Cross Validation

1 Prediction rules

Prediction problem typically begin with training set consist of N pairs

$$\mathcal{T} = \{(X_i, Y_i), i = 1, \ldots, N\},$$

where $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$. Based on this training data set, a prediction rule $\hat{f}_{\mathcal{T}}(X)$ is constructed such that a prediction $\hat{f}$ is produced for any point $X \in \mathcal{X}$,

$$\hat{Y} = \hat{f}_{\mathcal{T}}(X), \qquad X \in \mathcal{X}.$$

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 $\mathcal{T}_k$ from the data $\mathcal{T}$, and denote the remaining $k - 1$ parts of the data as $\mathcal{T}(k)$. We fit the model to $\mathcal{T}(k)$ and denote the corresponding model we obtained by $\hat{f}_{\mathcal{T}(k)}$.
   (b) Calculate the total prediction error of the fitted model $\hat{f}_{\mathcal{T}(k)}(\cdot)$ when predicting

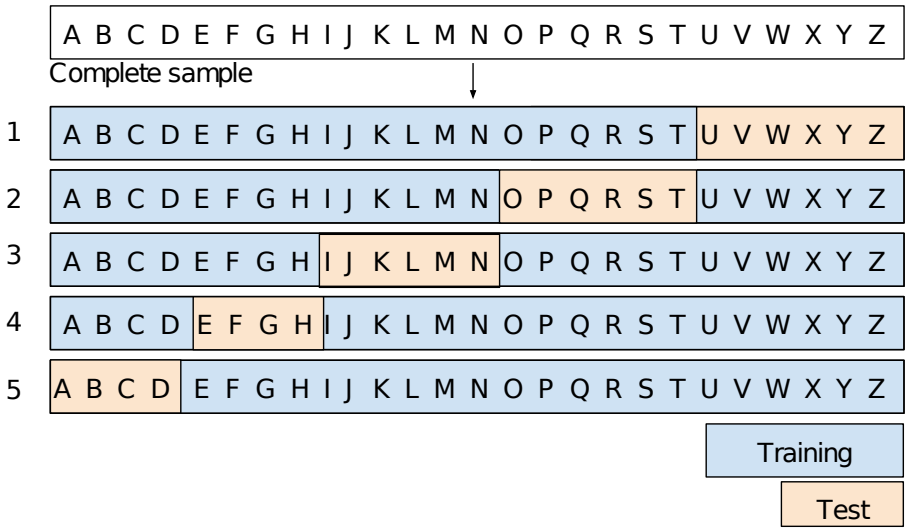
on the k-th part of the data $\mathcal{T}_k$

$$\text{cv}_k = \sum_{i \in \mathcal{T}_k} L(Y_i, \hat{f}_{\mathcal{T}(k)}(X_i))$$

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

$$\widehat{\text{Err}}_{\text{cv}} = \frac{1}{N} \sum_{k=1}^{K} \text{cv}_k = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} L(Y_i, \hat{f}_{\mathcal{T}(k)}(X_i)).$$

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



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 $\hat{f}(\cdot)$. In practice there are usually several competing rules

$$\hat{f}^1, \hat{f}^2, \ldots, \hat{f}^M$$

under consideration and the main question is determining which is best. Quantifying the prediction error of $\hat{f}_{\mathcal{T}}$ 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 ***test error*** $\mathrm{Err}_{\mathcal{T}}$ of rule $\hat{f}_{\mathcal{T}}(X)$ is the expected discrepancy

$$\mathrm{Err}_{\mathcal{T}} = E_{Y^0, X^0}[L(Y^0, \hat{f}_{\mathcal{T}}(X^0))|\mathcal{T}]$$

where the expectation is taken over a new pair (X^0, Y^0) drawn from F independently of $\mathcal{T}$. Here $\mathcal{T}$ is held fixed in expectation, only (X^0, Y^0) varying.

3.2 Validation error

We want to estimate $\mathrm{Err}_{\mathcal{T}}$. How about we use the ***training error***

$$\overline{\mathrm{err}} = \frac{1}{N} \sum_{i=1}^{N} L(Y_i, \hat{f}_{\mathcal{T}}(X_i)),$$

$\overline{\mathrm{err}}$ usually underestimates $\mathrm{Err}_{\mathcal{T}}$ since $\hat{f}_{\mathcal{T}}(X)$ 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) $\mathcal{T}_{\text{val}}$:

$$\mathcal{T}_{\text{val}} = \{(X_i^0, Y_i^0),\ i = 1, 2, \ldots, N_{\text{val}}\}.$$

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

$$\begin{aligned}\widehat{\mathrm{Err}}_{\mathrm{val}} &= \frac{1}{N_{\mathrm{val}}} \sum_{i \in \mathcal{T}_{\mathrm{val}}} L(Y_i^0, \hat{Y}_{0j}) \\ &= \frac{1}{N_{\mathrm{val}}} \sum_{i \in \mathcal{T}_{\mathrm{val}}} L(Y_i^0, \hat{f}_{\mathcal{T}}(X_i^0)) \end{aligned} \tag{1}$$

It is unbiased since

$$\begin{aligned} E_{Y^0, X^0}\left[\widehat{\mathrm{Err}}_{\mathrm{val}}\right] &= \frac{1}{N_{\mathrm{val}}} \sum_{i \in \mathcal{T}_{\mathrm{val}}} E\left[L(Y_i^0, \hat{f}_{\mathcal{T}}(X_i^0))\right] \\ &= \frac{1}{N_{\mathrm{val}}} \sum_{i \in \mathcal{T}_{\mathrm{val}}} \mathrm{Err}_{\mathcal{T}} \\ &= \mathrm{Err}_{\mathcal{T}} \end{aligned}$$

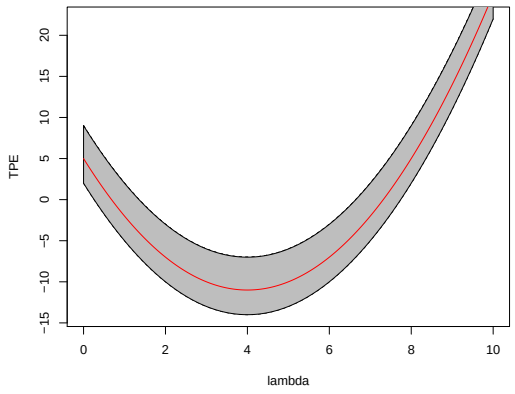


Figure 1: $\widehat{\mathrm{Err}}_{\mathrm{val}}$ is an unbiased estimator of $\mathrm{Err}_{\mathcal{T}}$.

3.3 Cross-validation error

Cross-validation attempts to mimic $\widehat{\text{Err}}_{\text{val}}$ without the need for a separate validation set. Define $\mathcal{T}(i)$ to be the reduced training set which the i-th pair (X_i, Y_i) has been removed. Let $\hat{f}_{\mathcal{T}}(i)(.)$ indicate the rule constructed on $\mathcal{T}(i)$. The ***cross-validation estimate*** of prediction error is

$$\begin{aligned}\widehat{\text{Err}}_{\text{cv}} &= \frac{1}{N}\sum_{i=1}^{N} L(Y_i, \hat{Y}_{(i)}) \\ &= \frac{1}{N}\sum_{i=1}^{N} L(Y_i, \hat{f}_{\mathcal{T}(i)}(X_i)),\end{aligned}$$

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. $\widehat{\text{Err}}_{\text{cv}}$ 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" $\mathcal{T}(i)$ 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.

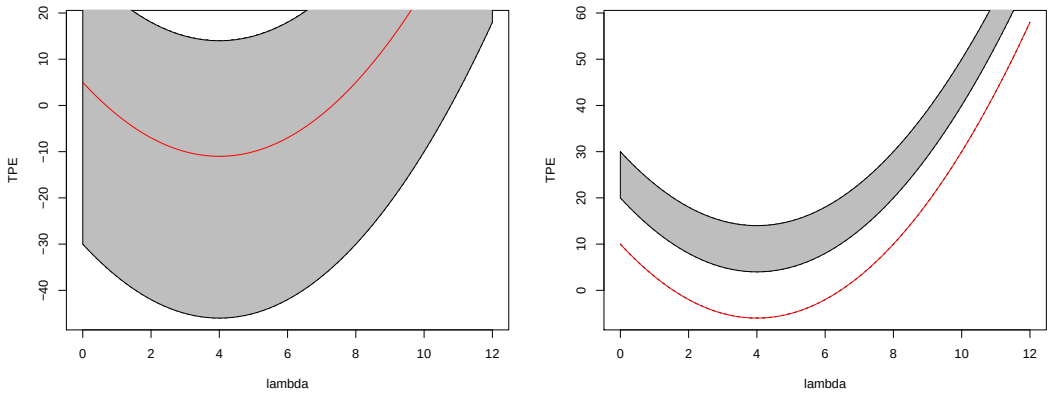


Figure 2: Left: $K = N$, almost unbiased but high variance. Right: $K = 5$, larger bias but small variance.

Figure 3 shows a hypothetical "learning curve" for a classifier on a given task, a plot of $\mathrm{Err}_{\mathcal{T}}$ 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 $\mathrm{Err}_{\mathcal{T}}$. Hence as an estimate of $\mathrm{Err}_{\mathcal{T}}$, cross-validation would be biased upward.

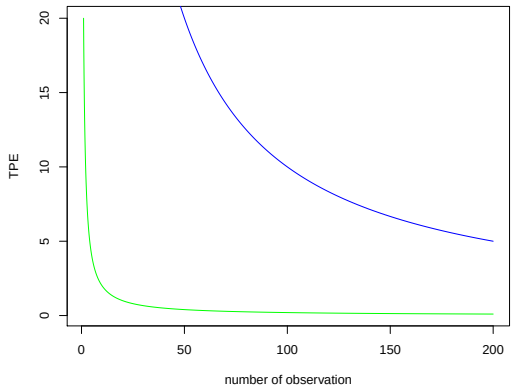


Figure 3: Hypothetical learning curve for a classifier on a given task: a plot of the true prediction error $\mathrm{Err}_{\mathcal{T}}$ versus the size of the training set N. With a dataset of 200 observations, 5-fold cross-validation would use training sets of size 160, which would behave much like 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.