1 Model Validation

1.1 Introduction

Model validation is an essential and perhaps the most important of machine learning process. This is because it helps in checking the stability of the model and stops it from $overfitting$ or $underfitting$ on training datasets. *See Evaluation metrics to understand overfitting and underfitting.* Machine learning algorithms are prone to underfitting or overfitting depending on the structure of data available, hence the need to $cross-validate$ model using special techniques.

1.2 Cross-Validation

Cross-validation is a machine learning technique used to check the performance of a machine learning algorithm, especially with regards to unseen data. It is based on re-sampling or shuffling of training data during learning, to ensure the algorithm captures the underlying structure of the data and does not overfit. This process leads to a better hypothesis capable predicting unseen data with some level of accuracy.

1.2.1 Cross-validation techniques

1. Train-Test split
   This technique essentially splits the whole dataset into random subsets of some particular ratio. In most cases the train-test random split is usually 70:30 (70% for training model and generalization & 30% for testing the resultant model.) This technique is perhaps the simplest case for cross-validating our model.

2. K-Folds Cross-validation
   k-Folds Cross-validation is by far the most used cross-validation technique in machine learning. This is because it is both computationally inexpensive and efficient. Unlike train-test split, k-fold splits the dataset into k equal sized subsets (or folds).

   k is chosen with respect to the size of the dataset (5-10 fold is ideally a good start for large dataset.) k samples is retained for training while k -1 used for testing the model. This is repeated k-times with each of the k samples used exactly once.

- Partition data into k-equal sized subsets.
- fit data using k-1 subsets and validate using kth fold.
- Repeat process k-times and average prediction scores.

3. Stratified K-Folds Cross-validation
   Very similar to the k-fold cross-validation except that it best works for binary/multiclass classification. the folds are selected so that the mean response value $\{0, 1\}$ is approximately equal in all the folds. In essence, each subset of the partition contains equal proportion of the classes.

4. Leave-one-out cross-validation
   Leave-one-out cross-validation (LOOCV) is a particular case of leave-p-out cross-validation with p = 1. This cross-validation technique is computational expensive than k-fold. A reason attributed to the iterative selection of one fold each of the entire dataset until k is reached.