DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Fast Linear Model Trees by PILOT

Jakob Raymaekers[1], Peter Rousseeuw[2], Tim Verdonck[3], and Ruicong Yao[2]

[1]Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands
[2]KU Leuven, Department of Mathematics, Celestijnenlaan 200B, Leuven 3001, Belgium
[3]University of Antwerp, Department of Mathematics, Middelheimlaan 1, Antwerp 2020, Belgium

Linear model trees are regression trees that incorporate linear models in the leaf nodes. This preserves the intuitive interpretation of decision trees and at the same time enables them to better capture linear relationships, which is hard for standard decision trees. But most existing methods for fitting linear model trees are time consuming and therefore not scalable to large data sets. In addition, they are more prone to overfitting and extrapolation issues than standard regression trees. In this paper we introduce PILOT, a new algorithm for linear model trees that is fast, regularized, stable and interpretable. PILOT trains in a greedy fashion like classic regression trees, but incorporates an L^2 boosting approach and a model selection rule for fitting linear models in the nodes. The abbreviation PILOT stands for **PI**ecewise **L**inear **O**rganic **T**ree, where 'organic' refers to the fact that no pruning is carried out. PILOT has the same low time and space complexity as CART without its pruning. An empirical study indicates that PILOT tends to outperform standard decision trees and other linear model trees on a variety of data sets. Moreover, we prove its consistency in an additive model setting under weak assumptions. When the data is generated by a linear model, the convergence rate is polynomial.