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

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

Question.3-06 2)의 결과에서 확인할 수 있듯이 같은 learning rate에 대해서도 $x^{(i)}$와 $y^{(i)}$의 크기에 따라 학습 중 θ를 발산시킬 수 있다.

이는 θ를 update하는 식

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

에서 $|x^{(i)}(y^{(i)} - \theta \cdot x^{(i)})|$가 지나치게 커서 발생하는 문제이다.

위의 dataset에서 θ를 θ^*의 방향으로 학습시키는 data sample들의 집합 P와 θ를 발산시키는 data sample들의 집합 Q를 다음과 같이 생각해보자.

$$P = \{(p_x^{(1)}, p_y^{(1)}), (p_x^{(2)}, p_y^{(2)}), \cdots, (p_x^{(n_p)}, p_y^{(n_p)})\}$$
$$Q = \{(q_x^{(1)}, q_y^{(1)}), (q_x^{(2)}, q_y^{(2)}), \cdots, (q_x^{(n_q)}, q^{(n_q)})\}$$

이때 각 P, Q의 cardinality의 관계는 다음과 같다.

$$|P| \gg |Q|$$

만약 그렇지 않다면 learning rate 전체적으로 크다는 의미이므로 learning rate를 줄여야 한다.

그리고 위의 data sample들을 이용하여 cost를 구하면 각각 P, Q에 대한 $\mathcal{L}_p, \mathcal{L}_q$에 대해

$$J = \frac{1}{n_p}\sum_{i=1}^{n_p}\mathcal{L}_p^{(i)} + \frac{1}{n_q}\sum_{i=1}^{n_q}\mathcal{L}_q^{(i)}$$

가 되고, $\frac{\partial J}{\partial \theta}$는

$$\frac{\partial J}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\frac{1}{n_p}\sum_{i=1}^{n_p}\mathcal{L}_p^{(i)} + \frac{1}{n_q}\sum_{i=1}^{n_q}\mathcal{L}_q^{(i)}\right]$$
$$= \frac{1}{n_p}\sum_{i=1}^{n_p}\left(\frac{\partial \mathcal{L}_p^{(i)}}{\partial \theta}\right) + \frac{1}{n_q}\sum_{i=1}^{n_q}\left(\frac{\partial \mathcal{L}_q^{(i)}}{\partial \theta}\right)$$
$$= \frac{1}{n_p}\sum_{i=1}^{n_p}\left[-2p_x^{(i)}(\hat{p}_y^{(i)} - \theta \cdot p_x^{(i)})\right] + \frac{1}{n_p}\sum_{i=1}^{n_p}\left[-2q_x^{(i)}(\hat{q}_y^{(i)} - \theta \cdot q_x^{(i)})\right]$$

가 된다. 여기 $\hat{p}_y^{(i)}, \hat{q}_y^{(i)}$는 각각 predictor의 $p_y^{(i)}, q_y^{(i)}$에 대한 prediction이다. 고로 이를 이용하여 θ를 update시키면

$$\theta := \theta - \alpha\left[\frac{1}{n_p}\sum_{i=1}^{n_p}\left[-2p_x^{(i)}(\hat{p}_y^{(i)} - \theta \cdot p_x^{(i)})\right] + \frac{1}{n_p}\sum_{i=1}^{n_p}\left[-2q_x^{(i)}(\hat{q}_y^{(i)} - \theta \cdot q_x^{(i)})\right]\right]$$

이 된다. 고로 파란 부분은 θ를 안정적으로 update시키고 빨간 부분은 θ를 발산시키는데 $|P| \gg |Q|$이므로 이 둘의 결과는 θ를 안정적으로 update시킬 가능성이 크다. 위의 이유로 θ에 대한 학습의 불안정성을 일으키는 outlier들의 영향력을 줄일 수 있다. 물론 outlier의 크기가 너무 크면 θ를 발산시킬 수 있지만 그 경우에도 outlier 하나만을 이용하여 θ를 update하는 경우보단 θ를 급격하게 변화시키진 않는다.

따라서 위의 식은 여러 개의 data sample을 이용하여 θ를 update시키면 outlier의 영향력을 최소화하여 predictor의 학습을 보다 안정적으로 만들 수 있다.