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

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

의 관계를 설명하시오.

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

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

두 과정의 차이점을 설명하시오.

1) 먼저 3개의 $\mathcal{L}^{(i)}$를 이용해 cost J를 구하는 식은 다음과 같다.

$$J(\theta) = \frac{1}{3}\sum_{i=1}^{3} \mathcal{L}^{(i)}(\theta)$$

이를 θ에 대해 미분하면

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\frac{1}{3}\sum_{i=1}^{3} \mathcal{L}^{(i)}(\theta)\right]$$
$$= \frac{1}{3}\sum_{i=1}^{3}\left[\frac{\partial \mathcal{L}^{(i)}(\theta)}{\partial \theta}\right]$$
$$= \frac{1}{3}\left[\frac{\partial \mathcal{L}^{(1)}(\theta)}{\partial \theta} + \frac{\partial \mathcal{L}^{(2)}(\theta)}{\partial \theta} + \frac{\partial \mathcal{L}^{(3)}(\theta)}{\partial \theta}\right]$$

이다. 즉, cost는 loss들의 평균값인 $\frac{\partial J(\theta)}{\partial \theta}$도 각 $\frac{\partial \mathcal{L}^{(i)}(\theta)}{\partial \theta}$들의 평균값이다.

이의 결과를 이용하면

$$\theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$$
$$= \theta - \frac{\alpha}{3}\left[\frac{\partial \mathcal{L}^{(1)}(\theta)}{\partial \theta} + \frac{\partial \mathcal{L}^{(2)}(\theta)}{\partial \theta} + \frac{\partial \mathcal{L}^{(3)}(\theta)}{\partial \theta}\right]$$

가 된다. 즉 cost를 이용하여 θ를 update하면 각 loss들에 대해 update 되는 값들을 평균적으로 반영하여 update한다.

2) Question.3-06에서 data sample (4,8)를 이용하여 θ를 update하면 다음과 같이 발산했다.

1st iteration: $\theta := 1 + 3.2(2-1) = 4.2$

2nd iteration: $\theta := 4.2 + 3.2(2-4.2) = -2.84$

3rd iteration: $\theta := -2.8 + 3.2(2+2.84) = 12.65$

2)와 1)에서 설명한대로 cost를 이용하여 θ를 update시키면 각 θ들의 평균을 이용하므로

	(1, 2)	(3, 6)	(4, 8)
$-\alpha \frac{\partial \mathcal{L}^{(i)}(\theta)}{\partial \theta}$	$0.2(2-\theta)$	$1.8(2-\theta)$	$3.2(2-\theta)$
1st iteration	$0.2(2-1) = 0.2$	$1.8(2-1) = 1.8$	$3.2(2-1) = 3.2$

에 대해

$$\theta := 1 + \frac{1}{3}(0.2 + 1.8 + 3.2) = 2.73$$

로 θ가 update된다. 즉, θ를 발산시키는 (4,8)의 영향력은 $\frac{1}{3}$로 줄게되고, θ를 제대로 학습시키는 (1,2), (3,6)의 영향력이 $\frac{2}{3}$를 차지하면서 θ의 발산을 막을 것이다. learning rate이 적절히 작다면 대부분의 data sample들은 θ를 제대로 학습시킬 것이고, 소수의 outlier들이 θ를 발산시킬 것이다. cost를 이용하면 outlier로 인한 학습의 불안정성을 줄일 수 있다.