# Cross-Validation

## Overview

Usually in Predicting, our main aim is to reduce the error , we divide our dataset in two parts train set and test set (sometimes we call test set as validation set , however they are different, but here we have used it interchangeably) so after dividing dataset, we estimate the parameters of the model on the train set. Then after that we calculate some type of performance measures, and asses the loss on test sets for comparisons of different models.

## Problem with train-test split Procedure

The Train Test Split works good for most of the data but sometimes its fails to assess the real loss by predicted model , specifically when dataset is small, actually for small datasets the performance measures for the test set may be too much optimized for the specific test set, This may be due to the fact that

$$E(MSE\ of\ Train\ Set) = \frac{n-p-1}{n+p+1}E(MSE\ of\ Test\ Set)$$

The second problem with train-test split is, the train set we used to optimize the model is not using the full potential of the information we have suppose we take 80:20 ratio for train test split , then we are using only 80% of the data rather than using 100% data. There is a chance that remaining 20% of the may play significant role in determining optimized parameters

## Cross-Validation

Cross-Validation is a method that reduces our dependency on how the data splits in train and test, there may be, only by chance that our performance metrics are representing our model as good, this is due to the fact we do not use all the data to calculate performance metrics. To eradicate the dependency on only one train test splits data we use Cross-Validation.

In CV , we do not split data only once , in this method we split our data a lot of time then after that we compute performance measures, as for example accuracy. Then after calculating we performance measures for all the splits we summarize the performance measure to our need some people take mean of the performance measures of all split some take mode. Its totally dependent on interest of statistician and type of data

It can be classified into following

1. **Exhaustive Cross-Validation :** In this type of Cross validation, We focus on Splitting our data in train test set in every possible way.

Following are the methods for Exhaustive CV

(a) *Leave p Out CV (LpOCV)*

In this type of Exhaustive CV , we choose "p" a number , then we take p observation from data as validation set and "n-p" as train set and this is repeated in all the possible combination in $\binom{n}{p}$ ways.

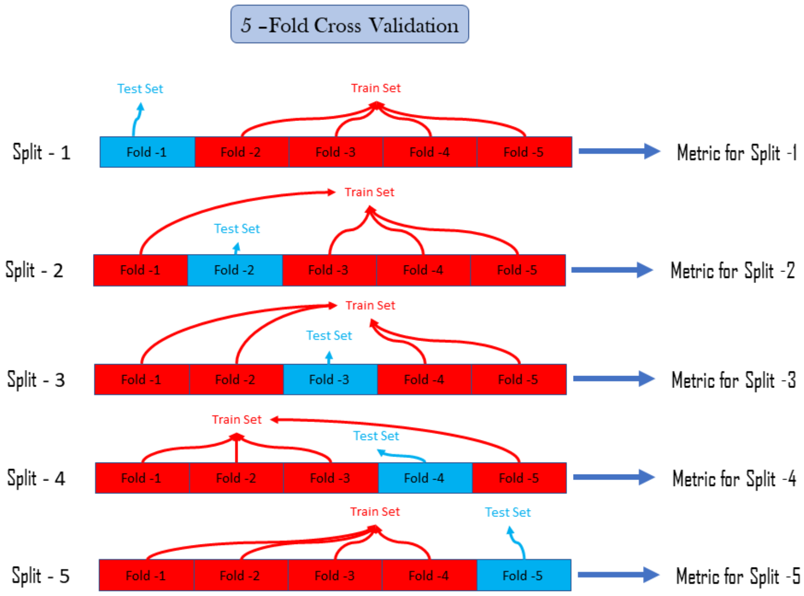
(a) *Leave one Out CV (LOOCV)*

It is a particular case of *LpOCV*, where  p =1 , in this method we take only one observation as a validation set and all others as a train set

2. **Non Exhaustive Cross-Validation :** In these type of cross validation we do not compute the all possible ways of splitting validation test and train set

(a) *k-Fold Cross Validation*:  k-fold Cross-validation where is k is a parameter and a positive integer suppose k=5 means there are 5-fold cross-validation, it simply divide our observations in our dataset into 5 groups commonly known as a fold, then we hold the first fold as a test set and all other folds are merged to create train set and then we calculate the performance metric we are interested in, and then do the same again by holding second fold as a test set and remaining as train set and after that calculate performance metric, this is known as a performance metric for the second split, similarly in k fold cross validation we calculate performance metric k times for k splits and further after calculating metric for every split we can calculate statistics of our interest such as the mean of these k performance metrics or mode, median or whatever statistic we want

5 -Fold Cross Validation Can be represented as follows


5 –Fold Cross Validation
Test Set
Train Set
Split - 1
Fold -1
Fold -2
Fold -3
Fold -4
Fold -5
Metric for Split -I
Split - 2
Metric for Split -2
Split - 3
Metric for Split -3
Split - 4
Metric for Split -4
Split - 5
Metric for Split -5