Resampling Methods

Sravani Vadlamani

Agenda

- Resampling methods – Definition and Why they are used
- Different Techniques
  - Cross Validation
  - Bootstrap

Definitions

- Resampling methods
  - Involve repeatedly drawing samples from a data set and refitting a model of interest on each sample with the goal of obtaining additional information about the fitted model.
  - Computationally expensive since the same model is fitted multiple times on different sets of data
  - Cross-validation and Bootstrap are the commonly used resampling techniques

Definitions

- Cross-validation is used to estimate the test error of a statistical learning method to evaluate its performance or to select the appropriate level of flexibility
- Model Assessment – the process of evaluating the performance of a model
- Model Selection – the process of selecting the appropriate flexibility level of a model
- Bootstrap is commonly used to provide a measure of accuracy of a parameter estimate or a given statistical learning method

Training and Test Data

- Model estimation is done using training data and the model performance is evaluated using test data
- The test error rate is the average error that results from using a statistical method to predict response on a new observation (i.e., data not used to train the model)
- Given a data set, the use of a particular statistical method is warranted if it results in low test error.
- The availability of a designated test set is an issue
- Techniques to use the available training data set and hold out a portion of the observations for testing will be discussed.

Validation Set Approach

- Randomly divide the available observations into a training set and a validation or hold out set.
- The model is fit on the training set which is then used to predict responses for the observations in the validation set.
- The resulting validation set error rate offers an estimate of the test error rate
- Although this approach is conceptually simple and easy to implement it has potential drawbacks:
  - The estimated test error is highly variable depending on which observations fall into the training and validation sets
  - The statistical method is trained with a fewer observations and hence the test error rate tends to be overestimated.

Leave-On-Out Cross-Validation (LOOCV)

- This approach attempts to address the drawbacks of the validation set approach
- Does not split the data evenly but withholds a single observation for the validation set and uses the remaining n-1 observations for fitting the model. This process is repeated n times with each observation held out once. The n mean squared errors are averaged to yield the LOOCV estimate of the test error

$$CV(n) = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

- Advantages
  - Less bias than the validation set approach
  - LOOCV does not overestimate the test error rate as much as the validation set approach
  - LOOCV is less variable – it always yields the same results since there is no randomness in the training/validation splits

K-Fold Cross-Validation

- Randomly divides the observations into k groups or folds of roughly equal size. Each of the k folds are treated as the validation set and the remaining k-1 folds are used to estimate the models. This results in k estimates of the test error which is computed as

$$CV(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

- LOOCV is a special case of k-fold cross validation where k=n
- Typical values of K are 5 or 10

K-Fold Cross-Validation

- Goal of Cross validation
  - To determine how well a given statistical learning method performs on independent data using the test MSE metric
  - To estimate the minimum point in the test MSE curve that is used to compare different statistical learning methods or compare different levels of flexibility for a single method

Bias Variance Trade-off for K-Fold Cross Validation

- K-fold has computational advantage to LOOCV
- K-fold gives more accurate estimates of the test error rate than LOOCV due to bias-variance trade off
- In terms of bias, LOOCV is preferable to k-fold and k-fold is preferable to validation set approach
- In terms of variance k-fold is preferable to LOOCV and LOOCV is preferable to validation set

Bootstrap

- A powerful and widely applicable tool to quantify the uncertainty associated with a given estimator or statistical learning method including those for which a variability estimate is difficult to obtain.
- Use a computer to emulate the process of obtaining new sample sets
- Bootstrap involves obtaining distinct data sets by repeatedly sampling from the original dataset
- Sampling is done with replacement which means that same observation can occur multiple times or some observations may not be included at all
- The process is repeated B times to yield B bootstrapped data sets which can be used to estimate quantities like standard error