Cross Validation

It’s a technique used to estimate or evaluate the performance or accuracy of Machine Learning Model.It used to protect the model from overfitting. Overfitting means you train the data very well but when it test in real life scenario it give poor result. Because the training data-set is not enough to cover all the possible scenario of model.

So, the main objective of Cross-Validation is to make model working well for the real-life data.

In this process, the original data is divided into several subsets. We fixed one subset as testing data and all other as training data. After training, the model do testing o testing data. Repeat this process by making testing and training subsets differently of overall data-set.

Types of Cross Validation:

1. Holdout method

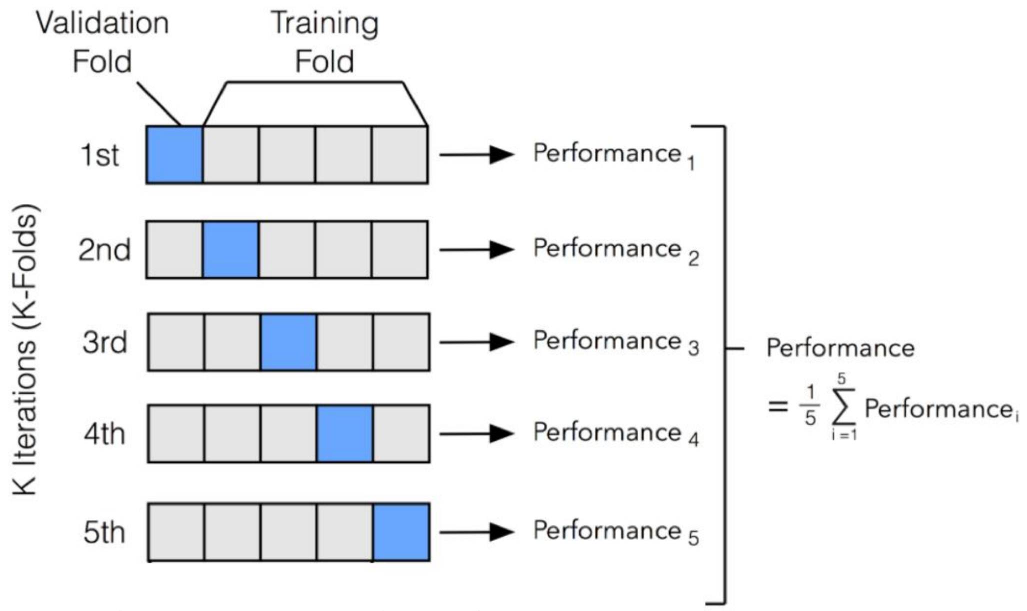
In this approach the data-set is divided into two subsets, Training data and Testing data. The model is trained on training data-set and assess on testing data-set. Mostly the size of training data-set is 80% or 70% of overall data-set. While the testing data-set size is about 20% or 30 % respectively. Before training and testing the whole data is randomly shuffled.

2. K-fold cross-validation

It’s a improved version of holdout method. We divide the data into k number of splits, and then perform Holdout methods k times.

For example:

k = 5 , means there are 5 subsets A, B, C, D and E. So for first iteration model is test on A subset and Train on B, C, D and E subset. For 2nd Iteration, Model is training on A, C, D and E subset and test on B subset. And So on.



The k-fold Cross Validation Technique produce a less biased models. Because every data point will appear in both the training and testing set. If you have limited data points then this method gives Optimal result.

3. Stratified k-fold cross-validation

Since we do randomly Shuffling data and Split them into k-folds cross validation, There is a possibility that we should have imbalanced subsets which cause inaccurate results.

While shuffling we have a training set which are mostly from same class. So, while training it gives good result but when test it, it gave bad result due to imbalance of data set.

For avoiding this Imbalance , ***Stratification*** is used. In this process the data is rearranged and make sure that each subset will give a good representation of whole data-set.

4. Leave-p-out cross-validation

In Leave-p-out cross-validation “p” number of data points are taken out from whole data and which can represented by n.

The model is tested on p data points trained on remaining data points. The same process is repeated for all possible combinations of p from the original data sample. Final result is averaged of all iterations result.

References:

1. https://machinelearningmastery.com/k-fold-cross-validation/
2. https://www.mygreatlearning.com/blog/cross-validation/
3. https://learn.g2.com/cross-validation
4. https://www.kaggle.com/general/204878