# Session 20 – Machine Learning 1

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

The three stages to build the hypotheses or model in machine learning are:

1. Model building
2. Model testing
3. Applying the model

2. What is the standard approach to supervised learning?

The standard approach to supervised learning is to split the set of example into the training set and the test set. Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly.

3. What is Training set and Test set?

**Training Set** – A training set is a data set used to train a model in order to prepare it for future simulations and outputs after feeding it with real data. In training the model, specific features are picked out from the training set. These features are then incorporated into the model.

**Test Set -** A test set is a dataset used to measure how well a model performs at making predictions on that test set, after it is prepared via training sets. If the prediction scores for the test set are unreasonable, we’ll have to make some adjustments to our model and try again. Test set should be different from the training set as it won’t help in assessing the model correctly, since the model itself was built using the training set. Hence, we should test the model with the data, it hasn’t seen before.

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?

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; especially if you have a small data set and you are forced to come with a model based on the limited data. 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 data points 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.