Chapter 5 Notes: Resampling Methods

Statistical Learning with R

Jon Geiger

February 5, 2022

Resampling Methods refer to techniques which sample specific subsets of a training set to reveal more information about the fitted model. The two most common resampling methods are *cross-validation* and *bootstrap*.

Model Assessment: The process of evaluating a model's performance.

Model Selection: The process of selecting the proper level of flexibility for a model.

5.1: Cross-Validation

When evaluating the accuracy of a statistical learning method, it's difficult to calculate the test error when there is no designated test set. One method we can use is to hold out a certain subset of training observations, then applying the method to those held out observations.

This idea of holding out some observations brings upon the idea of the *validation set approach*, in which the available observations are broken into two parts: a *training set* and a *validation set/hold-out set*.

Validating the model on the validation set allows us to get a better estimate of the MSE, thus if we are testing many models, we can get a sense of which model is the best at predicting by choosing the model which minimizes the test MSE.

Two issues with the validation set approach are:

1. The MSE on the validation set is highly variable and can be unreliable with smaller data sets.
2. The validation set error rate may tend to *overestimate* the test error rate for the model fit on the whole data set.

Leave-one-out Cross-Validation (LOOCV) is a type of cross-validation which tries to address the validation set approach's drawbacks.

LOOCV involves splitting the data (size n) into two chunks: one of size $n-1$, and the other of size 1. In other words, a single observation is used for validation. However, this is done with every possible value of (x_i, y_i) in the training set, up to (x_n, y_n).

In LOOCV, the model is fit using the $n-1$ data points, and the remaining data point is used to calculate the test error. This is done for each data point, and the average error is taken, such that:

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{MSE}_i$$

In this case, MSE_i is the individual error calculated from the ith data point, and CV_n simply takes the average of all those errors. This method is very *consistent* because it uses $n-1$ data points to fit the model each time, but in turn it is *computationally expensive* because it needs to fit the model n times.

For linear and polynomial regression, there's a fantastic shortcut which allows you to perform LOOCV with only one model fit:

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2 \quad \text{where} \quad h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \text{ is the leverage of point } x_i$$

While this particular form of the equation cannot be fit to every model, LOOCV is very general and can be fit to any kind of predictive model.

---

k-fold Cross-Validation is an alternative to LOOCV, where the set of observations is randomly split up into k *folds* of equal size. The first fold is used as the validation set, and the remaining $k-1$ folds are used to fit the model. This process is repeated such that every fold is used as the validation set, so it is repeated k times. We then have:

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i$$

By thinking for a second or two, we can notice that k-fold CV is a more general form of LOOCV, or rather, LOOCV is a special case of k-fold CV with $k = n$.

When we are trying to calculate the test MSE as a function of the flexibility, we are trying to find the amount of flexibility such that our test MSE is minimized. In this case, k-fold CV provides good estimates of the test MSE as compared to LOOCV, while being less computationally expensive.

It turns out that k-fold CV also provides better estimates of the test error rate than LOOCV does. Also, k-fold CV both reduces bias *and* variance in the procedure.

Generally, $k = 5$ or $k = 10$ are perfectly fine in estimating LOOCV.

Cross-Validation methods can also be used in classification problems, where the LOOCV error would be given by:

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\text{Err}_i$$

Where $\text{Err}_i = I(y_i \neq \hat{y}_i)$ is the indicator function taking on values of 0 and 1.

In the case where we were to use a polynomial logistic regression model, we could use Cross-validation to estimate the degree of polynomial for which the CV error would be minimized. Adding polynomial terms to a logistic regression model makes a curved decision boundary, and adding too many degrees onto the polynomial will begin to increase the test error rate.

We can also use k-fold Cross Validation for something such as KNN, where we can choose the value of K which minimizes the CV error.

5.2: The Bootstrap

The **Bootstrap** is a statistical tool used to quantify the uncertainty associated with a given estimator or statistical learning method. It is commonly used to assess the variability associated with coefficients in fitting a statistical model.

As an example, consider a problem where we want to invest money in financial assets with returns of X and Y, respectively, where X and Y are random quantities (RVs). We'll invest a fraction α of our money in X, and the remaining $1-\alpha$ in Y. We want to minimize the risk, or variance of the investment, so we minimize $Var(\alpha X + (1-\alpha)Y)$. The value of α which minimizes this variance is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Because we don't know these quantities, we can estimate them using samples, such that:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

Bootstrapping allows us to emulate the process of obtaining new sample sets, so that we can estimate the variability of a parameter without generating additional samples. In this case, we will repeatedly sample *with replacement* from the original data set, such that a single observation can be repeatedly chosen for the same sample.

From a data set Z, we can choose a value of B where we take B different bootstrap data sets $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ and calculate the sample statistics from those bootstrap sets, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$.

We can calculate the mean and standard error of these parameter estimates with:

$$\hat{\alpha} = \frac{1}{B}\sum_{r=1}^{B} \hat{\alpha}^{*r} \quad \text{and} \quad \mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

Essentially, the bootstrap method is a technique which takes a sample of data, creates B sub-samples of that data with replacement, and calculates parameters and their variabilities based on those sub-samples.