**Data Science Session 19 Assignment 1**

**1. What are the three stages to build the hypotheses or model in machine learning?**

**Answer:**

There are 3 stages to build hypotheses for model for machine learning.

a) Model Building:

b) Model Testing

c) Applying the model

**2. What is the standard approach to supervised learning?**

**Answer:**

Supervised learning is a type of machine learning that enables the model to predict future outcomes after they are trained based on past data. In order to train the model, set of input and out is required based upon which models get trained about data to predict the output. Every output that the model provides, along with the new data that facilitated the output, becomes the new input-output combination that is fed as training data into the model for learning.

For Supervised learning, standard approach is to split the data into **Training** and **Test** data. Then develop the model based upon Training data and Once model has been obtained then use Test data to predict the outcome using data on model.

**3. What is Training set and Test set?**

**Answer:**

**Training Set:** The training data set in Machine Learning is the actual dataset used to train the model for performing various actions. This is the actual data the ongoing development process models learn with various API and algorithm to train the machine to work automatically.

**Test Set:** The Testing data set in machine learning is used to provide and unbiased evaluation of model that are completed and fit upon training data set. Such test data set is used for testing the model that it is working appropriately or not. Hence if changes are required that can be done after evaluation of outcome predict by model

**4. What is the general principle of an ensemble method and what is bagging and**

**Boosting in ensemble method?**

**Answer:**

Ensemble model combines multiple ‘individual’ (diverse) models together and delivers superior prediction power. The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model.

**Bagging** (Bootstrap Aggregating) is an ensemble method. First, we create random samples of the training data set (subsets of training data set). Then, we build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting. Bagging helps to reduce the variance error.

**Boosting** provides sequential learning of the predictors. The first predictor is learned on the whole data set, while the following are learnt on the training set based on the performance of the previous one**.** Itstarts by classifying original data set and giving equal weights to each observation. If classes are predicted incorrectly using the first learner, then it gives higher weight to the missed classified observation. Being an iterative process, it continues to add classifier learner until a limit is reached in the number of models or accuracy. Boosting has shown better predictive accuracy than bagging, but it also tends to over-fit the training data as well.

**5. How can you avoid overfitting?**

**Answer:**

A statistical model is said to be over fitted, when we train it with a lot of data. When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too much of details and noise.

There are several ways to avoid overfitting, which are

1.Cross validation

2.Pre-pruning

3.Post-pruning

4. Regularization

5.Feature Selection

6.Take more data