**ASSIGNMENT 7.2**

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

a) Model building

b) Model testing

c) 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.

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

#### **Training Set**

In Machine Learning, a training set is a dataset used to train a model. In training the model, specific features are picked out from the training set. These features are then incorporated into the model. Thereby, if the training set is labeled correctly, the model should be able to learn something from these features.

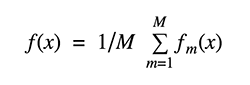
**Test Set**

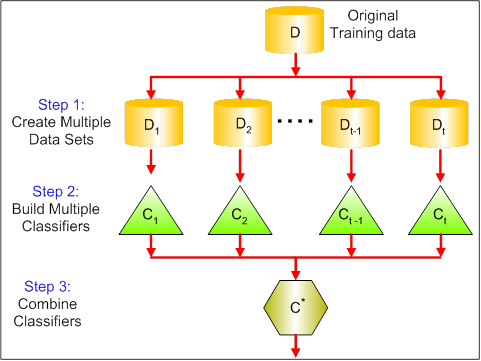
The test set is a dataset used to measure how well the model performs at making predictions on that test set.  In the case of sentiment analysis, a test set is a dataset of tweets that are distinct from the tweets in the training set. If the prediction scores (sentiment scores) for the test set is unreasonable, we’ll need to make some adjustments to our model and try again.

**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. The two paradigms of ensemble methods are:-Sequential ensemble methods and Parallel ensemble methods. Bagging is a method in ensemble for improving unstable estimation or classification schemes. Bagging both can reduce errors by reducing the variance term. Boosting method is used sequentially to reduce the bias of the combined model. Boosting can reduce errors by reducing the variance term.

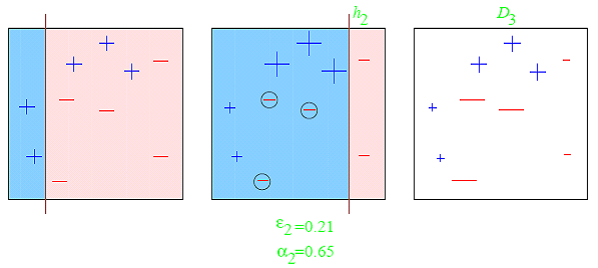
**Bagging** (Bootstrap Aggregating) is an ensemble method. First, we create random samples of the training data set (sub sets of training data set). Then, we build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting. Bagging helps to reduce the variance error. For example, we can train M different trees on different subsets of the data (chosen randomly with replacement) and compute the ensemble:



[](https://www.analyticsvidhya.com/wp-content/uploads/2015/09/bagging.png)

bagging uses voting for classification and averaging for regression.

**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.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/09/boosting1.png)

**5. How can you avoid over fitting?**

**Cross Validation:-**By using a lot of data over fitting can be avoided, over fitting 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 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.

**Early Stopping**: Early stopping rules provide guidance as to how many iterations can be run before the learner begins to over-fit.

**Pruning**: Pruning is used extensively while building CART models. It simply removes the nodes which add little predictive power for the problem in hand.

**Regularization**: This is the technique we are going to discuss in more details. Simply put, it introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term.