Assignment 11

1. What are the different validation techniques in Machine Learning?

- **Validation Set Approach:** The input dataset is divided into a training set and test or validation set in the validation set approach. Both the subsets are given 50% of the dataset.
- **Leave-P-out cross-validation:** In this approach, the p number of datasets are left out of the training data. It means, if there are total n data points in the original input dataset, then **n-p data points** will be used as the **training dataset** and the **p data points as the validation set.** This complete process is repeated for all the samples, and **the average error is calculated** to know the effectiveness of the model
- **Leave one out cross-validation:** Here instead of p, we need to take 1 dataset out of training. Only one data point is reserved, and the remaining dataset is used to train the model. This process repeats for each data point. Hence for n samples, we get n different training set and n test set.
- **K-fold cross-validation**: K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called folds. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set.
- **Stratified k-fold cross-validation:** This technique is similar to k-fold cross-validation with some little changes. This approach works on stratification concept, it is a process of rearranging the data to ensure that each fold or group is a good representative of the complete dataset. To deal with the bias and variance, it is one of the best approaches.

2. What are the different ways of improving the accuracy of a Machine Learning Model?

- Collect data: Increase the number of training examples.
- Feature processing: Add more variables, or select relevant variables for better feature processing.
- Model parameter tuning: Consider alternate values for the training parameters used for model training.
- Train your model using cross-validation.

3. What is stratified cross-validation and when should we use it?

Stratified k-fold cross-validation: This technique is similar to k-fold cross-validation with some little changes. This approach works on stratification concept, it is a process of rearranging the data to ensure that each fold or group is a good representative of the complete dataset.

It should be used to deal with the bias and variance. For some dataset variance in price can be high, to tackle such situation, a stratified k-fold cross-validation technique is useful.

4. What is the difference between the validation set and the holdout test set?

The validation set is a subset of input data set used to optimize the model parameters.

The validation dataset can be run number of times until the optimum parameters are achieved.

The test set/ holdout test set is used to provide an unbiased estimate of the final model. It is run only once to evaluate the model performance.

5. How to solve a multi-class classification problem vs multi-label classification available?

Multi-class classification & **multi-label classification** is that in multi-class problems the classes are mutually exclusive, whereas for multi-label problems each label represents a different classification task, but the tasks are somehow related.

Multi-class classification problem.

- Load dataset from the source.
- Split the dataset into “training” and “test” data.
- Train Decision tree, SVM, and KNN classifiers on the training data.
- Use the above classifiers to predict labels for the test data.
- Measure accuracy and visualize classification.

Multi-Label classification problem.

There are two main methods for tackling a multi-label classification problem: **problem transformation methods and algorithm adaptation methods(KNN, RF), Ensemble methods**. Problem transformation methods transform the multi-label problem into a set of binary classification problems, which can then be handled using single-class classifiers.

- Treat each label as a separate single classification problem.

- Each label will serve as an input feature for next label of classification. In this way is will for a chain of classifiers to preserve the label correlation.
- Then transform the problem into a multi-class classification problem by providing them with unique class for every label combination in our dataset.