Question.3-06

다음과 같이 Dataset이 주어졌다.

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

이때 다음 질문들에 답하시오.

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

1) learning rate이 0.1일때, 각각 θ의 loss에 대한 update equation을 구하시오.

2) 1)에 구한 update equation을 이용하여 3번의 iteration에 대해 각각 θ의 변화를 구하고, Question.3-05 3)의 관점에서 data sample에 따른 학습의 불안정성을 설명하시오.

3) learning rate이 0.01일 때, 각각 θ의 loss에 대한 update equation을 구하고, 3번의 iteration에 대해 각각 θ의 변화를 구하시오. 추가로 2)에서의 학습의 불안정성이 해결되는지 설명하시오.

먼저 일반적인 data sample (x,y)와 임의의 α, θ에 대한 parameter update equation을 구해보면 다음과 같다.

$$\theta := \theta - \alpha \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \theta - \alpha \frac{\partial}{\partial \theta}[(y-\theta x)^2] = \theta + 2\alpha x(y-\theta x)$$

1) 위의 식에서 $\alpha = 0.1$, $(x^{(i)}, y^{(i)})$를 적용하면 각각의 parameter update equation들은 다음과 같다.

$$\theta := \theta + 0.2 \cdot x^{(1)}(y^{(1)} - \theta \cdot x^{(1)}) = \theta + 0.2(2-\theta)$$

$$\theta := \theta + 0.2 \cdot x^{(2)}(y^{(2)} - \theta \cdot x^{(2)}) = \theta + 0.6(6-3\theta) = \theta + 1.8(2-\theta)$$

$$\theta := \theta + 0.2 \cdot x^{(3)}(y^{(3)} - \theta \cdot x^{(3)}) = \theta + 0.8(8-4\theta) = \theta + 3.2(2-\theta)$$

가 된다. 즉, $x^{(i)}$가 커질때 parameter update 양은 α^2배도 커진다. 즉 learning rate의 α^2배만큼 커진 효과를 갖는다.

2) 각각의 data sample들에 대해 θ의 변화를 구하면 다음과 같다.

	(1,2)	(3,6)	(4,8)
1st iteration:	$\theta := 1 + 0.2(2-1) = 1.2$	$\theta := 1 + 1.8(2-1) = 2.80$	$\theta := 1 + 3.2(2-1) = 4.2$
2nd iteration:	$\theta := 1.2 + 0.2(2-1.2) = 1.36$	$\theta := 2.80 + 1.8(2-2.80) = 1.36$	$\theta := 4.2 + 3.2(2-4.2) = -2.84$
3rd iteration:	$\theta := 1.36 + 0.2(2-1.36) = 1.49$	$\theta := 1.36 + 1.8(2-1.36) = 2.51$	$\theta := -2.84 + 3.2(2+2.84) = 12.65$

위의 결과에서 (1,2)는 안정적인 학습을 보여주지만, (3,6)에서부터 α가 조금만 더 커지면 발산할 위험성이 커진다. 그리고 (4,8)에선 실제로 (1,2)와 같은 α로 학습되었지만 발산하는 것을 볼 수 있다.

즉, 위의 경우에서 알 수 있듯이 learning rate이 충분히 작더라도 절대값이 큰 data를 이용해서 학습을 시킬경우 학습에 실패할 수 있다.

3) 세번째 학습을 $\alpha = 0.01$에 대해 parameter update equation을 구해보면

$$\theta := \theta + 0.02 \cdot x^{(1)}(y^{(1)} - \theta \cdot x^{(1)}) = \theta + 0.02(2-\theta)$$

$$\theta := \theta + 0.02 \cdot x^{(2)}(y^{(2)} - \theta \cdot x^{(2)}) = \theta + 0.06(6-3\theta) = \theta + 0.18(2-\theta)$$

$$\theta := \theta + 0.02 \cdot x^{(3)}(y^{(3)} - \theta \cdot x^{(3)}) = \theta + 0.08(8-4\theta) = \theta + 0.32(2-\theta)$$

가 된다. 이를 이용하여 2)에서와 같이 3 iterations 동안 θ의 변화를 구하면 다음과 같다.

	(1,2)	(3,6)	(4,8)
1st iteration:	$\theta := 1 + 0.02(2-1) = 1.02$	$\theta := 1 + 0.18(2-1) = 1.18$	$\theta := 1 + 0.32(2-1) = 1.32$
2nd iteration:	$\theta := 1.02 + 0.02(2-1.02) = 1.0396$	$\theta := 1.18 + 0.18(2-1.18) = 1.3276$	$\theta := 1.32 + 0.32(2-1.32) = 1.5376$
3rd iteration:	$\theta := 1.04 + 0.02(2-1.04) = 1.0588$	$\theta := 1.3276 + 0.18(2-1.3276) = 1.4486$	$\theta := 1.5376 + 0.32(2-1.5376) = 1.6856$

위의 결과에서 알 수 있듯이 (1,2)에서는 상당히 학습속도가 늦어졌지만 안정적인 학습이 되는 것을 알 수 있다.

또한 (3,6)에서는 더이상 불안정한 모습의 학습을 보여주지 않는다.

역시나 (4,8)에 대해서는 더이상 발산하는 모습을 보여주지 않는다. 즉, 2)에서의 불안정성이 해결되었다.

따라서 모든 data sample에 대해서 iteration마다 학습이 올바른 방향으로 일어난다.