The concept of cross-validation in the context of machine learning?

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Types of Cross Validation?

Non-exhaustive Methods:Non-exhaustive cross validation methods, as the name suggests do not compute all ways of splitting the original data. Let us go through the methods to get a clearer understanding.

- Holdout method - This is a quite basic and simple approach in which we divide our entire dataset into two parts viz- training data and testing data. As the name, we train the model on training data and then evaluate on the testing set.
- K fold cross validation - K-fold cross validation is one way to improve the holdout method. This method guarantees that the score of our model does not depend on the way we picked the train and test set.
- Stratified K Fold Cross Validation - Using K Fold on a classification problem can be tricky. Since we are randomly shuffling the data and then dividing it into folds, chances are we may get highly imbalanced folds which may cause our training to be biased.

Exhaustive Methods - Exhaustive cross validation methods and test on all possible ways to divide the original sample into a training and a validation set.

- Leave-P-Out cross validation - When using this exhaustive method, we take p number of points out from the total number of data points in the dataset (say n). While training the model we train it on these (n – p) data points and test the model on p data points. We repeat this process for all the possible combinations of p from the original dataset. Then to get the final accuracy, we average the accuracies from all these iterations.
- Leave-one-out cross validation - This is a simple variation of Leave-P-Out cross validation and the value of p is set as one. This makes the method much less exhaustive as now for n data points and p = 1, we have n number of combinations.