Random Forest

Kaibin Yin, Xiao Wang

Random Forest

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

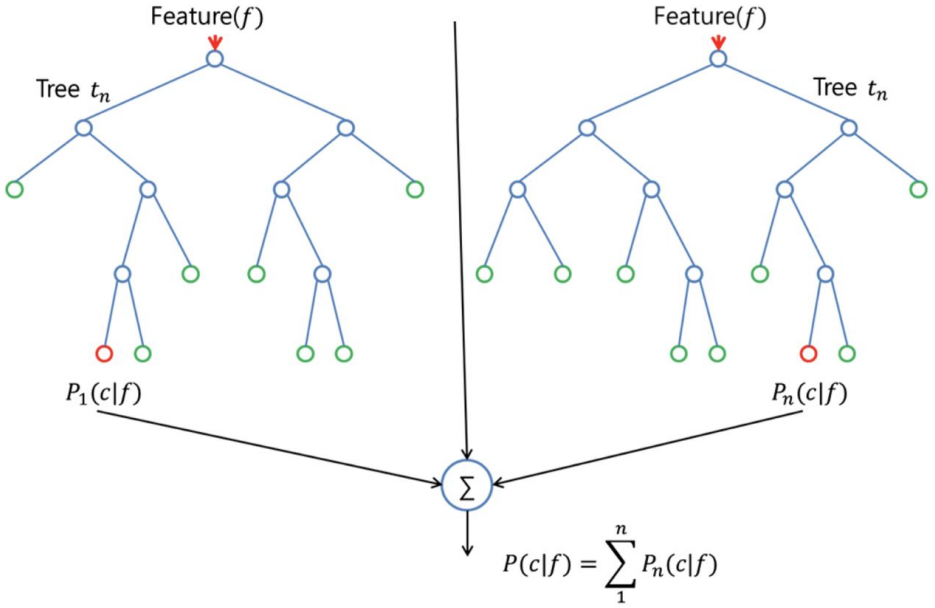

image: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Parameters

- **numClasses**: Number of classes. In this case is 2: 0 for background, and 1 for foreground.
- **numTrees**: Number of trees in the forest.
- **maxBins**: Number of bins used when discretizing continuous features.
- **impurity**: Impurity measure (discussed above) used to choose between candidate splits. This measure must match the algo parameter. In this case, we use gini index as impurity measurement.
- **maxDepth**: Maximum depth of each tree in the forest. Deeper trees are more expressive (potentially allowing higher accuracy), but they are also more costly to train and are more likely to overfit. As we have about more than 3000 features, it is reasonable to have trees with 10 or 11 depth at max.

Tuning

maxDepth numTrees	16	64
8	99.7404%	99.746%
10	99.7439%	99.759%

Possible improvements

- Train with bigger maxDepth and numTrees
- Rotate/mirror images before training
- Another possible algorithm for this case is Gradient Boosting Tree. But it could be more time consuming since it train one tree at one time.