ASSIGNMENT 19 – MACHINE LEARNING

1. **What are the three stages to build the hypotheses or model in machine learning?**
    - Model building
    - Model testing
    - Applying the model

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

    The standard approach to supervised learning is to split the set of examples into the training set and the test.

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

    In machine learning, a set of data is used to discover the potentially predictive relationship known as 'Training Set'. Training set is an example given to the learner. While Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of examples held back from the learner. Training set are distinct from Test set.

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

    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 is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

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

    Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

    There are two important techniques that we can use when evaluating machine learning algorithms to limit overfitting:

    - Use a resampling technique to estimate model accuracy.
    - Hold back a validation dataset.

ASSIGNMENT 19 – MACHINE LEARNING

The most popular resampling technique is k-fold cross validation. It allows you to train and test your model k-times on different subsets of training data and build up an estimate of the performance of a machine learning model on unseen data.

A validation dataset is simply a subset of your training data that you hold back from your machine learning algorithms until the very end of your project. After you have selected and tuned your machine learning algorithms on your training dataset you can evaluate the learned models on the validation dataset to get a final objective idea of how the models might perform on unseen data.

Using cross validation is a gold standard in applied machine learning for estimating model accuracy on unseen data. If you have the data, using a validation dataset is also an excellent practice.