Median-Based Splitting Rules For The Causal Tree

Jens Klenke[(1)], Lennard Maßmann[(1,2)]

[(1)] Chair of Econometrics, University of Duisburg-Essen; Germany
[(2)] Ruhr Graduate School in Economics; Germany

Introduction

Our paper contributes to the literature on tree-based methods for causal inference and treatment effect estimation in high-dimensional data settings. We investigate splitting rules of causal trees, which are a variation of standard regression trees (**atheyRecursive2016?**). Tree-based methods can be used for causal inference and treatment effect estimation, especially when the number of features is large in relation to the number of observations. We consider the potential outcome framework with i.i.d. sampled data (X_i, Y_i, D_i) for $i = 1, \dots, N$ with a binary treatment indicator D_i, the outcome variable Y_i and X the feature matrix of dimension $N \times K$ with $k = 1, \dots, K$. A CART regression tree Π divides the feature space into separate segments in a recursive manner until reaching a set of leafs $\Pi = \{l_1, \dots, l_L\}$, aiming at precisely estimating the expectation of the outcome variable conditional on the regressors, $\mu(X_i = x) = \mathbb{E}[Y_i|X_i]$. Given SUTVA, unconfoundedness and positivity assumptions, the CATE can be identify as

$$\tau(X_i = x) = \mathbb{E}[Y_i(D_i = 1) - Y_i(D_i = 0)|X_i = x].$$

Results of a simulation study suggest that the proposed median-based splitting rules improve CATE coverage if covariates are independent. When covariates are correlated, all considered splitting rules provide too wide conformal intervals leading to overcoverage.

Splitting Rules

Median-based splitting rules can be used as an alternative to spliting rules which are based on the mean-squared error. With focus on treatment effect analysis rather than predictive outcome accuracy, we develop three different splitting rules based on the Median Absolute Deviation (MAD), Median Squared Deviation (MSD) and Least Median Square (LMS). Since we are not able to observe both potential outcomes $(Y_i(1), Y_i(0))$ for each unit i, we use the Hodges-Lehmann estimator to estimate the median of the difference between the observed treatment and control groups (**roy_robustness_2012?**).

$$\widehat{\text{MAD}}_\tau(\mathcal{S}^{tr}, S^{tr}, \Pi) = \Big| \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}(X_i; \mathcal{S}^{tr}, \Pi) - \hat{\tau}_{HL}(X_i; \mathcal{S}^{tr}, \Pi) \Big|,$$

$$\widehat{\text{MSD}}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi) - 2\hat{\tau}(X_i; \mathcal{S}^{tr}, \Pi)\hat{\tau}_{HL}(X_i; \mathcal{S}^{tr}, \Pi),$$

$$\widehat{\text{LMS}}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) = \underset{i \in \mathcal{S}^{tr}}{\text{med}} \left[(Y_i - \hat{\mu}(D_i, X_i; \mathcal{S}^{tr}, \Pi))^2 \right].$$

Conclusion / Outlook

We implement three new median-based splitting rules into the causal tree for CATE estimation. Simulation results suggest competitive RMSE-results and improved conformal coverage rates in comparison to benchmark splitting rules, at least for LMS. An application to the Lalonde dataset demonstrates the usage in a more realistic setting. As next steps, we plan to examine CATE coverage more thoroughly across methods (conformal vs. bootstrap intervals) and settings (higher dimensions, more complex variable interactions).

References