Cross Validation and the 1se rule

Rob McCulloch

1. Cross Validation
2. 1se

1. Cross Validation

Let's consider the basic setup where we have a machine learning approach that, given training data, enables us to compute

$$\hat{f}(x, \theta)$$

Here, θ is a *tuning parameter of the approach*.
For example, with the LASSO, θ is λ.

The idea is that given x,
$\hat{f}(x, \theta)$ would be our prediction for y.

Let's use squared error loss.

Given training data and θ, and an x we want to predict at, our loss is

$$(y - \hat{f}(x, \theta))^2$$

How do we choose θ?

Cross validation is a basic approach.

We divide the training data up into K folds.

Let F_k be the set of indices for the k^{th} fold, $k = 1, 2, \ldots, K$.

That is

$$\cup_{k=1}^{K} F_k = \{1, 2, \ldots, n\}, \quad F_k \cap F_l = \emptyset, k \neq l.$$

where n is the number of observations in the training data.

Now let $\hat{f}^{-k}(x, \theta)$ be the f learned from the data using all of the folds *except* the k^{th}.

Then our CV estimate of our out of sample loss at the tuning parameter θ is

$$\hat{L}(\theta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} (y_i - \hat{f}^{-k}(x_i, \theta))^2.$$

Given a discrete set of possible θ values: $\theta \in \{\theta_c\}_{c=1}^{C}$, we could choose the value of θ that make $\hat{L}(\theta)$ the smallest.

2. 1se

The CV estimate of our loss is just a simple approach based on averaging.

How can we can a sense of how big the error might be?

Let's assume that the number of folds divides nicely into n, the training data sample size.

Then let $m = n/K$ be the number of observations in each fold.

Then $n = mK$ so that

$$\begin{aligned}\hat{L}(\theta) &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i\in F_k}(y_i - \hat{f}^{-k}(x_i,\theta))^2 \\ &= \frac{1}{K}\sum_{k=1}^{K} CV_k(\theta)\end{aligned}$$

Hence we can view our loss estimate as the average of the $CV_k(\theta)$ values and use the associated standard error to quantify possible error in estimation:

$$SE(\theta) = \frac{1}{\sqrt{K}} sd(\theta)$$

where $sd(\theta)$ is the sample standard deviation of the $CV_k(\theta)$ values,

$$sd(\theta) = \sqrt{\frac{1}{K-1} \sum (CV_k(\theta) - \hat{L}(\theta))^2}.$$

1se rule

The 1se rule suggests using the simplest model that has out of sample loss that could be similar to that of the model with the lowest estimated loss.

So, suppose we can order the θ_c so that as c increases, we expect a simpler model.

For example, in the case of the LASSO, if "θ is λ", then bigger θ would correspond to a large regularization penalty, and hence a simpler model.

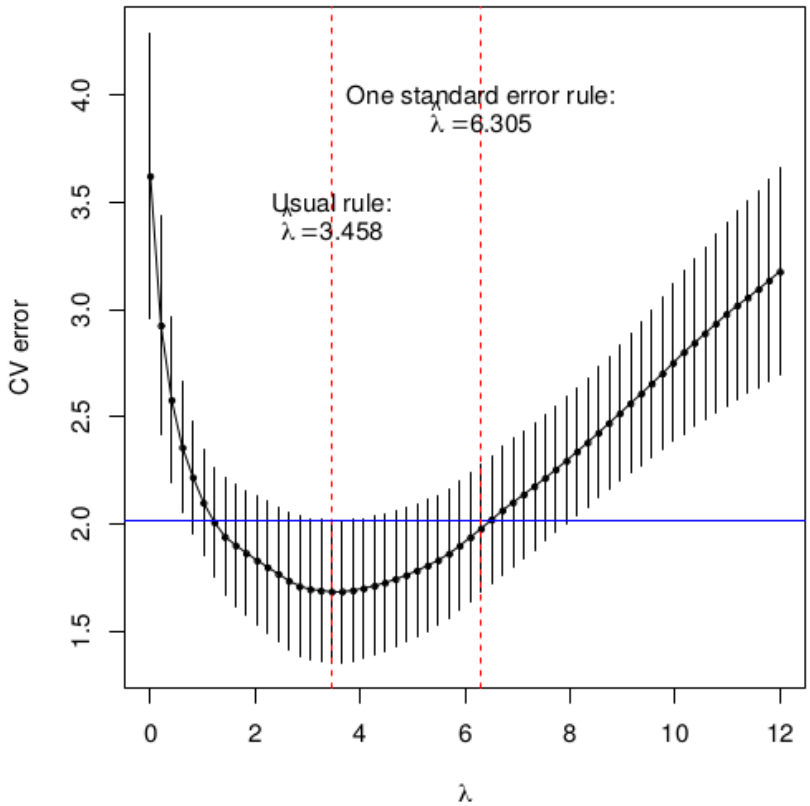
Then we choose the largest θ_c such that

$$CV(\theta_c) < CV(\theta^*) + SE(\theta^*)$$

where θ^* gives the mimimal $\hat{L}(\theta)$.

We use the simplest model whose loss estimate is within 1se of the smallest loss.

For the LASSO, it looks like this:



Increase λ from the loss mimimizer as long as you are within 1 se of the miminmal loss.

Increase because bigger λ means simpler models.

Clearly, the $SE(\theta)$ esimate is very rough.

It is base on just K "observations" and they are not independent.

Hence, the ad-hoc 1se rule.

A nice example of a simple but very useful idea that *very approximately* solves a tough problem.

Better a rough answer to the right question, than a precise answer to the wrong question !!!