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

Stage 1: Model building: In machine learning the engineer must be able to choose their model based on quality, quantity and accuracy that their model can bring on their data. Here the model is trained with a training set. Tested with the training set and re-trained by tweaking the learning parameters of the model. Which is 60-80% of the dataset.

Stage 2: Model Testing: The test data set is later tested on this model to check for the accuracy of the model to apply what it has learned. Which is 10-20% of the dataset. .(That can be changed). Validation 0 to 10% of the dataset.

Stage 3: Applying the Model: The model is then validated on a separate dataset.

1. What is the standard approach to supervised learning?

Splitting the dataset or corpus into training, testing and validation sets.

Usually 80% training and 20 % testing or

60% training, 10% testing and 10% validation.

1. What is Training set and Test set?

Training set is the data set on which the model is built on, where the learning parameters are set on based on the values of the features in the training dataset.

Test set is used to check the accuracy of the model’s learning.

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.  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.

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.