MATH 680 Computation Intensive Statistics

September 16, 2019

Cross Validation

Contents

1 Prediction rules

Prediction problem typically begin with training set consist of N pairs

$$\mathbf{D} = \{(x_i, y_i), i = 1, \ldots, N\},$$

where $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. Based on this training data set, a prediction rule $r_{\mathbf{D}}(x)$ is constructed such that a prediction $\hat{y}$ is produced for any point $x \in \mathcal{X}$,

$$\hat{y} = r_{\mathbf{D}}(x), \qquad x \in \mathcal{X}.$$

2 Algorithm

K-fold cross-validation uses part of the data to fit the model and a different part to test it.

1. Split the data into K roughly equal sizes parts $K = 5$.

2. For $k = 1, \ldots, K$ repeat Step (a)–(b):

   (a) We remove the k-th part $\mathbf{D}_k$ from the data $\mathbf{D}$, and denote the remaining $k - 1$ parts of the data as $\mathbf{D}(k)$. We fit the model to $\mathbf{D}(k)$ and denote the corresponding model we obtained by $r_{\mathbf{D}(k)}$.

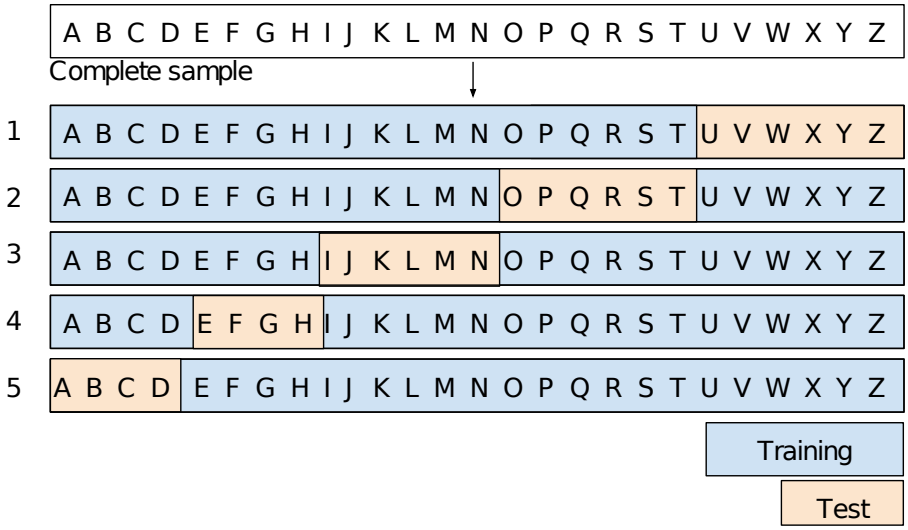
   (b) Calculate the prediction error of the fitted model $r_{\mathbf{D}(k)}(\cdot)$ when predicting on the k-th part of the data $\mathbf{D}_k$

$$\text{CV}_k = \sum_{i \in \mathbf{D}_k} L(y_i, r_{\mathbf{D}(k)}(x_i))$$

3. Then the **cross-validation estimate** of prediction error is

$$\widehat{\mathrm{Err}}_{\mathrm{cv}} = \frac{1}{N}\sum_{k=1}^{K} \mathrm{cv}_k = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in \mathbf{D}_k} L(y_i, r_{\mathbf{D}(k)}(x_i)).$$

If we are given M models $r^1, r^2, \ldots, r^M$ to choose from, we use cross-validation to compute $\widehat{\mathrm{Err}}_{\mathrm{cv}}(r^1), \widehat{\mathrm{Err}}_{\mathrm{cv}}(r^2), \ldots, \widehat{\mathrm{Err}}_{\mathrm{cv}}(r^M)$ and choose the model that return the smallest $\widehat{\mathrm{Err}}_{\mathrm{cv}}$.



3 Methodology

Question: having chosen a particular rule, how do we estimate its predictive accuracy?

Two quite distinct approaches to prediction error assessment developed in the 1970s. A narrower (but more efficient) model-based approach was the first, emerging in the form of Mallows' Cp estimate and the Akaike information criterion (AIC). The second, depending on the classical technique of cross-validation, was fully general and nonparametric.

3.1 Prediction error

We want to access the accuracy of $r(\cdot)$. In practice there are usually several competing rules

$$r^1, r^2, \ldots, r^M$$

under consideration and the main question is determining which is best. Quantifying the prediction error of $r_{\mathbf{D}}$ requires specification of the discrepancy $L(y, \hat{y})$ between a prediction $\hat{y}$ and the actual response y. The two most common choices are ***squared error***

$$L(y, \hat{y}) = (y - \hat{y})^2$$

for regression and ***classification error***

$$L(y, \hat{y}) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{if } y = \hat{y} \end{cases}$$

For error estimation assume that pairs (x_i, y_i) in the training set are obtained by random sampling from some probability distribution F

$$(x_i, y_i) \overset{\text{iid}}{\sim} F \qquad \text{for } i = 1, 2, \ldots, N.$$

The ***true error rate*** $\mathrm{Err}_{\mathbf{D}}$ of rule $r_{\mathbf{D}}(x)$ is the expected discrepancy

$$\begin{aligned}\mathrm{Err}_{\mathbf{D}} &= E_F[L(y_0, \hat{y}_0)] \\ &= E_F[L(y_0, r_{\mathbf{D}}(x_0))]\end{aligned}$$

where the expectation is taken over a new pair (x_0, y_0) drawn from F independently of $\mathbf{D}$. Here $\mathbf{D}$ is held fixed in expectation, only (x_0, y_0) varying.

3.2 Validation error

We want to estimate $\mathrm{Err}_{\mathbf{D}}$. How about we use the ***apparent error (in-sample error)***.

$$\mathrm{err}_{\mathbf{D}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_i),$$

$\mathrm{err}_{\mathbf{D}}$ usually underestimates $\mathrm{Err}_{\mathbf{D}}$ since $r_{\mathbf{D}}(x)$ has been constructed to fit $\{(x_i, y_i)\}_{i=1}^{N}$.

The ideal remedy is to have an independent ***validation set*** (or test set) $\mathbf{D}_{\mathrm{val}}$:

$$\mathbf{D}_{\mathrm{val}} = \{(x_{0j}, y_{0j}),\ j = 1, 2, \ldots, N_{\mathrm{val}}\}.$$

This would provide as an unbiased estimator of $\mathrm{Err}_{\mathbf{D}}$.

$$\begin{aligned}\widehat{\mathrm{Err}}_{\mathrm{val}} &= \frac{1}{N_{\mathrm{val}}}\sum_{i=1}^{N_{\mathrm{val}}} L(y_{0j}, \hat{y}_{0j}) \\ &= \frac{1}{N_{\mathrm{val}}}\sum_{i=1}^{N_{\mathrm{val}}} L(y_{0j}, r_{\mathbf{D}}(x_{0j}))\end{aligned} \tag{1}$$

It is unbiased since

$$\begin{aligned} E_F\left[\widehat{\mathrm{Err}}_{\mathrm{val}}\right] &= E_F\left[\frac{1}{N_{\mathrm{val}}}\sum_{i=1}^{N_{\mathrm{val}}} L(y_{0j}, \hat{y}_{0j})\right] \\ &= \frac{1}{N_{\mathrm{val}}}\sum_{i=1}^{N_{\mathrm{val}}} E\left[L(y_{0j}, r_{\mathbf{D}}(x_{0j}))\right] \\ &= \frac{1}{N_{\mathrm{val}}}\sum_{i=1}^{N_{\mathrm{val}}} \mathrm{Err}_{\mathbf{D}} \\ &= \mathrm{Err}_{\mathbf{D}} \end{aligned}$$

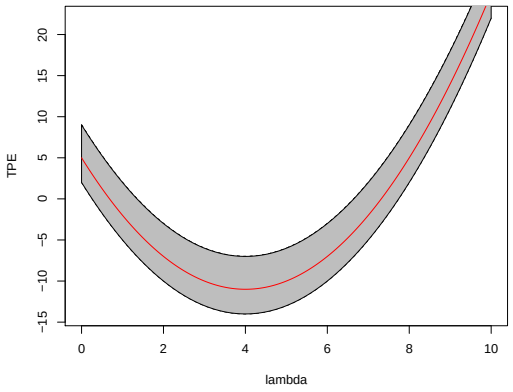


Figure 1: $\widehat{\mathrm{Err}}_{\mathrm{val}}$ is an unbiased estimator of $\mathrm{Err}_{\mathbf{D}}$.

3.3 Cross-validation error

Cross-validation attemps to mimic $\widehat{\mathrm{Err}}_{\mathrm{val}}$ without the need for a validation set. Define $\mathbf{D}(i)$ to be the reduced training set which the i-th pair (x_i, y_i) has been removed. Let $r_{\mathbf{D}}(i)(.)$

indicate the rule constructed on $\mathbf{D}(i)$. The ***cross-validation estimate*** of prediction error is

$$\begin{aligned}\widehat{\mathrm{Err}}_{\mathrm{cv}} &= \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_{(i)}) \\ &= \frac{1}{N}\sum_{i=1}^{N} L(y_i, r_{\mathbf{D}(i)}(x_i)),\end{aligned}$$

Compared with (1), now the i-th pair (x_i, y_i) is not involved in the construction of the prediction rule for y_i. $\widehat{\mathrm{Err}}_{\mathrm{cv}}$ is the "**leave one-out**" cross-validation.

4 What value should we choose for K?

It is interesting to wonder about what quantity K-fold cross-validation estimates.

- With $K = N$, the cross-validation estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because the N "training sets" $\mathbf{D}(i)$ are so similar to one another. The computational burden is also considerable, requiring N applications of the learning method.

- On the other hand, with $K = 5$ say, cross-validation has lower variance. But bias could be a problem, depending on how the performance of the learning method varies with the size of the training set.

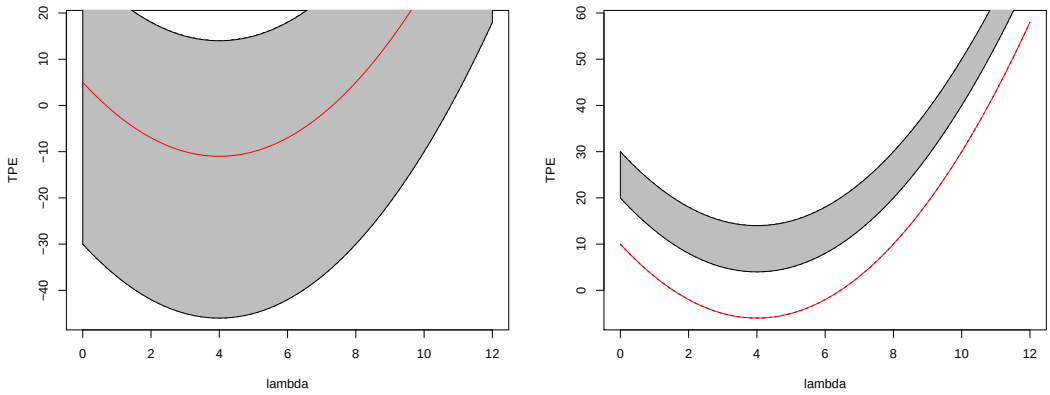


Figure 2: Left: $K = N$, almost unbiased but high variance. Right: $K = 5$, larger bias but small variance.

Figure 3 shows a hypothetical "learning curve" for a classifier on a given task, a plot of $\mathrm{Err}_{\mathbf{D}}$ versus the size of the training set N.

- For yellow curve, the performance of the classifier improves as the training set size increases to 50 observations; increasing the number further to 200 brings only a small benefit.

- If our training set had 200 observations, fivefold cross-validation would estimate the performance of our classifier over training sets of size 160, which from Figure 3 is virtually the same as the performance for training set size 200. Thus cross-validation would not suffer from much bias.

- However if the training set had 50 observations, fivefold cross-validation would estimate the performance of our classifier over training sets of size 40, and from the figure that would be an overestimate of $\mathrm{Err}_{\mathbf{D}}$. Hence as an estimate of $\mathrm{Err}_{\mathbf{D}}$, cross-validation would be biased upward.

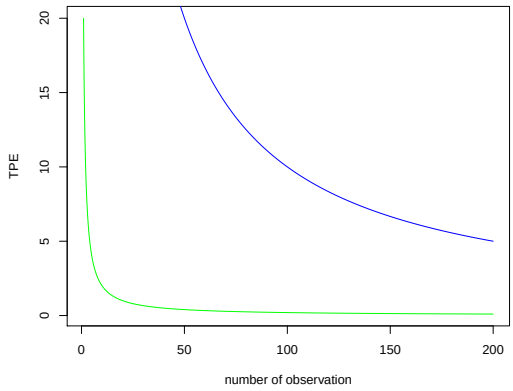


Figure 3: Hypothetical learning curve for a classifier on a given task: a plot of the true prediction error $\mathrm{Err}_{\mathbf{D}}$ versus the size of the training set N. With a dataset of 200 observations, 5-fold cross-validation would use training sets of size 160, which would behave much like the full set for the yellow curve. However, for the blue curve, this would result in a considerable overestimate of the true prediction error.

- If the classifier corresponds to the blue curve, fivefold cross-validation with training sets of size 160 would also suffer from bias.