n \rightarrow y - \theta x_n \rightarrow (y - \theta x_n)^2 \rightarrow * \frac{\partial L^{(i)}}{\partial \theta} = -2x_i(y_i - \theta x_i) \\ x_n &\rightarrow z_1^{(n)} \quad z_2^{(n)} \end{aligned}$$

$$* \frac{\partial J}{\partial \theta} = -\frac{2}{n} x_i(y_i - \theta x_i)$$

↓ Vector Notation

$$\begin{aligned} \theta &\rightarrow \frac{\partial z_1}{\partial \theta} = \vec{x} \quad \vec{y} \rightarrow \frac{\partial z_2}{\partial z_1} = -I_{(n \times n)} \quad \frac{\partial L}{\partial z_2} = \begin{bmatrix} 2z_2^{(1)} \\ \vdots \\ 2z_2^{(n)} \end{bmatrix} \\ \vec{x} &\rightarrow \theta \vec{x} = \vec{z}_1 \quad \vec{y} - \theta \vec{x} \rightarrow \vec{y} - \vec{z}_2 = \vec{z}_2 \quad \text{square-node} \end{aligned}$$

$$(\vec{y} - \theta \vec{x})^2 \rightarrow J = \frac{1}{n} \sum_{i=1}^n L^{(i)}$$

$$* \theta \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \theta \vec{x} \quad * \vec{y} - \theta \vec{x} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \theta \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} z_2^{(1)} \\ \vdots \\ z_2^{(n)} \end{bmatrix} \quad * L = \begin{bmatrix} (y_1 - z_2^{(1)})^2 \\ \vdots \\ (y_n - z_2^{(n)})^2 \end{bmatrix}$$

$$* \frac{\partial \vec{z}_2}{\partial \vec{z}_1} = \left[\frac{\partial z_2^{(1)}}{\partial z_1^{(1)}}, \dots, \frac{\partial z_2^{(n)}}{\partial z_1^{(n)}} \right] = -I_{(n \times n)}$$

$$* \frac{\partial \vec{L}}{\partial \vec{z}_2} = \left[\frac{\partial L^{(1)}}{\partial z_2^{(1)}}, \dots, \frac{\partial L^{(n)}}{\partial z_2^{(n)}} \right] = \begin{bmatrix} 2z_2^{(1)} \\ \vdots \\ 2z_2^{(n)} \end{bmatrix}$$

$$\frac{\partial J}{\partial \theta} = \frac{1}{n} \sum L^{(i)} = -\frac{2}{n} \sum_{i=1}^n x_i (y_i - \theta x_i) = \frac{1}{n} \frac{\partial L}{\partial \vec{z}_2} \frac{\partial \vec{z}_2}{\partial \theta} = -\frac{1}{n} \begin{bmatrix} 2z_2^{(1)} \\ \vdots \\ 2z_2^{(n)} \end{bmatrix} \begin{bmatrix} 1 & \dots & 0 \\ 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$* L^{(i)} = -2x_i (y_i - \theta x_i)$$

$$\theta = \theta + \frac{2}{n} \sum_{i=1}^n x_i (y_i - \theta x_i) \quad (\theta = \theta - \alpha \frac{\partial J}{\partial \theta} = \theta - \alpha \frac{1}{n} \sum_{i=1}^n -2x_i (y_i - \theta x_i))$$

Question.3-17

Question.3-18에서는 Jacobian들의 matrix multiplication을 이용하여 $y = \theta x$ 를 학습시키는 과정을 증명하였다.

그리고 이 과정에서 $\frac{\partial \vec{z}_2}{\partial z_1}, \frac{\partial \vec{z}}{\partial z_2}$ 는 모두 diagonal matrix인 것을 확인할 수 있었다.

이 관점에서 다음의 질문들에 답하시오.

1) n-dimentional row vector \vec{a} 와 nxn diagonal matrix B의 vector-matrix multiplication 결과에서 0이 아닌 값들을 Hadamard product를 이용하여 표현하시오.

2) n-dimentional vector \vec{a}, \vec{b} 의 dot product를 Hadamard product를 이용하여 표현하시오.

3) Question.3-16의 과정을 1)의 내용을 이용하여 Hadamard product으로 표현하시오.

$$1) \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}_{(n \times 1)} = \vec{a}, \begin{bmatrix} b_{11} & 0 & \cdots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & b_{nn} \end{bmatrix}_{(n \times n)} = B \Rightarrow \text{B의 diagonal entries} \geq \text{Column Vector } \vec{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

\Downarrow

$$\vec{a} \circ \vec{b} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \circ \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = a_1 b_1 + \cdots + a_n b_n$$

$$2) \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}_{(n \times 1)} = \vec{a}, \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}_{(n \times 1)} = \vec{b}, a_1 b_1 + \cdots + a_n b_n = \vec{a}^T \cdot \vec{b}$$

$$3) \theta \rightarrow \vec{z}_1 = \theta \vec{x} \rightarrow \vec{z}_2 = (\vec{y} - \vec{z}_1) \rightarrow L = \vec{z}_2 \circ \vec{z}_2 \rightarrow J = \frac{1}{n} \sum_{i=1}^n L^{(i)}$$

$$\frac{\partial \vec{z}_1}{\partial \theta} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \frac{\partial \vec{z}_2}{\partial \vec{x}} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}, \frac{\partial L}{\partial \vec{z}_2} = \begin{bmatrix} 2z_2^{(1)} & \cdots & 2z_2^{(n)} \end{bmatrix}, \frac{\partial J}{\partial \vec{x}} = \left[\frac{1}{n}, \dots, \frac{1}{n} \right]$$

$$\vec{F} = \begin{bmatrix} 1 \\ \vdots \\ -1 \end{bmatrix}_{(n \times 1)}, \beta = \begin{bmatrix} 2z_2^{(1)} \\ \vdots \\ 2z_2^{(n)} \end{bmatrix}_{(n \times 1)}$$

Hadamard 표현
역대 범주!
(n × 1) 골로!
"범주"는 예상치!

$$\alpha = \begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}_{(n \times 1)}$$

$$\therefore \frac{\partial J}{\partial \theta} = \frac{\partial J}{\partial L} \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial \theta} \frac{\partial z_2}{\partial \theta}$$

Scalar

$$= \text{sum} \left[(\alpha \circ \beta) \circ \vec{F} \circ \frac{\partial \vec{z}_1}{\partial \theta} \right] = -\frac{2}{n} \sum_i x_i (y_i - \theta x_i)$$

$$\theta := \theta - \alpha \frac{\partial J}{\partial \theta} = \theta - \alpha \left[\text{sum} \left[(\alpha \circ \beta) \circ \vec{F} \circ \frac{\partial \vec{z}_1}{\partial \theta} \right] \right]$$