**Solution for Assignment on Machine Learning-I**

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

**Ans** : a) Model building

b) Model testing

c) Applying the model

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

**Ans**: 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?**

**Ans** : 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.  In sentiment analysis, if the training set is labeled correctly, the model should be able to learn something from features. However, 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.

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, are unreasonable, we’ll need to make some adjustments to our model and try again. Whereas Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner.

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

**Ans** : The general principle of ensemble model is to combine multiple ‘individual’ (diverse) models together and delivers superior prediction power. Basically, an ensemble is a supervised learning technique for combining multiple weak learners/ models to produce a strong learner. Ensemble model works better, when we ensemble models with low correlation.

**Bagging**:

**Bagging** 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.

**Boosting**:

**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.

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

**Ans:** “Overfitting” is traditionally defined as training some flexible representation so that it memorizes the data but fails to predict well in the future.

There are two styles of general overfitting: over-representing performance on particular datasets and (implicitly) over-representing performance of a method on future datasets.

Following methods to avoid overfitting;

* Focus on hypothesis generation
* Data Cleaning and Exploration (Find the relation between features)
* Right selection of method (machine learning technique)
* Cross-Validation of model