Random Forest

Ensembled Learning

- Random Forest is an ensembled learning approach
- In ensembled learning approach, multiple predictive models are developed and results are aggregated to improve the precision

Random Forest

- In this algorithm, the observations as well as variables are sampled to create multiple decision trees
- Each observation is classified by each decision tree.
- The outcome is considered as per the majority in different trees

Algorithm
(Considering N observations & M variables)

1. Sample out of N cases with replacement from the training set, many samples. Consider each sample as root node for the decision trees to be constructed.
2. Choose some m < M number of variable by sampling at each node created in step 1.
3. Grow each tree without pruning with minimum node size as 1
4. Classify the validation / test set observations by traversing them through all the grown trees.
5. Classify each outcome by a majority vote of the trees.

OOB

- Each tree is constructed using a different bootstrap sample from the original data.
- About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.
- Put each case left out in the construction of the kth tree, traverse the kth tree to get a classification.
- In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was oob.
- The proportion of times that j is not equal to the true class of n averaged over all cases is the oob error estimate. This has proven to be unbiased in many tests.
- This is done internally with the training set

Variable Importance

- The importance gives the list of variables along with the measure of their importance in terms of the purity gained by them.