Boxi Huang's Reading Note bh563@cornell.edu

Page Covered: PP 175-190

The Validation Set Approach

It involves randomly dividing the available set of observations into two parts, a training set, and a validation set or hold-out set.

The model is fit on the **training set**.
The fitted model predict the response with the **validation set**.
The **validation set error rate** -- typically assessed using MSE in the case of a quantitative response -- provides an estimate of the test error rate.

Drawbacks of validation set:

1. The validation estimate of the test error rate can be highly variable due to the observations in training and validation.
2. Only a subset of the observations (training set) is used to train the model and with less model, we might be having a problem with overestimating result from the validation set about the test MSE.

Leaving-one-out Cross-Validation (LOOCV)

A single observation is used for the validation set and the remaining observations make up the training set. The statistical learning method is fit on the n-1 training observations and a prediction y_i is made for the excluded observation. We repeat the procedure by selecting one observation out to compute MSE. By repeating this approach n times produces n squared errors, MSE1, ...MSE_n. The LOOCV estimate for the test MSE is the average of these n test errors.

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

LOOCV has a couple of major advantages:

1. Far less bias
2. No randomness in the training/validation set splits. (always yield the same results)

K-fold Cross-Validation

Randomly dividing the set of observations into k groups, or folds, of approximately equal size. This procedure is repeated k times and produces k estimates of the test error. The k-fold CV estimate is computed by averaging these values:

Feb. 3rd, 2021
bh563@cornell.edu

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

Bias and Variance Trade-off for k-fold Cross-Validation

Typically, k is 5 or 10 and this is better due to a relatively inexpensive computation requirement.

LOOCV will always produce an **unbiased** estimate of the test errors since every training set contains n-1 observations. But LOOCV has a higher variance than K-fold CV since we might have an almost identical set of observations every time. So the output might be highly positively correlated.

Having K-fold cross-validation will lead to an intermediate level of bias since each training set contains (k-1)n/k observations, a lot less than LOOCV. When we perform k-fold CV with k<n, we are averaging the outputs of k fitted models that are somewhat less correlated with each other, prompting a test MSE with lower variance.

Bootstrap

The bootstrap can be used to quantify the uncertainty associated with a given estimator or statistical learning method. The sampling is performed with replacement and obtains distinct data sets by repeatedly sampling observations from the original data set, which means that the same observation can occur more than once in the bootstrap data set.