Decision Trees and Random Forest

Unfortunately there is no one clear algorithm on how to build binary decision trees and random forests. Rather eac

Preliminaries

We start with our training data $\mathcal{D} = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\}$ where $\mathbf{X}_i \in [0,1]^d$ and $Y_i \in R$ is a continuous response

Random forests are often used in setups with a large amount of input variables of which only a few actually determine t

Decision Tree A decision tree is a collection of recursive binary splits which aim to partition the data into more and

At each step, one variable and one split point are chosen which determine how the data will be partitioned. Consider th

The question arising now is how to choose the splitting dimension and the split point. Breiman et al. (1984) propos

The variance of a node $\mathbf{t}$ is given by

$\hat{\Delta}(\mathbf{t}) = \frac{1}{N(\mathbf{t})} \sum_{\mathbf{X}_i \in \mathbf{t}} (Y_i - \hat{Y}_{\mathbf{t}})^2$

where $\hat{Y}_{\mathbf{t}}$ is the sample mean of the responses and $N(\mathbf{t})$ is the number of observations in node $\mathbf{t}$ respectively. To ass

$\hat{\Delta}(s; \mathbf{t}) = \hat{\Delta}(\mathbf{t}) - [\hat{P}(\mathbf{t}_L)\hat{\Delta}(\mathbf{t}_L) + \hat{P}(\mathbf{t}_R)\hat{\Delta}(\mathbf{t}_R)]$

with $\hat{P}(\mathbf{t}_L) = \frac{N(\mathbf{t}_L)}{N(\mathbf{t})}$ and $\hat{P}(\mathbf{t}_R) = \frac{N(\mathbf{t}_R)}{N(\mathbf{t})}$ being the fractions of observations falling into the left and right child node

The estimator for a terminal node $\mathbf{t}$ is given by the sample mean of all observations falling into that node: $\hat{Y} = \bar{Y}_{\mathbf{t}}$.

While single decision trees have been empirically shown to have rather low biases, they do exhibit high variance. Th

Random Forest

Another remedy to reduce the variance of a decision tree while trying to keep the favorable property of low bias wa

Literature Overview

The theoretical properties of random forests have been studied quite intensively in the last couple of years. Decision

It is especially the data-dependency of the splitting points which complicates the analysis of random forest estimators. I

The conditions for consistency of data-independent estimators have been proposed by Stone (1977). Decision trees being

Biau (2012) was the first to show consistency results for an algorithm which more closely resembles the original random

Klusowski (2019) adds to the literature by proving consistency for a more general class of response surfaces which go be