Cross Validation

Involves

1. Estimate the parameters for the machine learning methods (training)
2. Evaluate how well the machine learning methods work (testing)

The terrible approach is to take all the data for the training. So, we won’t be left with any data to test. Reusing the same data for training and testing is a bad idea.

The better idea will be to use 75% of data for training and 20% of the data for testing.

What cross validation does is to split the data into chunks. For example, we have 100 data points, the first 25% i.e., 1-25 goes to test and the rest 75% (26-100) goes for training. In second iteration the next 25% 26-50 goes for testing and rest 75% (1-25,50-100) will be for training.

We have different cross validation techniques

1. K fold cross validation
2. Leave one out cross validation (Here each sample tested individually)

What we can find with cross validation?

1. Best model (Logistic, SVM, Decision Tree, etc.…) for our data.
2. Find best parameters using hyperparameter tuning.