Resampling Methods

Lu Haibo

Mondy, Mar 18, 2019

Introduction

Resampling methods are an indispensable tool in modern statistics

- repeatedly drawing samples from a training set
- refitting a model of interest on each sample in order to obtain additional information about the fitted model

Two most commonly used resampling methods:

- *cross-validation*
  - estimate the test error
- *bootstrap*
  - measure of accuracy of a parameter estimated

Cross-Validation

The Validation Set Approach

randomly dividing the available set of observations into two parts

- *training set*
- *validation set* or *hold-out set*

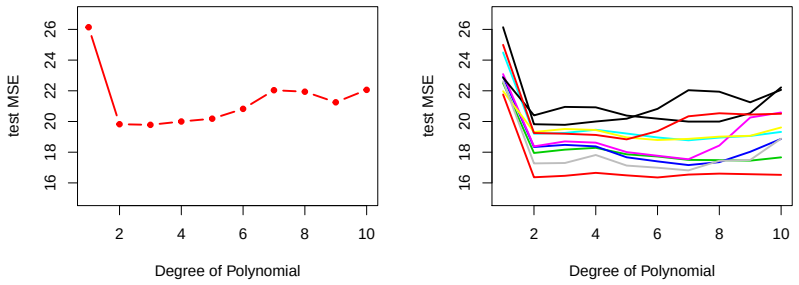

Figure 1: The validation set approach was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: Validation error estimates for a single split into training and validation data sets. Right: The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.

The validation set approach is conceptually simple and is easy to implement. But it has two potential drawbacks:

1. the validation estimate of *the test error rate can be highly variable*, depending on precisely which observations are included in the

training set and which observations are included in the validation set.

2. In the validation approach, only a subset of the observationsare used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to *overestimate the test error* rate for the model fit on the entire data set.

Leave-One-Out Cross-Validation

- A single observation (x_1, y_1) is used for the validation set, and the remaining observations $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ make up the training set. $MSE_1 = (y_1 - \hat{y}_1)^2$
- Repeat the procedure n times. The LOOCV estimate for the test MSE:

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

Major (dis)advantages

1. Far less bias
2. Always yield the same results
3. Could be time costly

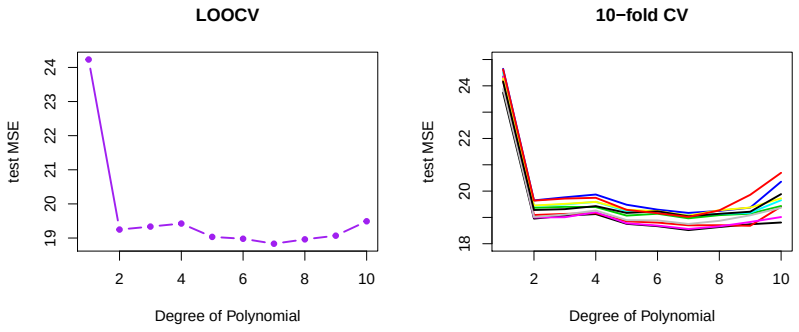

Figure 2: Cross-validation was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: The LOOCV error curve. Right: 10-fold CV was run 10 separate times, each with a different random split of the data into ten parts. The figure shows the 10 slightly different CV error curves.

k-fold Cross-Validation

- randomly dividing the set of observations into k *folds*, of approximately equal size
- pick one fold as a validation set, and fit on the remaining $k-1$ folds

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

- LOOCV is a special case of *k-fold* CV

- k-fold cv can be much more fast than LOOCV, and the variability is much lower than the validation set approach

Cross-Validation on Classification Problems

Rather than using MSE to quantify test error, we instead use the number of misclassified observations,

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

The Bootstrap

The *bootstrap* is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set. (*sampling from the sample*)