**Assignment7.2**

**Problem Statement**:

1. What are the three stages to build the hypotheses or model in machine learning?
2. What is the standard approach to supervised learning?
3. What is Training set and Test set?
4. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?
5. How can you avoid overfitting ?

**Answer:**

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

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

**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 labelled 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 are unreasonable, we’ll need to make some adjustments to our model and try again.

1. **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.

**Boosting**

**Boosting** (**machine learning**) **Boosting** is a **machine learning** ensemble meta-algorithm for primarily reducing bias, and also variance in supervised **learning**, and a family of **machine learning** algorithms that convert weak learners to strong ones.

**Bagging**

Bootstrap aggregating, also called **bagging**, is a **machine learning** ensemble meta-algorithm designed to improve the stability and accuracy of **machine learning**algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

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

**Cross-Validation** : Cross Validation in its simplest form is a one round validation, where we leave one sample as in-time validation and rest for training the model. But for keeping lower variance a higher fold cross validation is preferred.

**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.