

Question. 5-03

Dataset \mathcal{D} 가 다음과 같이 주어졌다.

$$\mathcal{D} = \{(x_2^{(1)}, x_1^{(1)}, y^{(1)}), (x_2^{(2)}, x_1^{(2)}, y^{(2)}), (x_2^{(3)}, x_1^{(3)}, y^{(3)})\} = \{(1,1,4), (3,2,5), (6,4,5)\}$$

Dataset을 $y = 2x_2 - 3x_1 + 5$ 에서부터 만들었기 때문에, 모델을 $\hat{y} = \theta_2 x_2 + \theta_1 x_1 + \theta_0$ 로 설정하였다.

initial $\vec{\theta}$ 가 $\theta_2 = \theta_1 = \theta_0 = 1$ 이고 learning rate = 0.1로 설정하였을 때, 다음 질문에 답하시오.

- 1) 각 Data sample별로 1 iteration씩 학습을 진행했을 때 각각 $\vec{\theta}$ 의 변화를 구하시오.
- 2) 1)에서의 결과를 참고하여 각 Data sample별로 $\frac{\partial \mathcal{L}(\vec{\theta})}{\partial \theta_2}, \frac{\partial \mathcal{L}(\vec{\theta})}{\partial \theta_1}, \frac{\partial \mathcal{L}(\vec{\theta})}{\partial \theta_0}$ 을 구하시오.
- 3) 위의 결과를 통해 $x_2^{(i)}$ 가 γ 배 되었을 때, $\theta_2, \theta_1, \theta_0$ 가 update되는 양이 몇 배 차이 나는지 구하시오.
- 4) 위의 결과를 통해 $x_1^{(i)}$ 가 γ 배 되었을 때, $\theta_2, \theta_1, \theta_0$ 가 update되는 양이 몇 배 차이 나는지 구하시오.

1)

$$\hat{y} = \theta_2 x_2 + \theta_1 x_1 + \theta_0 \text{ 이므로 } \mathcal{L} = (\hat{y} - y)^2 = ((4 - \hat{y})^2 = ((4 - (\theta_2 x_2 + \theta_1 x_1 + \theta_0))^2 \text{이다.}$$

이때 $\frac{\partial \mathcal{L}}{\partial \theta_2} = -2(4 - \hat{y}), \frac{\partial \mathcal{L}}{\partial \theta_1} = -2x_1(4 - \hat{y}), \frac{\partial \mathcal{L}}{\partial \theta_0} = -2(4 - \hat{y})$ 이므로 gradient descent method는

$$\theta_2 := \theta_2 - \alpha \frac{\partial \mathcal{L}}{\partial \theta_2} = \theta_2 + 2x_2(4 - \hat{y})$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial \mathcal{L}}{\partial \theta_1} = \theta_1 + 2x_1(4 - \hat{y})$$

$$\theta_0 := \theta_0 - \alpha \frac{\partial \mathcal{L}}{\partial \theta_0} = \theta_0 + 2(4 - \hat{y})$$

으로 표현할 수 있다.

따라서 $(x_2, x_1, y) = (1, 1, 4)$ 에 의한 $\vec{\theta}$ 의 학습은

$$\theta_2 := \theta_2 + 2x_2(4 - \hat{y}) = 1 + 0.2 \cdot 1 \cdot (4 - 3) = 1 + 0.2 = 1.2$$

$$\theta_1 := \theta_1 + 2x_1(4 - \hat{y}) = 1 + 0.2 \cdot 1 \cdot (4 - 3) = 1 + 0.2 = 1.2$$

$$\theta_0 := \theta_0 + 2(4 - \hat{y}) = 1 + 0.2 \cdot (4 - 3) = 1 + 0.2 = 1.2 \text{으로}$$

$(\theta_2, \theta_1, \theta_0) = (1, 1, 1)$ 에서 $(1.2, 1.2, 1.2)$ 로 +0.2씩 증가하는 것을 알 수 있다.

$(x_2, x_1, y) = (3, 2, 5)$ 에 의한 $\vec{\theta}$ 의 학습은

$$\theta_2 := \theta_2 + 2x_2(4 - \hat{y}) = 1 + 0.2 \cdot 3 \cdot (5 - 6) = 1 + (-0.6) = 0.4$$

$$\theta_1 := \theta_1 + 2x_1(4 - \hat{y}) = 1 + 0.2 \cdot 2 \cdot (5 - 6) = 1 + (-0.4) = 0.6$$

$$\theta_0 := \theta_0 + 2(4 - \hat{y}) = 1 + 0.2 \cdot (5 - 6) = 1 + (-0.2) = 0.8 \text{로}$$

$(\theta_2, \theta_1, \theta_0) = (1, 1, 1)$ 에서 $(0.4, 0.6, 0.8)$ 로 $(-0.6, -0.4, -0.2)$ 만큼 이동하는 것을 알 수 있다.

$(x_2, x_1, y) = (6, 4, 5)$ 에 의한 $\vec{\theta}$ 의 학습은

$$\theta_2 := \theta_2 + 2x_2(4 - \hat{y}) = 1 + 0.2 \cdot 6 \cdot (5 - 11) = 1 + (-7.2) = -6.2$$

$$\theta_1 := \theta_1 + 2x_1(4 - \hat{y}) = 1 + 0.2 \cdot 4 \cdot (5 - 11) = 1 + (-4.8) = -3.8$$

$$\theta_0 := \theta_0 + 2(4 - \hat{y}) = 1 + 0.2 \cdot (5 - 11) = 1 + (-1.2) = -0.2 \text{로}$$

$(\theta_2, \theta_1, \theta_0) = (1, 1, 1)$ 에서 $(-6.2, -3.8, -0.2)$ 로 $(-7.2, -4.8, -1.2)$ 만큼 이동하는 것을 알 수 있다.

- 2) $(\chi_2, \chi_1, y) = (1, 1, 4)$ 에서의 $\frac{\partial f}{\partial \theta_2}, \frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_0} = 0.2 : 0.2 : 0.2 = 1 : 1 : 1$ 이다.
 $(\chi_2, \chi_1, y) = (3, 2, 5)$ 에서의 $\frac{\partial f}{\partial \theta_2}, \frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_0} = -0.6 : -0.4 : -0.2 = 3 : 2 : 1$ 이다.
 $(\chi_2, \chi_1, y) = (6, 4, 5)$ 에서의 $\frac{\partial f}{\partial \theta_2}, \frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_0} = -1.2 : -4.8 : -1.2 = 6 : 4 : 1$ 이다.
- 즉, $\vec{\theta}$ 가 update되는 양의 배인 $\frac{\partial f}{\partial \theta_2}, \frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_0}$ 는 input \vec{x} 의 배를 따르는 것을 할 수 있다.

3) $\frac{\partial f}{\partial \theta_2} : \frac{\partial f}{\partial \theta_1} : \frac{\partial f}{\partial \theta_0} = -2\chi_2(y - \hat{y}) : -2\chi_1(y - \hat{y}) : -2(y - \hat{y}) = \chi_2 : \chi_1 : 1$ 이므로
 $\chi_2 := \gamma \chi_2$ 일 때 $\frac{\frac{\partial f}{\partial \theta_2}}{\frac{\partial f}{\partial \theta_1}} = \frac{\gamma \chi_2}{\chi_1}, \frac{\frac{\partial f}{\partial \theta_2}}{\frac{\partial f}{\partial \theta_0}} = \frac{\gamma \chi_2}{1}$ 이 된다.

따라서 θ_2 는 θ_2 의 $\frac{\gamma}{\chi_2}$ 배 만큼, θ_0 의 $\frac{\gamma}{\chi_2}$ 배 만큼 update 된다.

4) $\frac{\partial f}{\partial \theta_2} : \frac{\partial f}{\partial \theta_1} : \frac{\partial f}{\partial \theta_0} = -2\chi_2(y - \hat{y}) : -2\chi_1(y - \hat{y}) : -2(y - \hat{y}) = \chi_2 : \chi_1 : 1$ 이므로
 $\chi_1 := \gamma \chi_1$ 일 때 $\frac{\frac{\partial f}{\partial \theta_2}}{\frac{\partial f}{\partial \theta_1}} = \frac{\chi_1}{\chi_2}, \frac{\frac{\partial f}{\partial \theta_2}}{\frac{\partial f}{\partial \theta_0}} = \frac{\chi_1}{1}$ 이 된다.

따라서 θ_1 는 θ_2 의 $\frac{\chi_1}{\chi_2}$ 배 만큼, θ_0 의 $\frac{\chi_1}{\chi_2}$ 배 만큼 update 된다.