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

**Ans:** a) Model building b) Model testing c) Applying the model

**2. What is the standard approach to supervised learning?**

**Ans:** Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

**3. What is Training set and Test set?**

**Ans:** A **training** dataset is a dataset of examples used for learning, which is to fit the parameters (e.g., weghts) of, for example, a classifier.

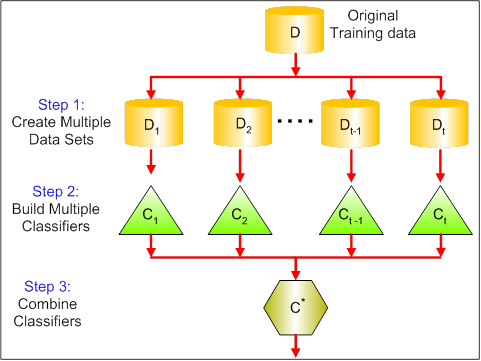
Most approaches that search through training data for empirical relationships tend to overfit the data, meaning that they can identify apparent relationships in the training data that do not hold in general.

A **test** dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place (see figure below). A better fitting of the training dataset as opposed to the test dataset usually points to overfitting.

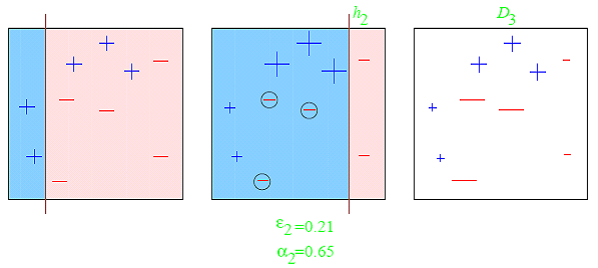
**4. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?**

**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.



Most common example of boosting is AdaBoost and Gradient Boosting

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

Ans: Methods to avoid Over-fitting:

Following are the commonly used methodologies:

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.