Lecture 11: Resampling Methods (Textbook 5.1)

Nayel Bettache

Department of Statistical Science, Cornell University

What is the Resampling Methods?

In this chapter, we discuss two of the most commonly used resampling methods, **cross-validation** and the **bootstrap**.

- Cross-validation is usually used to perform model selection.
- Bootstrap is used to measure accuracy of a parameter estimate or of a given statistical learning method.

Recap: Test Error and Training 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 is calculated by applying the statistical learning method to the observations used in its training.
- The training error rate often is quite different from the test error rate.
- The test error is a valid measure of model fit.

How to Calculate Test Error Rate?

- Ideally, the test error is calculated based on a test set.
- However, the test data are usually not available in practice. In practice, there are two classes of methods.
- Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate. This includes AIC and BIC in Chapter 6.
- 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.

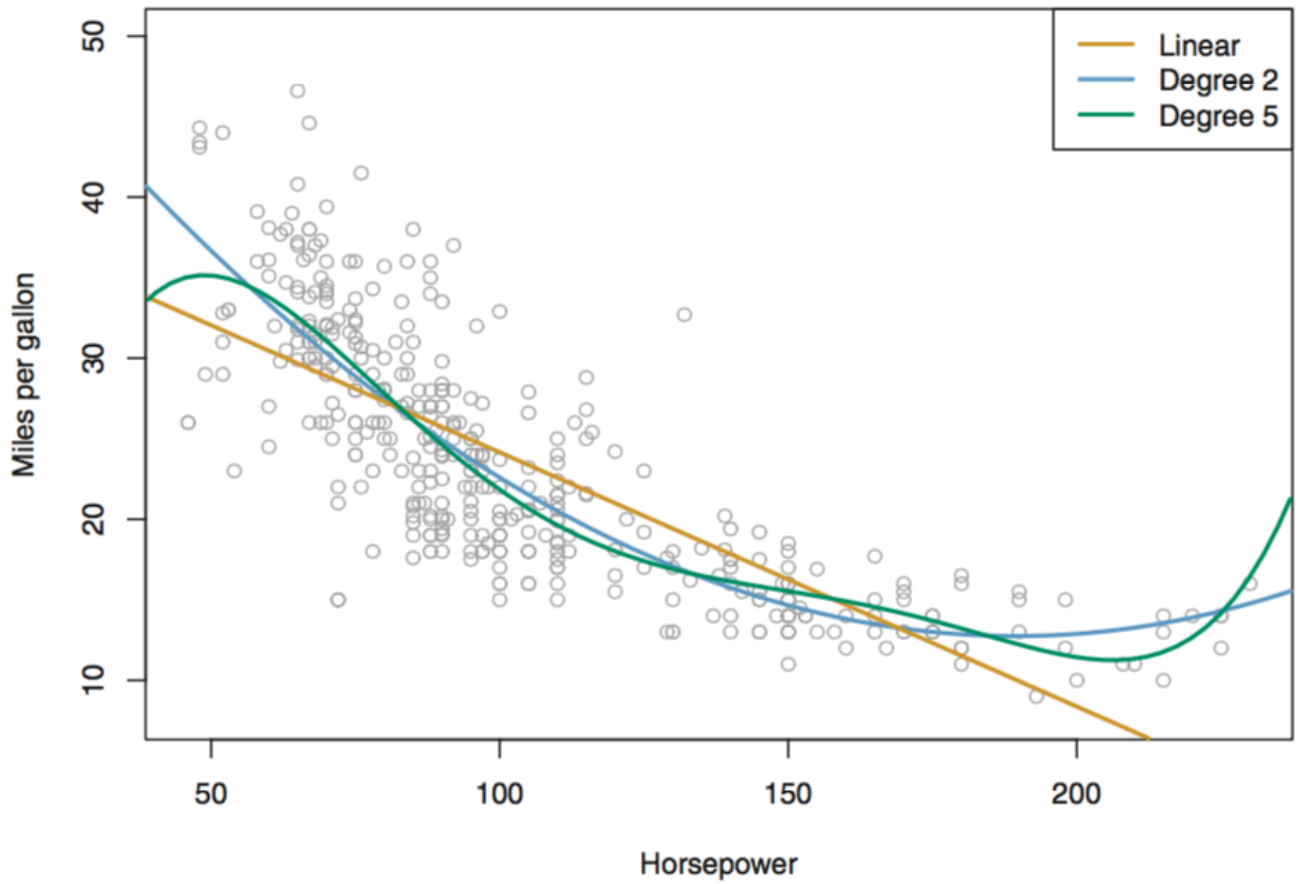
The Validation Set Approach

- 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 MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

A set of n observations are randomly split into a training set (left part) and a validation set (right part).

Example: Auto Data

In Chapter 3, we find there appears to be a non-linear relationship between mpg and horsepower.

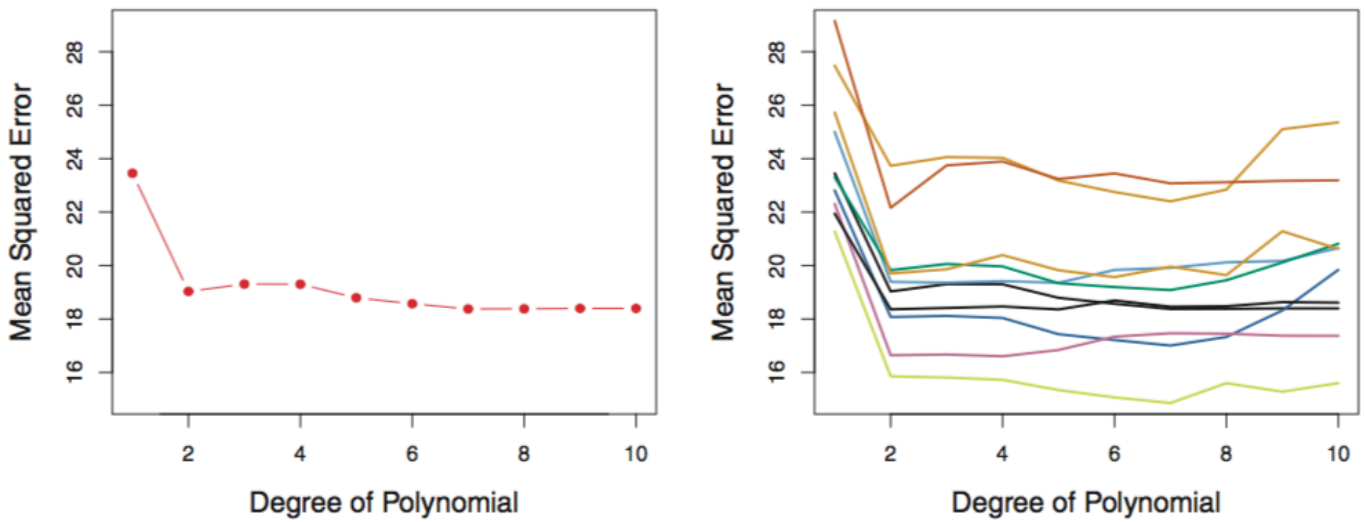


Whether a cubic or higher-order fit might provide a better fit?

Example: Auto Data

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 data.



Left: Validation error estimates for a single split into training and validation data sets. Right: The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. We can see it is not stable.

Drawbacks of Validation Set Approach

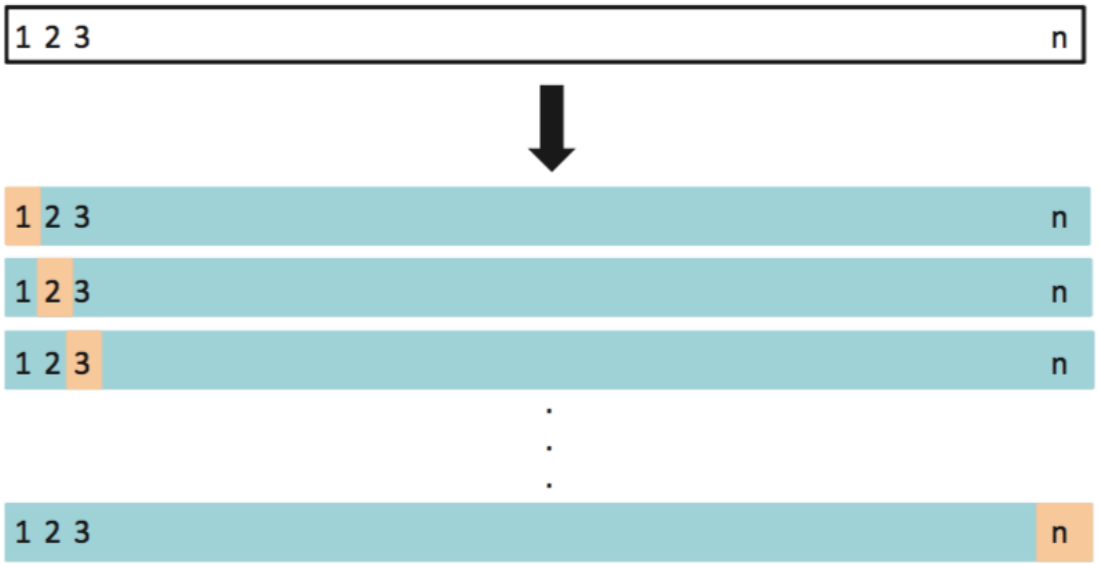
- The validation estimate of the test error can be highly unstable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations – those that are included in the training set rather than in the validation set – are used to fit the model. The estimate or classifier is worse!

Leave-One-Out Cross-Validation

- **Leave-one-out cross-validation** (LOOCV) splits the data into two parts: the validation set with a single observation (x_1, y_1), and training set with the remaining observations $(x_2, y_2), ..., (x_n, y_n)$.
- Based on the training data, we predict y_1 as $\hat{y}_1$ using the value x_1. The test error is approximated by $MSE_1 = (y_1 - \hat{y}_1)^2$. ($MSE_1$ not good enough!)
- We can repeat the procedure by selecting (x_2, y_2) for the validation data, training the statistical learning procedure on the $n - 1$ observations $(x_1, y_1), (x_3, y_3), ..., (x_n, y_n)$, and compute $MSE_2 = (y_2 - \hat{y}_2)^2$. Repeating this approach n times produces n squared errors, $MSE_1, ..., MSE_n$.
- The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i.$$

Leave-One-Out Cross-Validation



Validation data set in beige, and training set in blue.

LOOCV vs Validation Set Approach

LOOCV has the following advantage over the validation set approach.

- The training set of LOOCV is almost the same as the entire data set. The estimate or classifier is almost as good as that based on the entire data set.

- The validation approach yields different results when applied repeatedly, because the training/validation set is randomly divided. LOOCV has no randomness in the splitting.

However, LOOCV can be computationally expensive. (In linear model, the computation can be simplified, the formula is shown in book page 180).

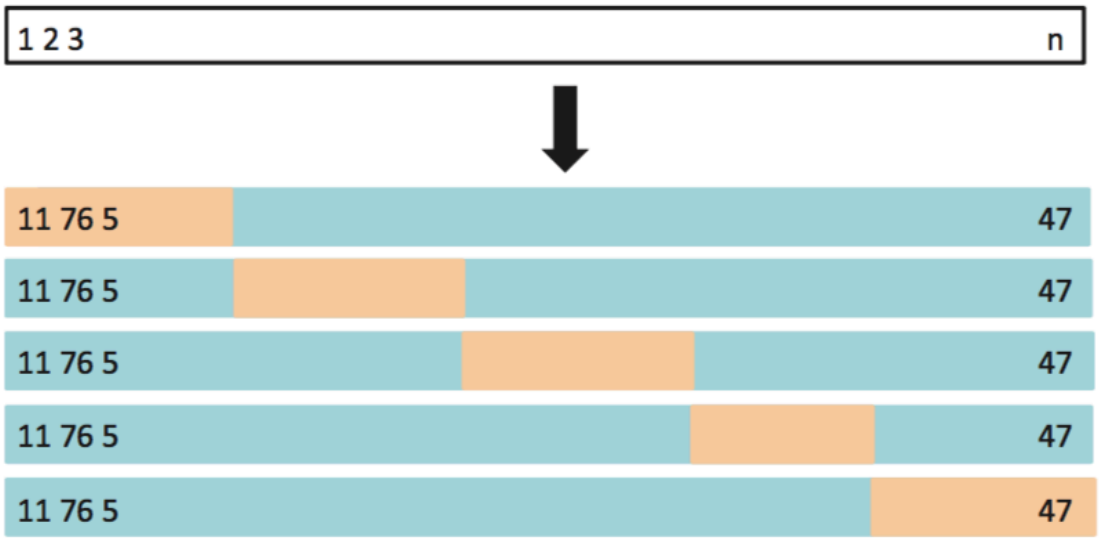
k-Fold Cross-Validation

- **k-fold CV** is to randomly divide the data into k (roughly) equal-sized groups or folds. The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. We compute the mean squared error, MSE_1, for the observations in the first fold.
- Then we repeat the procedure to fold 2, fold 3,..., fold k, and get MSE_2, MSE_3,...,MSE_k.
- The k-fold CV estimate is computed by averaging these values,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i.$$

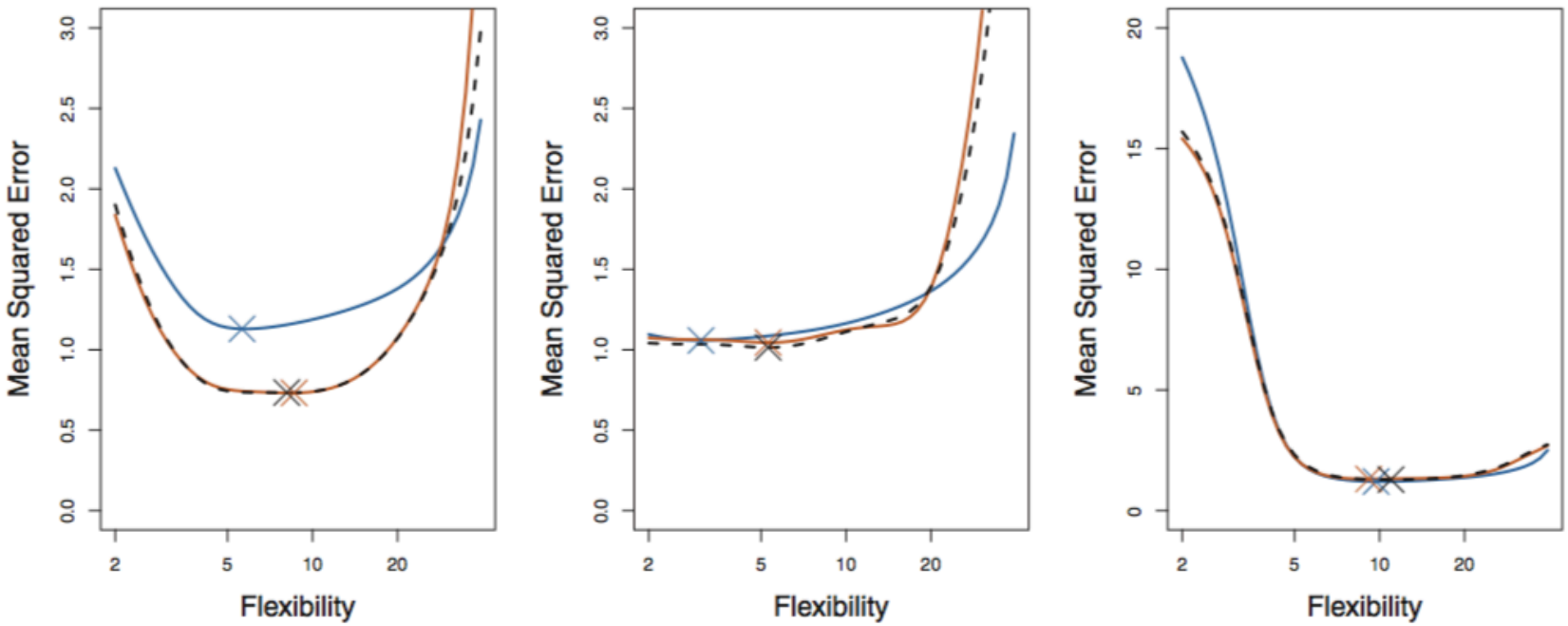
LOOCV is a special case of k-fold CV in which k is set to equal n.

k-Fold Cross-Validation



Validation data set in beige, and training set in blue.

k-Fold Cross-Validation



True test MSE (in blue), the LOOCV estimate (black dashed line), and the 10-fold CV estimate (in orange) for three simulated data sets.

Cross-Validation on Classification Problems

- Cross-validation also works for classification problems.
- For LOOCV, we split the data in the same way as before. We compute the error on the validation set, $Err_1 = I(y_1 \neq \hat{y}_1)$.
- Then we repeat the procedure n times, and get Err_2, Err_3,...,Err_n.
- The LOOCV estimate is computed by averaging these values,

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} Err_n.$$

Cross-Validation is sometimes tricky!

Consider a simple classifier applied to some two-class data:

1. Starting with 5000 predictors and 100 samples, find the 10 predictors having the largest correlation with the class labels.

2. We then apply a classifier such as logistic regression, using only these 10 predictors.

How do we estimate the test set performance of this classifier?