WEEK-6
MODEL EVALUATION AND TESTING
Session-5

CROSS VALIDATION

Cross-Validation is an essential tool in the Data Scientist toolbox. It allows us to utilize our data better.
An alternative is to partition the sample data into a training (or model-building) set, which we can use to develop the model, and a validation (or prediction) set, which is used to evaluate the predictive ability of the model. This is called cross-validation.

Why do we need Cross-Validation?

- To verify regression model's accuracy on multiple and different subsets of data. Therefore, ensure that it generalizes well to future data inputs.
- It is essential when we have a small dataset.
- To detect overfitting.
- To Optimize the model by tuning its parameters.

Cross validation Techniques

Some of them are commonly used Cross Validation Techniques are

1. Hold-out method
2. K-folds Cross validation
3. Leave-one-out Cross validation

1. Hold – out method

This is the simplest evaluation method and is widely used in Machine Learning projects. Here the entire dataset is divided into 2 sets such as train set and test set. The data can be divided into 70-30 or 60-40, 75-25 or 80-20, or even 50-50 depending on the use case. As a rule, the proportion of training data has to be larger than the test data.

The data split happens randomly and we cannot be sure which data ends up in the train and test bucket during the split unless we specify random_state. This can lead to extremely high variance and every time, the split changes, the accuracy will also change.
One of the major advantages of this method is that it is computationally inexpensive compared to other cross-validation techniques.

There are some drawbacks to this method:

The test error rates are highly variable (variance) and it totally depends on which observations end up in the training set and test set.

Only a part of the data is used to train the model (high bias) which is not a very good idea when data is not huge and this will lead to overestimation of test error.

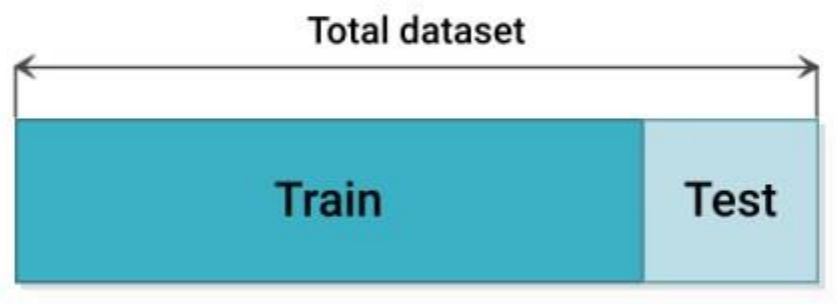

Fig: Hold – out method

2. Leave One Out Cross-Validation (LOOCV)

In this method, instead of dividing the data into 2 subsets, we select a single observation as test data, and everything else is labeled as training data and the model is trained. And then 2nd observation is selected as test data and the model is trained on the remaining data.

This process continues 'n' times and the average of all these iterations is calculated and estimated as the test set error.

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

When it comes to test-error estimates, LOOCV gives unbiased estimates (low bias). But bias is not the only matter of concern in estimation problems. We should also consider variance.

LOOCV has an extremely high variance because we are averaging the output of n-models which are fitted on an almost identical set of observations, and their outputs are highly positively correlated with each other. Here we can clearly see this is computationally expensive as the model is run 'n' times to test every observation in the data.

3. K-Fold Cross-Validation

In this technique, the whole data is divided into k sets of almost equal sizes. The first set is selected as the test set and the model is trained on the remaining k-1 sets. The test error rate is then calculated after fitting the model to the test data.

In the second iteration, the 2nd set is selected as a test set and the remaining k-1 sets are used to train the data and the error is calculated. This process continues for all the k sets. The mean of errors from all the iterations is calculated as the CV test error estimate.

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

In K-Fold CV, the no of folds’ k is less than the number of observations in the data (k<n) and we are averaging the outputs of k fitted models that are somewhat less correlated with each other since the overlap between the training sets in each model is smaller. This leads to low variance than LOOCV.

As the number of folds’ k increases, the variance also decreases (low variance). This method leads to intermediate bias because each training set contains fewer observations (k-1) n/k than the Leave One Out method but more than the Hold Out method.
Typically, K-fold Cross Validation is performed using k=5 or k=10 as these values have been empirically shown to yield test error estimates that neither have high bias nor high variance.

The major disadvantage of this method is that the model has to be run from scratch k-times and is computationally expensive than the Hold Out method but better than the Leave One Out method.

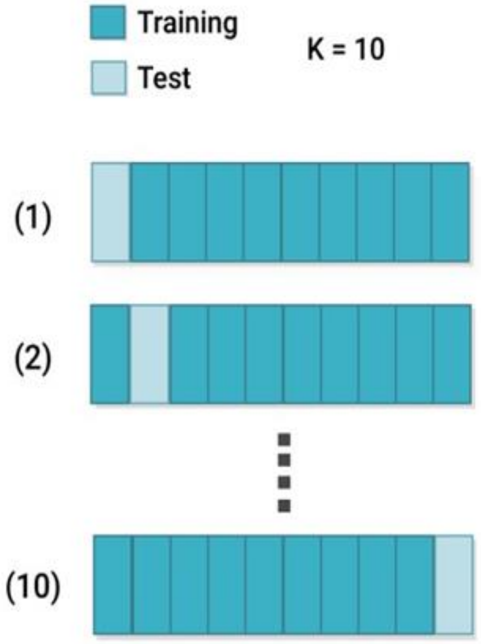

Fig: K-Fold Cross-Validation