Question.3-10

Linear regression을 위한 dataset이 다음과 같이 주어졌다. $y = 3x$인것

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

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

1) 1 epoch 동안 각 data sample들에 대해 loss와 θ가 update되는 절댓값을 구하시오.
그리고 loss의 감소 graph의 fluctuation이 생기는 이유를 설명하시오.

단, $\gamma > 1$

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

1) 먼저 square error를 이용한 loss와 θ를 update 시키는 gradient descent의 식은 data sample $(x^{(i)}, y^{(i)})$에 대해 다음과 같다.

$$\mathcal{L}^{(i)} = (y^{(i)} - \hat{y}^{(i)})^2 = (y^{(i)} - \theta x^{(i)})^2$$

$$\theta := \theta + 2\alpha x^{(i)}(y^{(i)} - \hat{y}^{(i)}) = \theta + 0.02 x^{(i)}(y^{(i)} - \theta x^{(i)})$$

이때 각 data sample들을 이용하여 θ를 update시키면 $\mathcal{L}$와 update되는 θ는 다음과 같다.

1st iteration: $\mathcal{L}^{(1)} = (y^{(1)} - \theta \cdot x^{(1)})^2 = (3 - 1\cdot 1)^2 = 4$ $\quad$ $\theta := \theta + 0.02\cdot x^{(1)}(y^{(1)} - \theta\cdot x^{(1)}) = 1 + 0.02\cdot 1(3 - 1\cdot 1) = 1.04$ $\quad +0.04$

2nd iteration: $\mathcal{L}^{(2)} = (y^{(2)} - \theta\cdot x^{(2)})^2 = (9 - 1.04\cdot 3)^2 = 34.57$ $\quad$ $\theta := \theta + 0.02 x^{(2)}(y^{(2)} - \theta\cdot x^{(2)}) = 1.04 + 0.02\cdot 3(9 - 1.04\cdot 3) = 1.39$ $\quad +0.35$

3rd iteration: $\mathcal{L}^{(3)} = (y^{(3)} - \theta\cdot x^{(3)})^2 = (6 - 1.39\cdot 2)^2 = 10.38$ $\quad$ $\theta := \theta + 0.02\cdot x^{(3)}(y^{(3)} - \theta\cdot x^{(3)}) = 1.39 + 0.02\cdot 2(6 - 1.39\cdot 2) = 1.52$ $\quad +0.13$

4th iteration $\mathcal{L}^{(4)} = (y^{(4)} - \theta\cdot x^{(4)})^2 = (-3 - 1.52\cdot(-1))^2 = 2.19$ $\quad$ $\theta := \theta + 0.02\cdot x^{(4)}(y^{(4)} - \theta\cdot x^{(4)}) = 1.52 + 0.02\cdot(-1)(-3 - 1.52\cdot(-1)) = 1.55$ $\quad +0.02$

위의 결과에서 알 수 있듯이 $(x^{(i)}, y^{(i)})$에 따라서 $\mathcal{L}^{(i)}$와 θ가 update되는 양은 fluctuation이 생긴다. 이는

$$\mathcal{L}^{(i)} = (y^{(i)} - \theta x^{(i)})^2 \qquad \Delta\theta = 2\alpha x^{(i)}(y^{(i)} - \theta x^{(i)})$$

에서 알 수 있듯이 $\mathcal{L}^{(i)}$는 $y^{(i)}$와 $\hat{y}^{(i)}$의 차이에 제곱에 비례하여 커지고 $\Delta\theta$는 $x^{(i)}$, $y^{(i)}$와 $\hat{y}^{(i)}$의 차이에 각각 비례한다. 따라서 하나의 data sample을 이용하여 θ를 학습시킬 때, $\mathcal{L}^{(i)}$는 $x^{(i)}, y^{(i)}$에 따라 fluctuating하게 되고, $\mathcal{L}^{(i)}$가 fluctuating 함에따라 $\Delta\theta$도 fluctuating하게 된다.

2) 위의 식

$$\mathcal{L}^{(i)} = (y^{(i)} - \theta x^{(i)})^2 \qquad \Delta\theta = 2\alpha x^{(i)}(y^{(i)} - \theta x^{(i)})$$

에서 $x^{(i)}$가 γ배 증가하여 $\gamma\cdot x^{(i)}$가 되었다고 가정해보자. 그러면 dataset은 $y = 3x$에서부터 만들어졌으므로 $y^{(i)}$는 $\gamma\cdot y^{(i)}$가 된다. 이 상황에서 data sample $(\gamma\cdot x^{(i)}, \gamma\cdot y^{(i)})$에 대한 loss $\mathcal{L}'$와 θ의 update되는 양 $\Delta\theta'$는

$$\mathcal{L}' = (\gamma\cdot y^{(i)} - \gamma\cdot x^{(i)})^2 = \gamma^2 (y^{(i)} - x^{(i)})^2$$

$$\Delta\theta' = 2\alpha(\gamma x^{(i)})(\gamma\cdot y^{(i)} - \theta(\gamma x^{(i)})) = \gamma^2\cdot 2\alpha x^{(i)}(y^{(i)} - \theta x^{(i)})$$

가 되므로 각각 γ^2배씩 증가하게 된다. 따라서 θ는 같은 α에 대해서 $(x^{(i)}, y^{(i)})$보다 $(\gamma\cdot x^{(i)}, \gamma\cdot y^{(i)})$를 이용하여 update 될 때 발산할 가능성이 커진다.