Training/Validation Split Theorem

October 10, 2016

We are interested in bounding the error of the selected model when tuning penalty parameters by a training validation split. We will concern ourselves with the error over the observed covariates in the validation set. Under sufficient entropy conditions, the error of the selected model will converge to the error of the oracle.

We will suppose that the data is generated from the model:

$$y = g^*(x) + \epsilon$$

where ϵ are independent, sub-gaussian errors. The penalized regression models are

$$\hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g \in \mathcal{G}} L_T(g|\boldsymbol{\lambda})$$

Let the model class after fitting on the training data be

$$\mathcal{G}(T) = \{\hat{g}(\cdot|\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Lambda\}$$

The selected penalty parameters are

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \|y - \hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2$$

Suppose the "oracle" penalty parameters are

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V$$

We will provide sharp oracle inequalities of the form

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + \delta$$

The document is organized as follows

1. Theorem 3 proves the training/validation split under general entropy conditions.
2. Theorem 1 applies Theorem 3 to the special case when the fitted functions are Lipschitz in the penalty parameters

1 Theorem 3

Suppose that if $\|\epsilon\|_T \leq 2\sigma$, then $\mathcal{G}(T)$ satisfies the entropy condition

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi_T(R)$$

Furthermore, suppose that

$$\frac{\psi_T\left(a+u\right)}{u^2}$$

is nonincreasing wrt to u for all $u, a > 0$ such that $a + u > \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$.
Then there is some constant C (only dependent on the characteristics of the sub-gassian errors) such that for all $\delta > 0$ such that

$$\sqrt{n_V}\delta^2 \geq 2C\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2\delta\right) \vee \left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2\delta\right)\right]$$

then

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \geq \delta \wedge \|\epsilon\|_V \leq 2\sigma\right) \quad \leq \quad c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right) + Pr\left(\|\epsilon\|_T \leq 2\sigma\right)$$

for a constant c.

Proof

The basic inequality gives us

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 + 2\left\langle\epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V$$

Note that since $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \leq \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V$, then

$$\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right)^2 \quad \leq \quad \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2$$

By a peeling argument, we have

$$\begin{aligned}
Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\geq\delta\right) &= \sum_{s=0}^{\infty}Pr\left(2^s\delta\leq\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\leq 2^{s+1}\delta\right)\\
&\leq \sum_{s=0}^{\infty}Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\geq 2^s\delta\wedge\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\\
&= \sum_{s=0}^{\infty}Pr\left(\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\right)^2\geq 2^{2s}\delta^2\wedge\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\\
&\leq \sum_{s=0}^{\infty}Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V^2-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2\geq 2^{2s}\delta^2\wedge\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\\
&\leq \sum_{s=0}^{\infty}Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta}\left\langle\epsilon,\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V\geq 2^{2s-1}\delta^2\right)
\end{aligned}$$

To apply the lemma based on vandegeer corollary 8.3 (see below), we must check all the conditions are satisfied.
We choose δ such that

$$\begin{aligned}
\frac{\sqrt{n_V}}{8} &\geq \frac{C}{4\delta^2}\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2\delta\right)\vee\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2\delta\right)\right]\\
&\geq \frac{C}{2^{2s+2}\delta^2}\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\vee\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2\delta\right)\right]
\end{aligned}$$

where the second line follows from the assumption that $\psi_T(a+u)/u^2$ is nonincreasing wrt u. Hence we have satisfied the condition in corollary 8.3. So for all $s=0,1,...$ since

$$\sqrt{n_V}2^{2s-1}\delta^2\geq C\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\vee\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\right]$$

we have

$$Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta}\left\langle\epsilon,\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V\geq 2^{2s-1}\delta^2\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\epsilon\|_T\leq 2\sigma\right)\leq\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4C^2\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)^2}\right)$$

Hence we have

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\geq\delta\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\epsilon\|_T\leq 2\sigma\right)\leq C\sum_{s=0}^{\infty}\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4C^2\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)^2}\right)$$

$$\leq C\sum_{s=0}^{\infty}\exp\left(-n_V\frac{2^{4s-2}\delta^4}{64C^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2}\right)\vee\exp\left(-n_V\frac{2^{2s}\delta^2}{196C^2}\right)$$

$$\leq c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2}\right)+c\exp\left(-\frac{n_V\delta^2}{c^2}\right)$$

for some constant c.

Hence we have found for the given δ choice, we have

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\geq\delta\wedge\|\epsilon\|_V\leq 2\sigma\right)\leq c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2}\right)+c\exp\left(-\frac{n_V\delta^2}{c^2}\right)+Pr\left(\|\epsilon\|_T\leq 2\sigma\right)$$

2 Theorem 1

Let $\Lambda=[n_V^{-t_{min}},n_V^{t_{max}}]^J$.

Suppose that if $\|\epsilon\|_T\leq 2\sigma$, there are constants C,κ such that for any $u>0$, we have for all $\lambda\in\Lambda$

$$\left\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})-\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)})\right\|_V\leq Cn^{\kappa}\|\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2\|$$

Then there are constants c,c_1,c_2 s.t. with high probability,

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V\leq\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+c_1\left(\frac{J\left(1+t_{max}+\kappa\right)\log n_V+c_4}{n_V}\right)^{1/2}+c_2\sqrt{\left(\frac{J\left(1+t_{max}+\kappa\right)\log n_V+c_4}{n_V}\right)^{1/2}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V}$$

Proof

1. Determine entropy bound and properties

Under the given Lipschitz condition, a δ-cover for Λ is a $Cn^{\kappa}\delta$-cover for $\mathcal{G}(T)$. We can therefore calculate a covering number for $\mathcal{G}(T)$ wrt $\|\cdot\|_V$ by using the covering number for Λ.

$$N\left(u,\mathcal{G}(T),\|\cdot\|_V\right)\leq N\left(\frac{u}{Cn^{\kappa}},\Lambda,\|\cdot\|_2\right)$$

By Lemma param_covering_cube, we know that

$$\begin{aligned} N\left(u,\Lambda,\|\cdot\|_2\right) &\leq \frac{1}{C_J}\left(\frac{4\left(\lambda_{max}-\lambda_{min}\right)+2\frac{u}{Cn^{\kappa}}}{\frac{u}{Cn^{\kappa}}}\right)^J \\ &= \frac{1}{C_J}\left(\frac{4\left(n_V^{t_{max}}-n_V^{-t_{min}}\right)Cn^{\kappa}+2u}{u}\right)^J \\ &\leq \frac{1}{C_J}\left(\frac{4Cn_V^{t_{\max}+\kappa}+2u}{u}\right)^J \end{aligned}$$

Hence

$$H(u,\mathcal{G}(T),\|\cdot\|_V) \leq \log\left[\frac{1}{C_J}\left(\frac{4Cn_V^{t_{\max}+\kappa}+2u}{u}\right)^J\right]$$

Then

$$\begin{aligned} \int_0^R H^{1/2}(u,\mathcal{G}(T),\|\cdot\|_V)du &\leq \int_0^R\left[\log\frac{1}{C_J}+J\log\left(\frac{4Cn_V^{t_{\max}+\kappa}+2u}{u}\right)\right]^{1/2}du \\ &\leq \int_0^R\left[\log\frac{1}{C_J}+J\log 4+J\log\left(\frac{8Cn_V^{t_{max}+\kappa}}{u}\right)\right]^{1/2}du \\ &= R\int_0^1\left[\log\frac{1}{C_J}+J\log 4+J\log\left(\frac{8Cn_V^{t_{max}+\kappa}}{Rv}\right)\right]^{1/2}dv \\ &\leq R\left[\int_0^1\log\frac{1}{C_J}+J\log 4+J\log\left(\frac{8Cn_V^{t_{max}+\kappa}}{R}\right)+J\log\frac{1}{v}dv\right]^{1/2} \\ &= R\left[\log\frac{1}{C_J}+J(1+\log 4+\log 8C)+J\log\left(n_V^{t_{max}+\kappa}\right)+J\log\frac{1}{R}\right]^{1/2} \end{aligned}$$

The second bound is crazy loose but I think it is okay. It comes from the fact that

$$\log\left(a+b\right)<\log\left(2a\right)+\log(2b)$$

The third inequality follows from concavity of the square root.

2. Apply Theorem 3

By Theorem 3, there is a constant $C_0 > 0$ such that the oracle inequality

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + \delta$$

holds with high probability for all $\delta > 0$ such that

$$\sqrt{n_V}\delta^2 \geq 2CR\left(\left[\log\frac{1}{C_J} + J(1+\log 4+\log 8C) + J\log\left(n_V^{t_{max}+\kappa}\right) + J\log\frac{1}{R}\right]^{1/2} \vee 1\right) \tag{1}$$

where $R = 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2\delta$.

Solving for δ is hard because of the $\log\frac{1}{R}$ term. Suppose we constrain ourselves to a choice of $\delta > 1/n_V$ (which is true for sufficiently large sample sizes since δ is on the order of $n_V^{-1/2}$ for parametric problems). Then

$$R = 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2\delta \geq 2/n_V$$

Therefore $\log\frac{1}{R} < \log n_V$. This allows us to replace the term $\log\frac{1}{R}$ with $\log n_V$.
Then we need to find a $\delta > 1/n_V$ such that

$$\sqrt{n_V}\delta^2 \geq 2C_0\left(\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + \delta\right)\left[\left[\log\frac{1}{C_J} + J(1+\log 4+\log 8C) + J\log\left(n_V^{t_{max}+\kappa}\right) + J\log n_V\right]^{1/2} \vee 1\right] \tag{2}$$

Now we can solve for δ. Let

$$\begin{aligned} K &= 2C_0\left(\left[\log\frac{1}{C_J} + J(1+\log 4+\log 8C) + J\log\left(n_V^{t_{max}+\kappa}\right) + J\log n_V\right]^{1/2}\right) \vee 1 \\ &= 2C_0\left(\left[\log\frac{1}{C_J} + J(1+\log 4+\log 8C) + J\left(1+t_{max}+\kappa\right)\log n_V\right]^{1/2}\right) \vee 1 \end{aligned}$$

and

$$\omega = \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$$

(2) can be expressed as

$$\sqrt{n_V}\delta^2 - K\delta - K\omega \geq 0$$

We notice that (2) is precisely the quadratic inequality and is satisfied for δ such that

$$\delta \geq \frac{K + \sqrt{K^2 + 4K\omega\sqrt{n_V}}}{2\sqrt{n_V}}$$

3. Re-writing the condition on δ to be more intuitive

For a more intuitive understanding of this lower bound for δ, we can use a slightly bigger lower bound:

$$\delta \geq c_1 \frac{K}{\sqrt{n_V}} + c_2 \sqrt{\frac{K}{n_V}\omega}$$

for universal constants $c_1 = 1 + \frac{1}{\sqrt{2}}$ and $c_2 = 2\sqrt{2}$.

To see why this holds, note that there are two cases:

Case 1. $K > 4\omega\sqrt{n_V}$

$$\begin{aligned} \frac{K + \sqrt{K^2 + 4K\omega\sqrt{n_V}}}{2\sqrt{n_V}} &\leq \frac{K + \sqrt{2K^2}}{2\sqrt{n_V}} \\ &\leq \frac{\left(1 + \sqrt{2}\right) K}{2\sqrt{n_V}} \end{aligned}$$

Case 2. $K < 4\omega\sqrt{n_V}$

$$\begin{aligned} \frac{K + \sqrt{K^2 + 4K\omega\sqrt{n_V}}}{2\sqrt{n_V}} &\leq \frac{K + \sqrt{8K\omega\sqrt{n_V}}}{2\sqrt{n_V}} \\ &\leq \frac{K}{2\sqrt{n_V}} + \sqrt{\frac{8K}{\sqrt{n_V}}\omega} \end{aligned}$$

Plug this inequality back into (1) to gives us the final result.

Theorem 3 gives us that with high probability

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + c_1 \frac{2C_0 \left([c + J\left(1 + t_{max} + \kappa\right)\log n_V]^{1/2}\right) \vee 1}{\sqrt{n_V}} + c_2 \sqrt{\frac{2C_0 \left([c + J\left(1 + t_{max} + \kappa\right)\log n_V]^{1/2}\right) \vee 1}{n_V} \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V}$$

where $c = \log \frac{1}{C_J} + J(1 + \log 4 + \log 8C)$

3 Appendix

Lemma Vandegeer (Based on Vandegeer Corollary 8.3)

(This lemma is directly out of Vandegeer's Empirical Process book.)

Let Q_m be the empirical distributon of m observations at covariates x_i.

Suppose ϵ are m independent sub-gaussian errors. Suppose the model class $\mathcal{F}(T)$ has elements $\sup_{f \in \mathcal{F}_n(T)} \|f\|_{Q_m} \leq R$ and satisfies

$$\psi_T(R) \geq \int_0^R H^{1/2}(u, \mathcal{F}(T), \|\cdot\|_{Q_m}) du$$

There is C dependent only on the sub-gaussian constants such that for all $\delta > 0$ such that

$$\sqrt{m}\delta \geq C(\psi_T(R) \vee R)$$

we have

$$Pr\left(\sup_{f \in \mathcal{F}_n(T)} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i f(x_i) \right| \geq \delta \wedge \|\epsilon\|_{Q_m} \leq \sigma \right) \leq C \exp\left(-\frac{m\delta^2}{4C^2R^2}\right)$$

Lemma param_covering_cube

Suppose $\Lambda = [\lambda_{min}, \lambda_{max}]^J$. Then the δ-covering number is bounded as follows

$$N(\delta, \Lambda, \|\cdot\|_2) \leq \frac{1}{C_J} \left(\frac{4(\lambda_{max} - \lambda_{min}) + 2\delta}{\delta} \right)^J$$

where $C_J = \frac{\text{volume of ball of radius } \rho}{\rho^J}$.

Proof

(Essentially the same proof as that for Lemma 2.5 in vandegeer)

Let $C = \{c_j\}_{j=1}^N \subset \Lambda$ be the largest set s.t. two distinct points c_{j_1}, c_{j_2} are at least δ apart. Then balls with radius δ centered at C cover Λ. Hence

$$N(\delta, \Lambda, \|\cdot\|_2) \leq N$$

If we instead consider the balls centered at C but with radius $\delta/4$, all of these smaller balls must be disjoint and are completely contained in the box $\Lambda_{bigger} = [\lambda_{min} - \delta/4, \lambda_{max} + \delta/4]^J$. So we know the aggregate volume of these smaller balls is less than the volume of Λ_{bigger}.

Hence

$$NC_J(\delta/4)^J \leq (\lambda_{max} - \lambda_{min} + \delta/2)^J$$