STAT 480 Statistical Computing Applications

Unit 5. Resampling Methods

Lecture 4. Cross-Validation

Department of Statistics, Iowa State University

Spring 2019

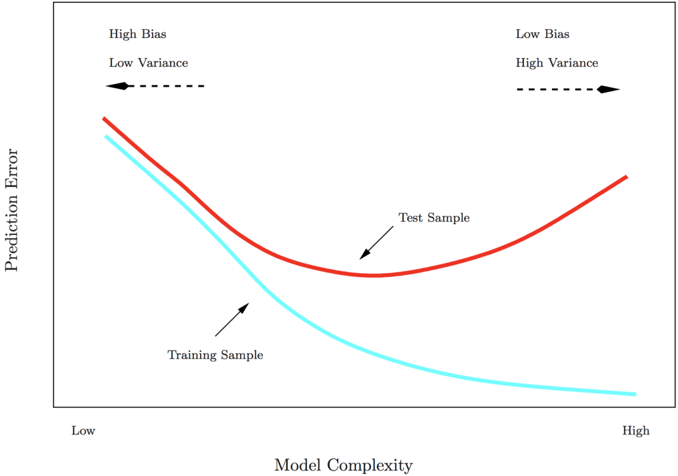
Cross-Validation and Bootstrap

- Cross-validation is another resampling methods.
- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates.

Prediction Error vs. Test Error

- Recall the distinction between the test error (prediction error) and the training error (estimation error):
- The test error is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

Training vs. Testing Performance



Prediction-Error Estimates

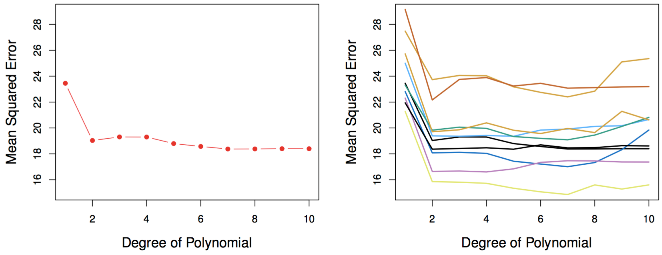
- Best solution: a large designated test set. Often not available!
- Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate. These include the C_p statistic, AIC and BIC.
- Here we instead consider a class of methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

Validation-Set Approach

- Here we randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. This is typically assessed using
  - mean squared error (MSE) for a quantitative response;
  - misclassification rate for a qualitative (discrete) response.

Example: Automobile Data

- We would like to compare linear vs higher-order polynomial terms in a linear regression.
- We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.



Left panel shows single split; right panel shows multiple splits

Drawbacks of Validation Set Approach

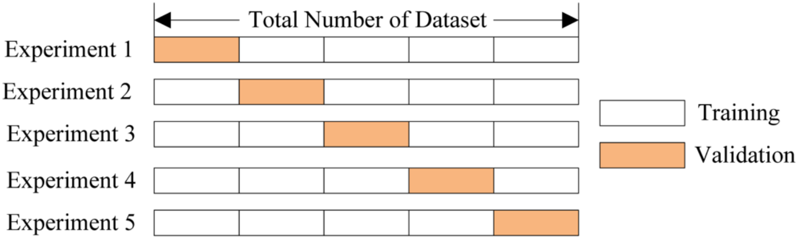
- The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations – those that are included in the training set rather than in the validation set – are used to fit the model.
- This suggests that the validation set error may tend to overestimate the test error for the model fit on the entire data set.

K–fold Cross–Validation

- Widely used approach for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K - 1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \dots, K$ and then the results are combined.

K–fold Cross–Validation (Cont.)

Divide data into K roughly equal-sized parts ($K = 5$ here).



The Details

- Let the K parts be $C_1, C_2, \ldots, C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if n is a multiple of K, then $n_k = n/K$.
- Compute
$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$
where $MSE_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k}$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
- Setting $K = n$ yields n-fold or leave-one out cross-validation (LOOCV).

Special Case

- With least-squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit! The following formula holds:

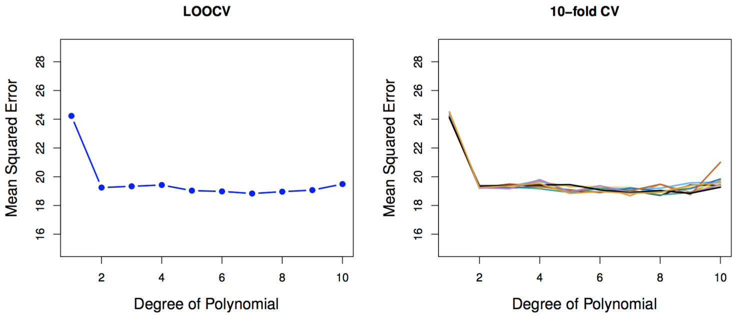
$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2 ,$$

  where $\hat{y}_i$ is the ith fitted value from the orginal least squares fit, and h_i is the leverage (diagonal of the "hat" matrix).
- This is like the ordinary MSE, except the ith residual is divided by $1 - h_i$.

K–fold Cross–Validation (Cont.)

- LOOCV sometimes useful, but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.
- A better choice is $K = 5$ or 10.

Auto Data (Cont.)



Other Issues with Cross–Validation

- Since each training set is only $(K-1)/K$ as big as the original training set, the estimates of prediction error will typically be biased upward.
- This bias is minimized when $K = n$ (LOOCV), but this estimate has high variance, as noted earlier.
- $K = 5$ or 10 usually provide a good compromise for this bias-variance tradeoff.

Cross–Validation for Classification Problems

- We divide the data into K roughly equal-sized parts $C_1, C_2, \ldots, C_K$.
- Compute

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{n} Err_k,$$

where $Err_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

Cross–Validation: Right and Wrong

- Consider a simple classifier applied to some two-class data:
    1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
    2. We then apply a classifier such as logistic regression, using only these 100 predictors.
- How do we estimate the test set performance of this classifier?
- Can we apply cross-validation in step 2, forgetting about step 1?

No!

- This would ignore the fact that in Step 1, the procedure has already seen the labels of the training data, and made use of them. This is a form of training and must be included in the validation process.
- It is easy to simulate realistic data with the class labels independent of the outcome, so that true test error $=50\%$, but the CV error estimate that ignores Step 1 is zero!
- Wrong: Apply cross–validation in step 2.
- Right: Apply cross–validation to steps 1 and 2.