Lecture 18

Ensemble learning

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use.

Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance.

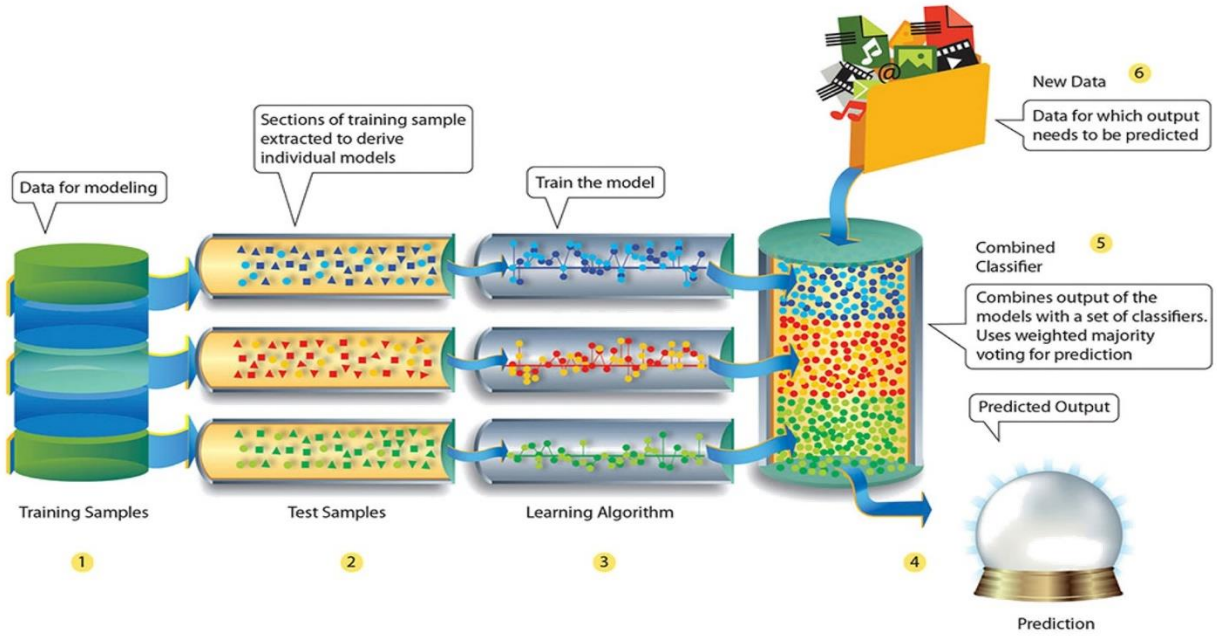
What is an ensemble method?

The main causes of error in learning models are due to **noise, bias and variance**. Ensemble methods help to minimize these factors. These methods are designed to improve the stability and the accuracy of Machine Learning algorithms.

Ensemble learning Technique:

Ensembles are a divide and conquer approach used to improve performance.

This technique is used to create multiple Machine Learning models, which are then combined to produce more accurate results. A general Machine Learning model is built by using the entire training data set. However, in Ensemble Learning the training data set is split into multiple subsets, wherein each subset is used to build a separate model. After the models are trained, they are then combined to predict an outcome in such a way that the variance in the output is reduced.



Different types of Ensemble learning methods:

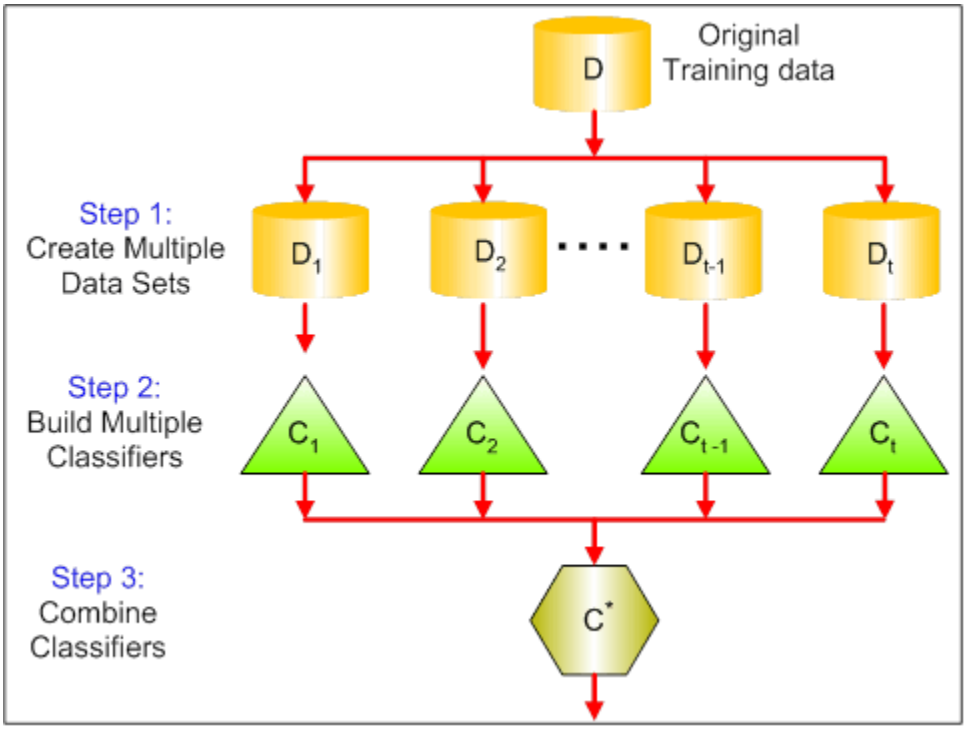
Although there are several types of Ensemble learning methods, the following two are the most-used ones in the industry.

Bagging based Ensemble learning:

Bagging is one of the Ensemble construction techniques which is also known as ***Bootstrap Aggregation***. Bootstrap Aggregating method works as following-

- First, we create random samples of the training data set with replacement (sub sets of training data set).
- Then, we build a model (classifier or Decision tree) for each sample.
- Finally, results of these multiple models are combined using average or majority voting.

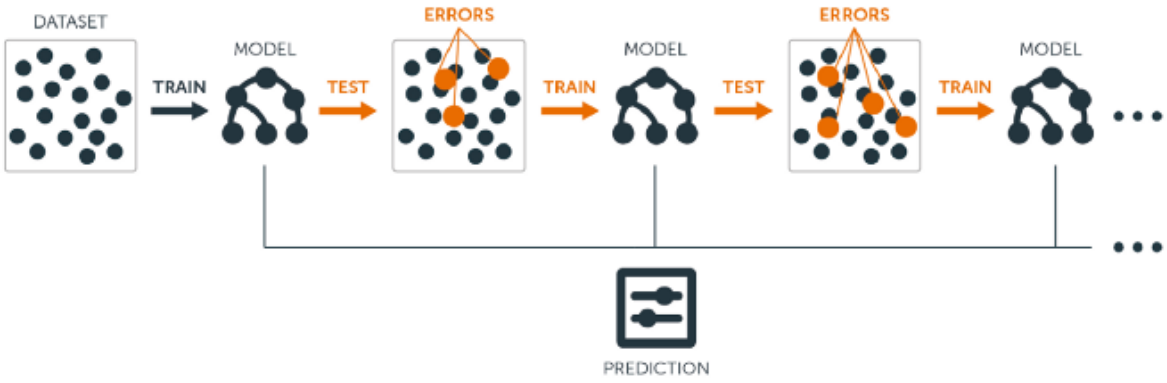
The following infographic gives a brief idea of Bagging:



Boosting-based Ensemble learning:

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa.

Boosting is a form of *sequential learning* technique. The algorithm works by training a model with the entire training set, and subsequent models are constructed by fitting the residual error values of the initial model. In this way, Boosting attempts to give higher weight to those observations that were poorly estimated by the previous model. Once the sequence of the models are created the predictions made by models are weighted by their accuracy scores and the results are combined to create a final estimation. Models that are typically used in Boosting technique are XGBoost (Extreme Gradient Boosting), GBM (Gradient Boosting Machine), ADABoost (Adaptive Boosting), etc.



Summary of differences between Bagging and Boosting

	Bagging	Boosting
Similarities	• Uses voting • Combines models of the same type	
Differences	Individual models are built separately	Each new model is influenced by the performance of those built previously
	Equal weight is given to all models	Weights a model's contribution by its performance

V. Experimental Design