Ensemble Methods

Let us consider a real world situation which uses Ensemble Methods, which is, when a user wants to buy a new product. Many users who have already purchased that product will have given either positive or negative ratings. If in the group, many users have given positive ratings, then the combined rating will be positive. Instead of a single rating, the ratings of the group of users is considered. The product is bought by the user when the combined ratings of the group is positive. The user gets a fairer idea about the product when all the ratings are combined.
Here, the combination of ratings is done so that the decision making process of the user is made easy.
Ensemble Methods refer to combining many different machine learning models in order to get a more powerful prediction.
Thus, ensemble methods increase the accuracy of the predictions.

Why use Ensemble Methods?

Ensemble Methods are used in order to:

- decrease variance (bagging)
- decrease bias (boosting)
- improve predictions (stacking)

Bagging

Bagging actually refers to Bootstrap Aggregators.
Bagging tests multiple models on the data by sampling and replacing data i.e it utilizes bootstrapping. In turn, this reduces the noise and variance by utilizing multiple samples. Each hypothesis has the same weight as all the others. Now, aggregating of the outputs of various models is done.

Boosting

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 in general decreases the bias error and builds strong predictive models.

Variance

Variance quantifies how the predictions made on same observation are different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training. Thus, we aim at low variance.

Bias

Bias error is useful to quantify how much on on average are the predicted values different from the actual value. A high bias error means we have a under-performing model. Thus, we aim at low bias.
A commonly used class of ensemble methods are forests of randomized trees.
In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e. a bootstrap sample) from the training set. In addition, instead of using all the features, a random subset of features is selected, further randomizing the tree.

As a result, the bias of the forest increases slightly, but due to the averaging of less correlated trees, its variance decreases, resulting in an overall better model.