Resampling method

Fangfang Song

June 2018

1 k-fold cross validation

k fold cross validation is a widely used approach for estimating test error. The 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, and then obtain predictions for the left-out kth part. This is done for each part $k = 1, 2, ...K$ and then the results are combined.

The cross validation error:

$$CV_{(K)} = \frac{1}{K}\sum_{k=1}^{K} MSE_k \tag{1}$$

1. K-fold cross validation is use the CV(cross validation) error to estimate the test error

2. There is a bias variance trade off associated with the choice of k in k fold 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 $K = n(leaveoneoutcrossvalidation)$, but this estimate has high variance. In LOOCV, the training samples are not shaked up enough and are highly correlated, only differ by one data, thus test error estimate tends to have a high variance and small bias.

3. K=5 or 10 is a good compromise for this bias-variance tradeoff

1.1 The right way to use CV

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

Step 1: Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.

Step 2: We then apply a classifier such as logistic regression, using only these 100 predictors.

This is the **wrong** way to use CV. Since this 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. The right way is as follows: we first define our folds, before we do any fitting, we remove one of the folds, all the data for that fold, the predictors and the response variable. and now we can do whatever we want on the other remaining parts. we can filter and fit.

2 Bootstrap

Bootstrap is a powerful tool for assessing uncertainty in estimates, particularly good for getting standard error of an estimate and getting confidence limits. Bootstrap convey the idea of tying to pull yourself up from what you've got. We are going to use the data itself to try to get more information about our estimator.

Bootstrap, Rather than repeatedly obtaining independent data sets from the population, which we cannot do without access to the population(all we have is a sample from the population), we're going to sample from the data itself with replacement.And then use those samples to get an idea of the variability, in the same way that we use the samples from the population to produce the histogram. Each of these "Bootstrap data sets" is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.