**Assignment 11.1**

**Machine Learning**

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

**Answer**: Below are the 3 stages to build the hypotheses in Machine Learning

a)      Model building  
b)      Model testing  
c)      Applying the model

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

**Answer**:**Machine learning technique** for building predictive models from known input and response data. **Supervised learning** is a type of **machine learning** algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values.

The standard approach to supervised learning is to split the set of example into the training set and the test. Model once trained on trained data can be used to verify/predict the results on test data.

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

**Answer:   
Training Set:**In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. These are the examples given to a learner to get itself trained on the same.

**Test Set**  
Once the Model has been trained, it is tested against the Test Set to test the accuracy or the prediction capability of the model. Test Data is held back from the learner.

Training set are distinct from Test set.

If the prediction scores for the test set are unreasonable, we’ll have to make some adjustments to our modeland try again.

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

**Answer:**

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.  **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.

**Bagging**is used typically when you want to reduce the variance while retaining the bias. This happens when you average the predictions in different spaces of the input feature space. In bagging, first you will have to sample the input data (with replacement) to generate multiple sets of input data. For each of those sets, the same baseline predictor is run to get a trained model for each of the training set.

**Boosting**Boosting is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.

**Question 4.How can you avoid overfitting ?**

**Answer:**By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as **cross validation**. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.

In this technique,  a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to “test” the model in the training phase.