Random Forest

Aaryan
CO21BTECH11001

Let number of features of dataset = n

Let number of training examples = m

Data consists of matrices X and y where i^{th} column of X represents the i^{th} feature of dataset and i^{th} element of y represents the value of variable dependent on set of features listed in i^{th} row of X.

Random forest is a bagging algorithm which consists of bagged decision trees, with a slightly modified splitting criteria.

The algorithm works as follows –

1. Sample p datasets $D_1, D_2, \dots, D_p$ with replacement.
2. For each D_j, train a full decision tree $h_j(\,)$ with one small modification: before each split, randomly subsample $k \leq n$ features (without replacement) and only consider these for split i.e., the feature with least impurity among these for split.
3. The final classifier $h(x) = mode_j\{h_j(x)\}$

Questions –

1. What kind of algorithm is Random Forest?
   Ans. It is a supervised learning algorithm widely used for classification/labeling problems.

2. How do we choose the parameter k?
   Ans. k is supposed to be chosen by handpicking i.e., change and find the appropriate value suitable for the problem. But a good estimate for starting is to take the round off value of $\sqrt{n}$

3. How do we choose the parameter m?
   Ans. This depends on requirement of problem, but for a higher accuracy, m is taken as a couple of thousands. And it can be as large as one can afford.

4. What are the advantages of using Random Forest?
   Ans. It doesn't require any kind of pre-processing of data. Since it is a splitting algorithm, it works the same irrespective of scale. Also the result is extremely insensitive to parameters m and k.

5. What are the disadvantages of using Random Forest?
   Ans. The main limitation is that large number of trees can make the algorithm too slow and ineffective for real-time predictions.

6. How can we increase accuracy and decrease time complexity of Random Forest?
   Ans. We should not grow each tree to its full depth, instead prune based on the leave out samples. This can further improve your bias/variance trade-off.