Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2021:11**
http://www.math.su.se

Carving Out Borders and Searching for Roots Among the Forest Kingdoms

Aron Södergren*

June 2021

Abstract

We compare the three tree-based gradient boosting methods XGBoost, LightGBM and CatBoost for binary classification tasks. We compare these by AUC scores and training time on data sets simulated from a logistic regression model with varying number of instances, categorical features and different degrees of cardinality in these categorical features. We use Bayesian hyperparameter optimization for hyperparameter tuning for all boosting methods and data sets. The goal of the study is to bring some light to why the performance of these boosting methods varies when they are applied to real-world data. This is of importance since gradient boosting growths ever more popular for binary classification tasks, but also because the relationship between the performance of these methods and different data characteristics is poorly researched at the moment. Furthermore we exchanged different components in the boosting methods to identify which parts that cause the variation in results, the goal here was to get a deeper understanding of how these methods work. For simulation scenarios with a high number of instances (100.000) and no categorical features of high cardinality, XGBoost and CatBoost was more accurate than LightGBM. For scenarios with a lower number of instances or with categorical features of high cardinality, CatBoost proved the most accurate, however when both number of instances was high and the cardinality of categorical features was high, LightGBM was equally accurate to CatBoost. When comparing components we found that gradient-based one-side sampling increased the speed for all scenarios, but accuracy was compromised for small data sets with categorical features. Exclusive feature bundling reduced training time when used with one-hot encoded categorical features of high cardinality. We found no significant difference in accuracy or training time between leaf-wise and level-wise splitting. Weighted quantile sketch improved the accuracy of histogram search. Naive target statistics increased accuracy for data sets with high cardinality categorical features and large number of instances when compared to one-hot encoding, this effect was reversed when number of instances was small, in both cases naive target statistics decreased training time. Similarly, ordered target statistics increased accuracy for all data sets with high cardinality categorical features, this however came at the cost of higher training time.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: aron.s.sodergren@gmail.com. Supervisor: Taras Bodnar and Pieter Trapman.