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 examples into the training set and test set.

3.What is Training set and Test set?

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 example 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 examples 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?

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. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

5. How can you avoid overfitting?

By using a lot of data overfitting can be avoided, overfitting 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 datapoints 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.