The effect of adding a small ridge penalty

November 3, 2016

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

1. We show that as $w \to 0$, the fitted model to the perturbed training criterion converges to the fitted model for the original training criterion. This uses the implicit function theorem. The result applies to parametric models where the training criterion can contain smooth or nonsmooth penalties. The proof technique can probably be extended to (certain) nonparametric models (using an implicit function theorem for banach spaces).

2. This older proof shows that the fitted model is actually Lipschitz in w. However, it requires the strong assumption that the training criterion was strongly convex. The proof is still here because... it is cute.

1 Continuity of ridge perturbation

We will consider the case of p-dimensional parametric models. Let

$$\hat{\boldsymbol{\theta}}_w = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|y - f\left(\cdot|\boldsymbol{\theta}\right)\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(\boldsymbol{\theta}) + \frac{w}{2}\|\boldsymbol{\theta}\|^2\right)$$

Let

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

Suppose that $P_j(\boldsymbol{\theta})$ and $f(\cdot|\boldsymbol{\theta})$ are continuously differentiable for all $\boldsymbol{\theta}$. Then there is a $W > 0$ such that $\hat{\boldsymbol{\theta}}_w$ is a continuous mapping from $[0, W)$ into some open neighborhood $B \subseteq \mathbb{R}^p$.

Proof

Consider the function

$$D(w, \boldsymbol{\theta}) = \nabla_\theta \left[L_T(\boldsymbol{\theta}|\boldsymbol{\lambda}) + \sum_{j=1}^{J} \lambda_j \frac{w}{2}\|\boldsymbol{\theta}\|^2\right]$$

Since $D(\cdot, \boldsymbol{\theta}) : \mathbb{R} \mapsto \mathbb{R}$ is one-to-one, then by the Implicit Function Theorem (Kumagai 1980), there is a unique solution to $D\left(w, \cdot\right) = 0$. By the gradient optimality conditions, we know that the solution must be $\hat{\boldsymbol{\theta}}_w$. Moreover, the Implicit Function Theorem states that there is some $W > 0$ such that $\hat{\boldsymbol{\theta}}_w$ is a continuous mapping from $[0, W)$ to some open subset in $\mathbb{R}^p$.

Source: Kumagai, 1980. An Implicit Function Theorem: Comment

Extension to Nonsmooth case

Let the differentiable space at $\hat{\boldsymbol{\theta}}_0$ be defined as

$$\Omega = \left\{ \boldsymbol{\eta} | \lim_{\epsilon \to 0} \frac{L_T(\hat{\boldsymbol{\theta}}_0 + \epsilon \boldsymbol{\eta} | \boldsymbol{\lambda}) - L_T(\hat{\boldsymbol{\theta}}_0 | \boldsymbol{\lambda})}{\epsilon} \text{ exists} \right\}$$

Let U be an orthonormal basis of Ω where U has rank $q \leq p$.
Suppose that for all $w < W'$, we have that

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2} \|y - f(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(\boldsymbol{\theta}) + w \|\boldsymbol{\theta}\|^2 \right) = \min_{\beta \in \mathbb{R}^q} \frac{1}{2} \|y - f(\cdot|U\boldsymbol{\beta})\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(U\boldsymbol{\beta}) + w \|U\boldsymbol{\beta}\|^2 \right)$$

Suppose that $P_j(U\boldsymbol{\beta})$ and $f(\cdot|U\boldsymbol{\beta})$ are continuously differentiable along the directions in U. Then there is a $0 < W < W'$ such that $\hat{\boldsymbol{\theta}}_w = U\hat{\boldsymbol{\beta}}_w$ is a continuous mapping from $[0, W)$ into some open neighborhood $B \subseteq \mathbb{R}^p$.

2 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

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

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

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

and

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

Plugging in the inequality, we get

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