Random Forest

Houcine Senoussi

October 8, 2019

What is it about ?

- A **supervised** learning technique introduced by L. Breiman in the early 2000's.
- Used for **classification** and regression.
- It aims to build a classifier (the RF) consisting of a collection of decision trees grown on subsets of the original data.
- Each classifier is defined by :
    1. an horizontal radom selection (on observations, a bootstrap).
    2. a vertical random selection (on variables) at each node.
- The random forest prediction is obtained by taking the **majority vote** of the trees in the case of classification and the average over their predictions in the case of regression.
- In this chapter, we consider exclusively classification problems in which all the variables are **categorical**.

Definitions

the idea of random forest can be formalized as follows :

1. Input :
   1. The training set $D = \{O_i = (X_i, Y_i)\ i = 1, ..., n\}$. The instances X_i are described by p features (categorical or binary variable) $F_1, .., F_p$. The class Y is a categorical or binary variable which takes its value in a set $\{C_1, ..., C_s\}$.
   2. *ntree* : the number of the trees to build.
   3. *mtry* : the number of features to try for each node.
2. Output :
   1. The random forest *RF* : a set of *ntree* trees.

Definitions

1. Algorithm :
   1. Draw *ntree* bootstrap samples D_k from D.
   2. For $k = 1, ..., ntree$ use D_k to build the decision tree T_k by recursively repeating the following steps for each node until the stopping criterion is met :
      - Randomly select m_{try} features.
      - Pick the best feature according to the splitting criterion (see below) among the m_{try}.
      - Use this feature to split the node into two nodes.
2. To make a prediction at a new instance x :
   - $RF(x) = \text{majority vote } \{T_k(x)\}$

Splitting criterion

- During the building of a tree, each node N represents a subset D_N of D.
- Splitting N means partitionning D_N into two subsets D_{NL} and D_{NR}, each one corresponding to some values of F, the feature we use in splitting.
- We aim to reach leaves of the tree as quickly as possible.
  - Leaves are nodes corresponding to "pure" subsets (subsets in which majority or totality of observations belong to the same class).
- $\Rightarrow$ A "good" feature as a feature that improves purity. In other words, F is good in N, if using it to split N results in two subsets as pure as possible.
- We need a splitting criterion that characterizes node/subset purity.

Splitting criterion-2

- Let us denote by n_N the cardinality of D_N (node size). For each value C_j of the class Y, let $p_N j$ the proportion of class C_j observations in D_N.
- In other words :
  - $p_{Nj} = \frac{1}{n_N} \sum_{O_i \in D_N} I(Y_i = C_j)$

  where $I()$ is the indicator function.

Splitting criterion-2

This proportion is used to define impurity measures $Q(D_N)$. One of the most important is Gini index defined as follows :

- $Q(D_N) = 1\text{-}\sum_{j=1}^{q} p_{Nj}^2$

Using this impurity measure, we define the best feature at each node as the one that maximizes the Gini Information Gain defined as follows :

- $IG(N, F) = Q(D_N)\text{ - }(\frac{n_{NL}}{n_N}Q(D_{NL}) + \frac{n_{NR}}{n_N}Q(D_{NR}))$

In other words, the best feature is the one that maximizes the impurity reduction.

Stopping criterion

It is simply defined by the minimum node size. This parameter will be noted *ndsize*.

Output

1. *OOB*-Error : for each observation $O_i = (X_i, Y_i)$, let us aggregate the votes only over those trees T_k whose bootstrap sample D_k does not contain O_i. The classifier thus obtained is called the *out-of-bag* (OOB) classifier. The error rate of this classifier on the training set is called the *Out-of-bag* error.
2. Test-Error : obtained by applying the RF to a test Set.
3. Variable importance : There are several ways to measure variable importance but the most widely used is also called permutation importance. When the tree T_k is created, its prediction accuracy is estimated using its OOB sample. Then, the values for each feature F are randomly permuted and the new prediction accuracy of T_k is computed. A measure of importance of F is obtained by averaging the decrease in accuracy due to these permutation.

References

- T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning.