The effect of adding a small ridge penalty

October 9, 2016

We will show that adding a small ridge penalty scaled by some constant w does not change the fitted model by very much.

The proof will presume a parametric model space and that the training criterion is strongly convex.

Unfortunately, it is unclear how to extend this proof technique to the non-smooth case. In addition, showing this result for non-parametric regression models is quite difficult. More assumptions are probably needed. It may be easier to consider specific regression problem examples.

1 Parametric Models: Strongly Convex Penalized Objective

Let the training criterion be denoted L_T

$$L_T(\boldsymbol{\theta}) = \frac{1}{2}\|y - f(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta})$$

Suppose $\nabla^2 L_T(\theta)$ exists and the training criterion is m-strongly convex in $\boldsymbol{\theta}$. That is, there is some constant $m > 0$ such that

$$\nabla^2 L_T(\boldsymbol{\theta}) \succeq mI$$

Consider the minimizer of the perturbed problem

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w) = \arg\min_{\theta} L_T(\boldsymbol{\theta}) + \sum_{j=1}^{J} \lambda_j \frac{w}{2}\|\boldsymbol{\theta}\|^2$$

So $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)$ is the solution to the original penalized regression problem.

Then for any w, we have

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w) - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|_2 \quad \leq \quad \frac{2}{m} w \left(\sum_{j=1}^{J} \lambda_j \right) \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|$$

Proof

By page 460 of Boyd, we know that for strongly convex loss functions, we have that

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w) - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|_2 \leq \frac{2}{m} \|\nabla L_T(\boldsymbol{\theta})\|_{\theta = \hat{\theta}_\lambda(w)}$$

By the gradient optimality conditions, we have that

$$\nabla L_T(\boldsymbol{\theta})|_{\theta = \hat{\theta}_\lambda(w)} + \sum_{j=1}^{J} \lambda_j w \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w) = 0$$

So

$$\|\nabla L_T(\boldsymbol{\theta})\|_{\theta=\hat{\theta}_\lambda(w)} = \left(\sum_{j=1}^J \lambda_j\right) w\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)\|$$

We can show that

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)\|^2 \le \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|^2$$

To see this, use the definitions of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)$:

$$L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)) + \sum_{j=1}^J \lambda_j \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)\|^2 \le L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)) + \sum_{j=1}^J \lambda_j \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|^2$$

and

$$L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)) \le L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w))$$

Plugging in the inequality, we get

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w) - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|_2 &\le \frac{2}{m} w \left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)\| \\
&\le \frac{2}{m} w \left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|
\end{aligned}$$