**Response for Assignment 14.1**

Q1) What are the three stages of building a hypothesis or model in machine learning ?

Ans: Below are the 3 stages

1. **Data Preparation and Training Set generation** – This includes capturing the data and putting it in a specific format. This data is generally called training data or training set.
2. **Modelling by selecting appropriate Algorithm** – Evaluating which algorithm suits the requirement and fine tuning parameters to finally select the appropriate algorithm. This becomes the model
3. **Predicting using the model** – Using the selected model, we feed the data which needs output. This output is referred as predicted data or prediction.

Q2) What is the standard approach to Supervised Learning ?

Ans: The standard approach to Supervised Learning is as below

1. Training Data should have input (referred as X) and expected output (referred as y). It should be clear as what is to be predicted which means what is the desired output (y).
2. Input and output (within Training/Test Sets) can be series/list of multiple data points.
3. Appropriate model (Linear, Statistical, Probabilistic, Tree, etc) is applied on input (X) to predict output (y). This model is generally a function
4. Predicted Output (y of test set) is then compared with given output (y of test set) to get the accuracy.

Q3) What is Training Set and Test Set ?

Ans: During the stage of data preparation, it is always recommended to divide the data set into 2 parts – Training Set and Test Set. Usually, we divide in 70-30 ratio where Training Set is 70% of total data and remaining 30% is Test Set.

**Step 1**

Generally, initial model training happens on training set (70%)

**Step 2**

Test set (30%) is used to test the accuracy of model.

**Step 3**

Basis Over-fitting or Under-fitting, model is tweaked on training set. It is then further re-tested on test set to again calculate accuracy.

**Redo Step 1**

This process is re-iterated until we reach expected accuracy. Hence modeling is done on training set and tested on test set.

Q4) What is the general principle of ensemble method and what is bagging and boosting in ensemble method ?

Ans: Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Supervised algorithms are most commonly describedas performing the task of searching through a hypothesis space to find a suitable hypothesis that will make good predictions with a particular problem. Ensembles combine multiple hypotheses to form a (hopefully) better hypothesis. The term **ensemble** is usually reserved for methods that generate multiple hypotheses using the same base learner. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation.

**Bagging**: Bagging (Bootstrap aggregating) was originally proposed to improve classification by combining classifications of randomly generated training sets. However it is one of the ensemble techniques which also reduces the problems of overfitting of the training data. It creates additional training sets by sampling uniformly and with replacement. These samples are called bootstrap samples. Different models are fitted on these bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

**Boosting**: Boosting is an algorithm for primarily reducing bias, and also variancein supervised learning (and a family of machine learning algorithms) that convert weak learners to strong ones. Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models mis-classified. Most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When they are added, they are typically weighted in some way that is usually related to the weak learners’ accuracy. After a weak learner is added, the data are re-weighted: data points that are misclassified do gain weight and data points that are classified correctly do lose weight.

Q5) How can you avoid overfitting ?

Ans: During modelling or model creation, accuracy is determined by applying the model on test set. If the accuracy is high on training data and low on test set, it means that model is overly fitting the training data. This means that model is designed closely as per the training data. This also means that decision boundary is quite specific to training data. However, this model doesn’t give accurate results the moment it is applied on any unseen data (test set). This means that **model is overfitting** the training data. Below are **few methods to avoid Overfitting**

1. **Add more data**: If we add more data to training data, it will give variation and will be generalized well
2. **Remove Features**: Studying data, it is generally realized that there are multiple features which may not be required but just influencing the model/algorithm
3. **Regularization**: Use regularization parameter to simplify the algorithm
4. **Cross Validation**: Convert trainings sets into Cross Validation data sets to iteratively train the model
5. **Ensembling**: It combines predictions from multiple models using Bagging or Boosting (explained above).