Chapter 5: Resampling Methods

Resampling methods are very important tools in building the best predictive models. They involve repeatedly drawing samples from a training set and refitting a model on each sample in order to get the best possible prediction.

For example, if we want to measure the variability of a linear regression fit, we can repeatedly draw different samples from the training data, fit a linear regression to each new sample, and then review and compare the results of fit from these samples.

Resampling approaches can be computationally expensive since it involves fitting the same statistical method multiple times, using different subsets of the training data.

However, due to recent advances in computer power, the computational requirements of resampling methods are not prohibitive.

Two of the most frequently used resampling methods are cross-validation and bootstrap.

Cross-validation can be used to estimate the test error associated with a given statistical learning method or select the appropriate level of flexibility.

The bootstrap is used primarily to assess the accuracy of a parameter estimate or for model selection.

Cross-Validation

Let's revisit the definition of test error rate. The test error is the average error that results from using a prediction model to predict the response of a new observation, which is a measurement that wasn't used in training and building the model.

The test error rate can be easily calculated if a designated test data set is available. The model with the smallest test error rate is the model of choice for a giver research objective.

However, a test dataset is frequently not available. Cross-validation methods are a class of methods that estimate the test error rate 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.

Note that the various models under consideration were built based on the training datasets. The resampling methods remain the same regardless of whether the response Y is qualitative or quantitative.

Cross-Validation: The Validation Set Approach

Suppose we would like to estimate the test error associated with fitting a particular predictive model on a set of observations.

A very simple strategy is the validation set approach. It involves taking a dataset and randomly split the observations into a training set and a validation set. The predictive is fit on the training set and its performance is evaluated on the validation set.

The fitted model is used to predict the responses for the observations in the validation set. The resulting validation set error rate (we can use the mean squared error for quantitative data) provides an estimate of the test error rate.

What are the appropriate sample sizes between the training set and the validation set? The rule of thumb for this simple strategy is to split the data into halves where half of the observations will be the training set and the other half will be the test set.

Statistical models tend to perform worse when trained on fewer observations. The validation estimate of the test error rate can be highly variable, depending on which observations are included in the training set and which observations are included in the validation set.

Next, we present refinements to the validation set method.

Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) involves splitting the set of observations into two parts. But unlike the validation set approach, instead of creating two data subsets of comparable size, a single observation is used for the validation set, and the remaining (n-1) observations make up the training set.

The prediction model is fit on the n-1 training observations, and a prediction is made for the excluded observation.

This approach produces an unbiased estimate for the test error, but it could be a poor prediction estimate since it could be highly variable based upon a single test observation. How can we minimize the variance? By repeating the selection of the test observation.

We can repeat the procedure by selecting another observation to leave out for test purposes, and then the remaining observations will become the training data set.

In summary, the first training set contains all but observation 1, the second training set contains all but observation 2, and so on.

The LOOCV estimate for the test MSE is the average of these n test error estimates. There are a couple of advantages using the LOOCV method over the validation set approach: It has less bias in the prediction estimates and there is no randomness in the training/validation splits.

However, LOOCV has the potential to be expensive to implement, since the prediction model has to be fit n times. If n is large, then the LOOCV process can be time consuming and each individual model is slow to fit!

There is one exception: A least squares linear or polynomial prediction model. An amazing shortcut, utilizing the residuals of high leverage points, makes the time cost of LOOCV the same as that of a single model fit. But the magic formula here is not applicable to any other predictive models.

k-Fold Cross-Validation

An alternative to LOOCV is the k-fold CV. This approach involves randomly dividing the set of observations into k groups or k folds, of approximately equal size. The first fold is treated as a validation set and the model is fit on the remaining k-1 folds.

The MSE for the first fold is then computed on the observations in the held-out fold. This procedure is repeated k times. Each time, a different group of observations is treated as the validation set. This process results in k estimates of test errors.

The k-fold CV estimate is computed by averaging these values.

The k-fold CV approach is a special case of LOOCV in which k=n. In practice, we use k=5 or k=10. It has an obvious computational advantage!

Some statistical learning methods have computationally intensive fitting procedures, and so performing LOOCV has tremendous computational problems if the data size is extremely large. In contrast, a 5-fold or 10-fold CV requires fitting the procedure only 5 or 10 times, respectively.

There is also a bias-variance trade-off associated with the choice of k in k-fold cross-validation. For k=5 or k=10, it has been shown empirically to yield estimates of error that have smaller bias and variance terms than other CV approaches.

Cross-Validation on Classification Problems

We can apply the same cross-validation methods to validate prediction models when Y is a qualitative variable. We use the number of misclassified observations to determine the average error rates across the CV methods discussed earlier in this chapter.

The Bootstrap

The bootstrap is a widely applicable and extremely powerful tool that can be used to quantify the uncertainty associated with a given estimate.

The procedure randomly selects n observations from the data set in order to produce a new bootstrap sample data. The sampling is performed with replacement, which means that the same

observation can occur more than once in the bootstrap data set. This procedure is repeated a large number of times (1,000 is a frequently used number), in order to produce B different bootstrap sample data sets.

Let's look closely at the following example:

We are using the bootstrap approach to assess the variability of the coefficient estimates and predictions from a linear regression model. For each bootstrap sample, we fit a linear regression model and calculate the standard errors for the slope and the intercept. The average of the standard errors across the B bootstrap sample is the estimate of variance for the slope and intercept in the data set.

We can easily fit other statistical learning methods and compare the variability of the respective coefficients by utilizing the bootstrap approach.

Assignments to discuss in class

Review of Lab in Chapter 5

Discussion of the R code: Examples of Cross-Validation Approaches and the Bootstrap