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

Ans 1.a) Model building b) Model testing c) Applying the model

2. What is the standard approach to supervised learning?

Ans 2. 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 3. In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an examples given to the learner, while 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. Training set are distinct from Test set.

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

Ans: Ensemble model combines multiple ‘individual’ (diverse) models together and delivers superior prediction power.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.

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. It starts 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.

5. How can you avoid overfitting ?

Ans: One of the important reason and possibility of over fitting is because the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model ,We can avoid over fitting by using:

• Lots of data

• Cross-validation