Cross Validated

Derivation of Regularized Linear Regression Cost Function per Coursera Machine Learning Course

Asked 9 years, 1 month ago Modified 11 months ago Viewed 21k times

15

I took Andrew Ng's course "Machine Learning" via Coursera a few months back, not paying attention to most of the math/derivations and instead focusing on implementation and practicality. Since then I have started going back to study some of the underlying theory, and have revisited some of Prof. Ng's lectures. I was reading through his lecture on "Regularized Linear Regression", and saw that he gave the following cost function:

$$J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2]$$

Then, he gives the following gradient for this cost function:

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} - \lambda\theta_j]$$

I am a little confused about how he gets from one to the other. When I tried to do my own derivation, I had the following result:

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) + y^{(i)})x_j^{(i)} + \lambda\theta_j]$$

The difference is the 'plus' sign between the original cost function and the regularization parameter in Prof. Ng's formula changing into a 'minus' sign in his gradient function, whereas that is not happening in my result.

Intuitively I understand why it's negative: we are reducing the theta parameter by the gradient figure, and we want the regularization parameter to reduce the amount that we are changing the parameter to avoid overfitting. I am just a little stuck on the calculus that backs this intuition.

FYI, you can find the deck here, on slides 15 and 16.

regression self-study

Share Cite Improve this question Follow

edited Aug 10, 2014 at 22:01

asked Aug 10, 2014 at 21:54

wellington
689 2 6 13

1 In your result you have a "**+**" preceding the y^(i)--is that a typo? – Steve S Aug 10, 2014 at 22:26

3 Answers

Sorted by: Highest score (default)

7

Actually if you check the lecture notes just after the video , it shows the formula correctly . The slides that you have lined here shows the exact slide of the video.

Repeat {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \right] \qquad j \in \{1, 2 \ldots n\}$$

}

Share Cite Improve this answer Follow

answered Jul 19, 2018 at 7:49

Piyush
86 1 1

[coursera.org/learn/machine-learning/supplement/pKAsc/…](coursera.org/learn/machine-learning/supplement/pKAsc/) here its the link to the notes right after the video showing correct formula. – RoundPi Aug 12, 2018 at 11:49

14

$J(\theta) = \frac{1}{2m} [\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2]$

Now

$\frac{\partial}{\partial \theta_j} (h_\theta(x^{(i)}) - y^{(i)})^2 = 2[(h_\theta(x^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_j} \{h_\theta(x^{(i)})\}]$

Note that in a linear model (being discussed on the pages you mention), $\frac{\partial}{\partial \theta_j} (h_\theta(x^{(i)}) = [x^{(i)}]_j$

$\frac{\partial}{\partial \theta_j} \lambda \sum_{j=1}^n \theta_j^2 = 2\lambda\theta_j$

So for the linear case

$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} [\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} + \lambda\theta_j]$

Looks like perhaps both you and Andrew might have typos. Well, at least two of the three of us seem to.

Share Cite Improve this answer Follow

edited Oct 23, 2022 at 9:10

answered Aug 11, 2014 at 0:22

Glen_b
275k 37 605 1005

1 its confirmed, just a typo on the Andrew's note, it should be a + sign. And Prof correctly explain everything correctly including the intuition θ(1-α(λ/m)) meaning every time this shrink θ then minus the usual part before regularisation being introduced. – RoundPi Aug 12, 2018 at 11:55

How does the derivative of the regularization term work out to 2λθj? I understand the power rule, but I'm not sure what happens to the inner function because I thought the ∑ notation would appear in the resulting derivative. – bbk611 Oct 23, 2022 at 7:42

2 $\frac{\partial}{\partial x}(x^2+y^2+z^2)=2x$ ... why would you expect to see a sum involving the terms in y and z? – Glen_b Oct 23, 2022 at 9:10

That makes sense. Thank you. – bbk611 Oct 23, 2022 at 10:06

To add more context on how ∑ goes off: `(d/dw_j) Σ[j=1 to n] w_j^2 = (d/dw_j) (w_1^2 + w_2^2 + ... + w_j^2 + ... + w_n^2) = 0 + 0 + ... + 2w_j + ... + 0 = 2w_j` – Abhishek E H Apr 24 at 8:01

2

Actually, I think that's just a typo.

On slide #16 he writes the derivative of the cost function (with the regularization term) with respect to theta *but it's in the context of the **Gradient Descent*** algorithm. Hence, he's also multiplying this derivative by $-\alpha$. Notice: On the second line (of slide 16) he has $-\lambda\theta$ (as you've written), multiplied by $-\alpha$. However, by the third line the multiplied term *is still negative* even though--if the second line were correct--the negative signs would've cancelled out.

Make sense?

Share Cite Improve this answer Follow

answered Aug 10, 2014 at 22:53

Steve S
1,094 8 17