A Novel Approach to Forest-Level Splitting for Regression Models

Abstract

1 Introduction

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 introduces Dynaforest's methodology, Section 4 describes our experimental setup, and Section 5 presents results and analysis. Finally, Section 6 concludes with insights and future directions.

2 Related Work

3 Dynaforest Methodology

3.1 Overview

3.2 Mathematical Foundation

In regression tree analysis, determining the optimal split point within a leaf node involves selecting a value, denoted as α, that minimizes the sum of loss functions over all observations in the node. This is mathematically represented as minimizing

$$\sum_{i=1}^{n} L(y_i, \alpha),$$

where L denotes the loss function and y_i are the observed values. For regression tasks employing the mean squared error (MSE) as the loss function, $\alpha = \bar{y}$.

Building upon this, it has been demonstrated [BY WHOM] that the objective function can be simplified to:

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}.$$

This formulation is computationally advantageous, as it allows for the calculation of optimal sums in $O(n)$ time after sorting, thereby enhancing efficiency. I n our approach, we extend this methodology by incorporating predictions from an ensemble of M trees within a specified window. The computation is performed as follows:

$$
\begin{aligned}
E = \\
\sum_{i=1}^{n}(y_i - \frac{Mo_i + \frac{1}{n}\sum_{i=1}^{n} y_i}{M+1})^2 = \\
\sum_{i=1}^{n}(y_i - \frac{Mo_i + \bar{y}_n}{M+1})^2 = \\
\sum_{i=1}^{n} y_i^2 - \frac{2My_io_i + 2y_i\bar{y}_n}{M+1} + \frac{(Mo_i + \bar{y}_n)^2}{(M+1)^2} = \\
\sum_{i=1}^{n} y_i^2 - \frac{2M}{M+1}\sum_{i=1}^{n} y_io_i - \frac{2}{M+1}n\bar{y}_n^2 + \frac{M^2}{(M+1)^2}\sum_{i=1}^{n} o_i^2 + \frac{2M}{(M+1)^2}\bar{y}_n\sum_{i=1}^{n} o_i + \frac{\sum_{i=1}^{n} u_n^2}{(M+1)^2} = \\
\sum_{i=1}^{n} y_i^2 - \frac{2M}{M+1}\sum_{i=1}^{n} y_io_i - \frac{2}{M+1}n\bar{y}_n^2 + \frac{M^2}{(M+1)^2}\sum_{i=1}^{n} o_i^2 + \frac{2M}{(M+1)^2}\bar{y}_n\sum_{i=1}^{n} o_i + \frac{(\sum_{i=1}^{n} y_i)^2}{n(M+1)^2} = \\
\sum_{i=1}^{n} y_i^2 - \frac{2M}{M+1}\sum_{i=1}^{n} y_io_i - \frac{2(\sum_{i=1}^{n} y_i)^2}{n(M+1)} + \frac{M^2}{(M+1)^2}\sum_{i=1}^{n} o_i^2 + \frac{2M\sum_{i=1}^{n} y_i}{n(M+1)^2}\sum_{i=1}^{n} o_i + \frac{(\sum_{i=1}^{n} y_i)^2}{n(M+1)^2}
\end{aligned}
$$

Again, we only need to pre-calculate cumulative sums for

$$\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} o_i, \sum_{i=1}^{n} o_iy_i, \sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} o_i^2$$

[Add a few sentences here, AND mention the possibility of replacing the mean with some specified value]