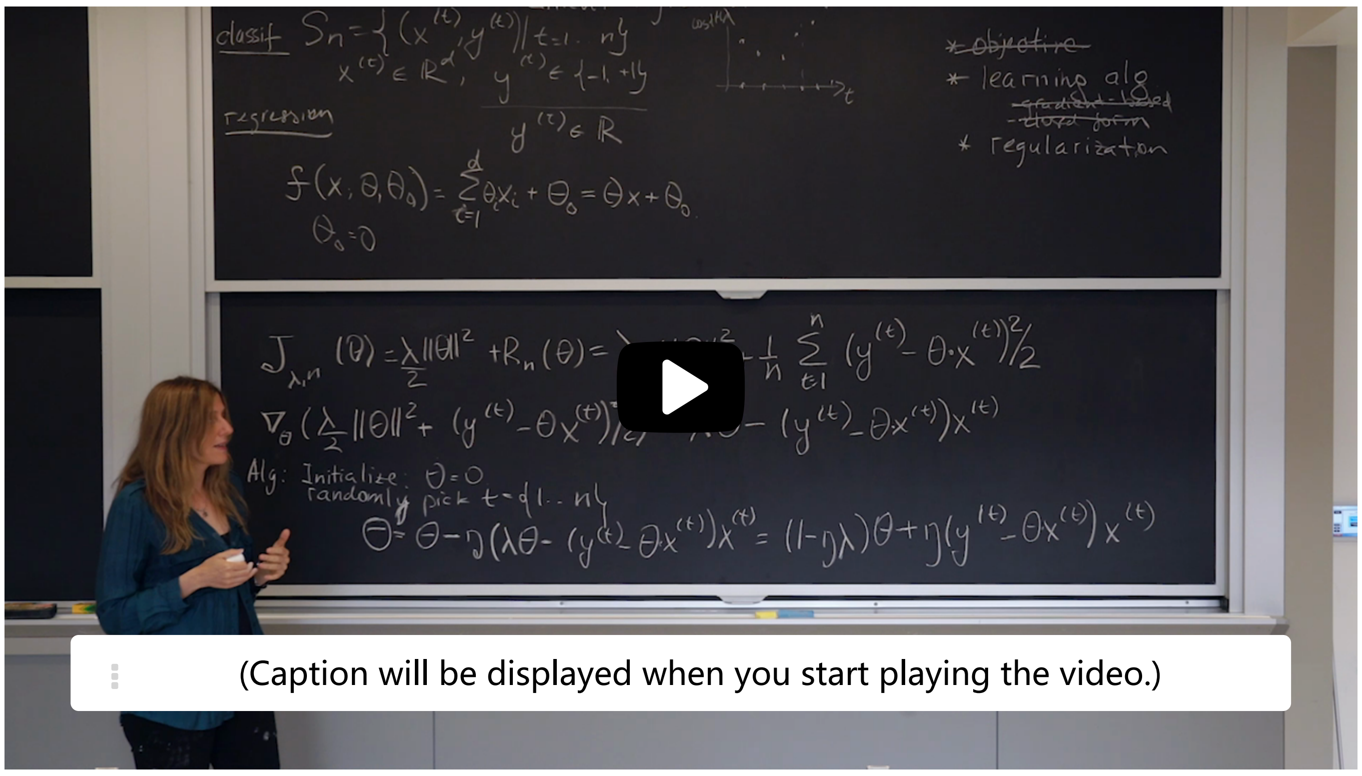
9. Closing Comment

Closing Comment

Start of transcript. Skip to the end.



0:00 / 0:00 1.0x

Now what I want to do before we close today's lecture is

actually is to say jointly what this

regularization is doing.

It doesn't matter how, at this point,

which algorithm do you use.

I want to bring you back to this formula,

to the Suivche regression formula

and think together with me, what does it do?

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

(Optional) Equivalance of regularization to a Gaussian Prior on Weights

(Optional) Equivalance of regularization to a Gaussian Prior on Weights

The regularized linear regression can be interpreted from a probabilistic point of view. Suppose we are fitting a linear regression model with n data points $((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n))$, assume the ground truth is that y is linearly related to x but we also observed some noise ϵ for y:

$$y_t = \theta \cdot x_t + \epsilon$$

where $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

Then the likelihood of our observed data is

$$\prod_{t=1}^{n} \mathcal{N}\left(y_t | \theta x_t, \sigma^2\right).$$

Now, if we impose a Gaussian prior $\mathcal{N}\left(\theta | 0, \lambda^{-1}\right)$, the likelihood will change to

$$\prod_{t=1}^{n} \mathcal{N}(y_t|\theta x_t, \sigma^2)\,\mathcal{N}(\theta|0, \lambda^{-1})\,.$$

Take the logarithim of the likelihood, we will end up with

$$\sum_{t=1}^{n} -\frac{1}{2\sigma^2}(y_t - \theta x_t)^2 - \frac{1}{2}\lambda\|\theta\|^2 + \text{constant}.$$

Try to derive this result by yourself. Can you conclude that maximizing this loglikelihood equivalent to minimizing the regularized loss in the linear regression? What does larger λ mean in this probabilistic interpretation? (Think of the error decomposition we discussed.)

Hide

Discussion

Hide Discussion

Topic: Unit 2 Nonlinear Classification, Linear regression, Collaborative Filtering (2 weeks):Lecture 5. Linear Regression / 9. Closing Comment

Add a Post

< All Posts

Derivation of loglikelihood inside, spoiler alert

discussion posted 3 days ago by Cool7 (Community TA)

As title. This is easier than last one. Just put it here in case somebody interested. I'm practicing my latex writing, lol.

$$\log\left(\prod_{t=1}^{n}\mathcal{N}(y_t|\theta x_t, \sigma^2)\mathcal{N}(\theta|0, \lambda^{-1})\right)$$
$$=\sum_{t=1}^{n}\left(\log(\mathcal{N}(y_t|\theta x_t, \sigma^2)) + \log(\mathcal{N}(\theta|0, \lambda^{-1}))\right)$$
$$=n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{t=1}^{n} log\left(e^{-\frac{(y_t-\theta x_t)^2}{2\sigma^2}}\right) + n\log\left(\sqrt{\frac{\lambda}{2\pi}}\right) + \sum_{t=1}^{n} log\left(e^{-\frac{\lambda\|\theta\|^2}{2}}\right)$$
$$=\sum_{t=1}^{n}\left(-\frac{1}{2\sigma^2}(y_t - \theta x_t)^2 - \frac{\lambda}{2}\|\theta\|^2\right) + n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + n\log\left(\sqrt{\frac{\lambda}{2\pi}}\right)$$
$$=\sum_{t=1}^{n} -\frac{1}{2\sigma^2}(y_t - \theta x_t)^2 - \frac{1}{2}\lambda\|\theta\|^2 + \text{constant}$$

My understanding is

- First term is related to posterior distribution, it represents the accuracy of the estimation/training loss/bias.
- Second term is related to prior distribution, it represents the regularization(recall we imposed it on) / variance.

Thus λ as hyper parameter is to adjust the weights between bias and variance, inline with the error decomposition discussed a few pages before.

This post is visible to everyone.

Add a Response

1 response

Alexander_Konstantinidis
2 days ago

Another way to view this, is to consider λ as expressing the degree of our certainty (prior belief) that there is no real explanatory value in the model or stated differently very few if any of the predictors truly matter. (This is because λ is the inverse of the variance of probabilistic theta). The higher the λ the more evidence will be required to arrive to a complex model and vice versa.

Indeed, this is a very interesting interpretation. In the extreme case, where lambda is infinity, it means your prior belief is so strong that no matter what data is presented, the hard coded parameters do not change. On the other extreme, when lambda is 0, variance is infinity and thus you don't have a prior belief. Data takes control of everything, even if there're a lot of noise. So a moderate lambda lets the model to learn from data, but regularizes the parameters so that do not deviate too much from the prior belief.

posted 2 days ago by **FutureStar**

extending this line of thought to the gradient approach explained by the professor, the regularization term is only explained by the multiplier $(1 - \eta.\lambda)$. If λ were to take the value $1/\eta$ then the regularization term would disappear and the gradient expression would then only depend on the data.

Another instance would be if $\eta.\lambda$ were >1, the regularization parameter turns negative. How do we intuitively explain this? So kind of struggling to reconcile the two views (gradient vs log-likelihood).

posted about 8 hours ago by **RanganN**

Add a comment

Showing all responses

Add a response:

Preview

Submit