Resampling Methods

Earl Patrick B. Macalam

December 19, 2020

Resampling methods are an indispensable tool in modern statistics. They involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. Goal here is to estimate the test error which would allow us to know if the model is usable (only if it has low error rate) or not. Ofcourse, test data are not available so we need to find ways on generating them reasonably.

1 Cross - Validation

Test Error - is the average error that results from using a statistical learning method to predict the response on a new observation — that is, a measurement that was not used in training the method.

Training Error - is the average error that results from using a statistical learning method to predict the response on the training data.

2 The Validation Set Approach

- Used to estimate the test error associated with fitting a particular statistical learning method on a set of observations.
- Involves randomly dividing the available set of observations into two parts namely: *training set* and *validation set*.
- The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.
- The resulting validation set error rate—typically assessed using MSE in the case of a quantitative response—provides an estimate of the test error rate.

Drawbacks:

1. The validation esti- mate of the test error rate can be highly variable.

2. 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. 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.

3 Leave-One-Out Cross-Validation

- Attempts to address that method's out drawbacks.
- 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. We repeat this until we utilized all single observations in the data as test set. We then compute the MSE on each process producing n MSE in all. Then compute the mean MSE.
- Again, the method addresses the two drawbacks of validation method.
- Disadvantage of this is it's computationally heavy.

4 k-Fold Cross-Validation

- Divide the available observations to k parts with equal sizes. For every process of division we compute MSE, thus resulting to n MSE. Then average it.
- Usuall values of k are 5 and 10 as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

5 Bias-Variance Trade-Off for k-Fold Cross-Validation

Each succeeding single observation we picked as validation set in LOOCV is highly correlated and thus would produce to a positively correlated observation(high variance). In terms of bias reduction, LOOCV is favorable since it consider all single elements in our availabe observations and thus poducing low bias. As we increase k variance decreases but bias increases.

6 Cross-Validation on Classification Problems

Analogous with the two preceding methods but here we use the number of misclassified observations instead of MSE as an estimate to test error.

7 The Bootstrap

This is use in checking the accuracy of our parameter estimate where we generate n data sets and compute the parameter estimate for each data set resulting us to n parameter estimates. Then compute the mean of these n estimates and compare to the true value. We can also, compute for the standard deviation to know the accuracy of the estimates produced in each dataset.

Two Cases:

1. The case where the true population is available
   - Just generate n datasets from that population.
2. True Population is not available
   - Resample till you obtain the desired n datasets (with replacement).