Bivariate Decision Trees: Smaller, Interpretable, More Accurate

Rasul Kairgeldin and **Miguel Á. Carreira-Perpiñán**
Dept. of Computer Science and Engineering, UC Merced

KDD2024 BARCELONA, SPAIN

1 Introduction

Univariate decision trees, commonly used since the 1950s, predict by asking questions about a single feature in each decision node. While they are interpretable, they often lack competitive predictive accuracy due to their inability to model feature correlations. Multivariate (oblique) trees use multiple features in each node, capturing high-dimensional correlations better, but sometimes they can be difficult to interpret.We advocate for a model that strikes a useful middle ground: bivariate decision trees, which use two features in each node. This typically produces trees that not only are more accurate than univariate trees, but much smaller, which offsets the small increase in node complexity and keeps them interpretable. They also help data mining by constructing new features that are useful for discrimination, and by providing a form of supervised, hierarchical 2D visualization that reveals patterns such as clusters or linear structure. We give two new algorithms to learn bivariate trees: a fast one based on CART; and a slower one based on alternating optimization with a feature regularization term, which produces the best trees while still scaling to large datasets.

Work partially supported by NSF award IIS–2007147.

2 Learning bivariate trees with TAO

We establish the following objective function over all parameters of a tree:

$$\min_{\Theta} E(\Theta) = \sum_{n=1}^{N} L(y_n, T(\mathbf{x}_n; \Theta)) + \lambda \sum_{i \in \mathcal{N}_{\text{dec}}} \phi(\mathbf{w}_i), \text{s.t. } \|\mathbf{w}_i\|_0 \leq 2,\ b_i \in \mathbb{R},\ i \in \mathcal{N}_{\text{dec}}; \quad (1)$$
$$c_j \in \{1, \dots, K\},\ j \in \mathcal{N}_{\text{leaf}}$$

where $L(\cdot,\cdot)$ is $0/1$ loss function. Furthermore, we introduce the following regularization:

$$\phi(\mathbf{w}_i) = \begin{cases} C, & \text{if } \|\mathbf{w}_i\|_0 = 2 \\ \|\mathbf{w}_i\|_0, & \text{if } \|\mathbf{w}_i\|_0 < 2 \end{cases}$$

Separability condition implies that equation 1 can be separated and optimized over parameters of any non-descendant nodes (located on the same depth) independently and in parallel. **Reduced problem over a node** (RP) states that optimizing equation 1 over parameters of the given node $i \in \mathcal{N}$ reduces to simpler, well-defined problem involving its reduced set $\mathcal{R}_i$.
For leaf $i \in \mathcal{N}_{\text{leaf}}$ the exact solution of RP is a majority class of samples in $\mathcal{R}_i$.
For decision node $i \in \mathcal{N}_{\text{dec}}$ RP is $0/1$ *loss binary classification problem*:

$$E_i(\mathbf{w}_i, b_i) = \sum_{n \in \mathcal{R}_i} L(\bar{y}_n, f_i(\mathbf{x}_n; \mathbf{w}_i, b_i)) + \lambda\phi(\mathbf{w}_i),\ \text{s.t. } \|\mathbf{w}_i\|_0 \leq 2,\ b_i \in \mathbb{R} \quad (2)$$

where L is a $0/1$ loss and $\bar{y}_n \in \{\texttt{left}, \texttt{right}\}$ corresponds to a pseudolabel assigned to a training instance x_n, signifying the child that yields a lower loss value. The loss is computed by propagating a sample through the corresponding child.

input training set $\{\mathbf{x}_n, \bar{y}_n\}_{n \in \mathcal{R}_i}$ of decision node $i \in \mathcal{N}_{\text{dec}}$,
  matrix of orientations $\mathbf{W} \in \mathbb{R}^{2 \times H}$
for each pair of features $j, k \in D$
  for $\mathbf{w}_l \in \mathbf{W}$
    $\mathbf{x}_l^{j,k} \leftarrow$ project selected features onto w_l
    $b_l^{j,k} \leftarrow$ optimal thresholding over $\mathbf{x}_l^{j,k}$
    if $j, k, \mathbf{w}_l, b_l^{j,k}$ produce lowest value of eq. 2
      $\theta_i^{biv} \leftarrow \{\mathbf{w}^*, b_l^{j,k}\}$, where $\mathbf{w}^*$ is a sparse vector of all zeros with corresponding value of $\mathbf{w}_l$ at j, k
    end if
  end for
end for
return θ_i^{biv}

Figure: Pseudocode of bivariate solution.

3 Effect of regularization and Interpretability

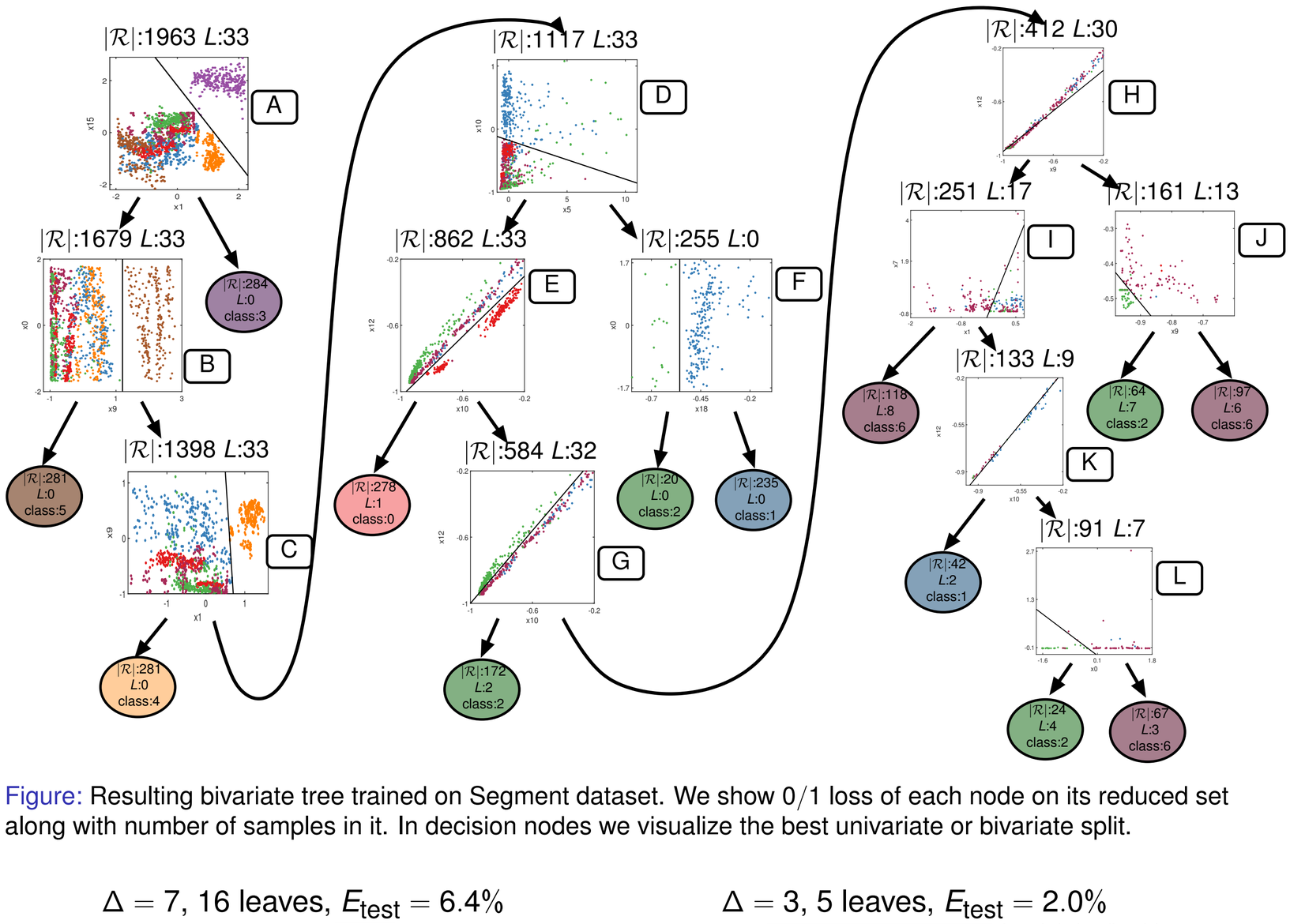


Figure: Resulting bivariate tree trained on Segment dataset. We show $0/1$ loss of each node on its reduced set along with number of samples in it. In decision nodes we visualize the best univariate or bivariate split.

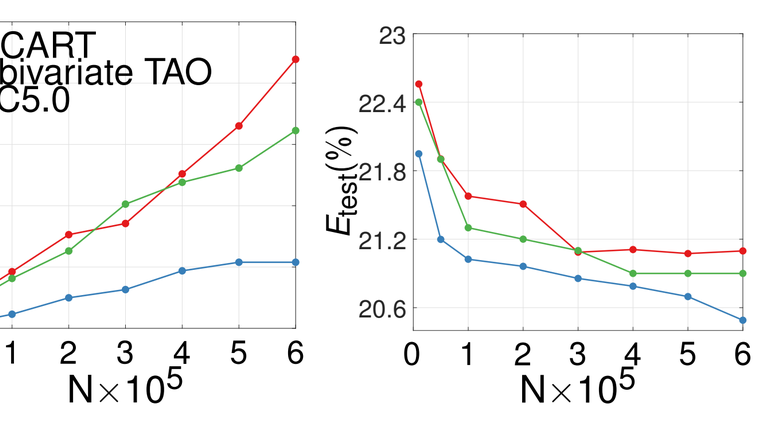
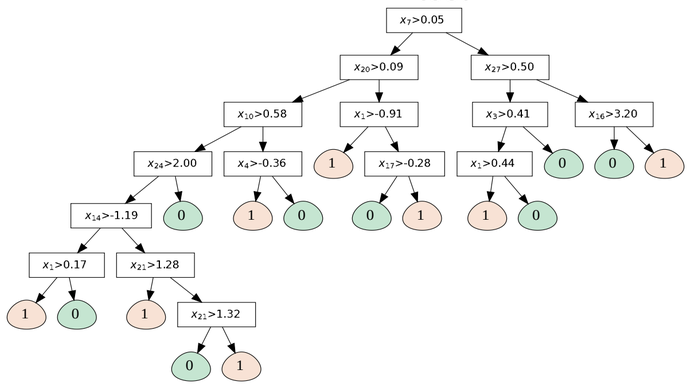
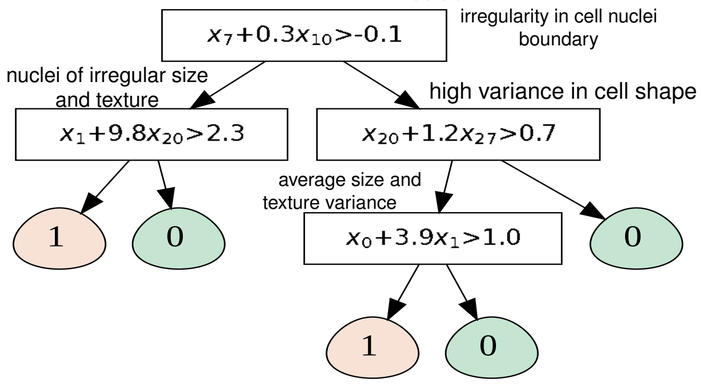


Figure: Illustration of our approximate solution of the RP at a decision node. The instances in the reduced set of the node are labeled according to their pseudolabels (preferred child, left ∘ or right +). The thin red lines are all the possible thresholds (passing through midpoints between projected instances) for the red orientation.

$\Delta = 7$, 16 leaves, $E_{\text{test}} = 6.4\%$



$\Delta = 3$, 5 leaves, $E_{\text{test}} = 2.0\%$



1: **if** $(x_7 > 0.05)$ & $(x_{20} > 0.09)$ & $(x_{10} > 0.58)$ & $(x_{24} > 2.00)$ & $(x_{14} > -1.19)$ & $(x_1 > 0.17)$ **then** PREDICT 1
2: **if** $(x_7 > 0.05)$ & $(x_{20} > 0.09)$ & $(x_{10} > 0.58)$ & $(x_{24} > 2.00)$ & $(x_{14} > -1.19)$ & $(x_1 \leq 0.17)$ **then** PREDICT 0
3: **if** $(x_7 \leq 0.05)$ & $(x_{27} \leq 0.50)$ & $(x_{16} \leq 3.20)$ **then** PREDICT 1
4: **if** $(x_7 \leq 0.05)$ & $(x_{27} \leq 0.50)$ & $(x_{16} > 3.20)$ **then** PREDICT 0
5: ...//**12 MORE RULES**...

1: **if** $(x_7 + 0.3x_{10} > -0.1)$ & $(x_1 + 9.8x_{20} > 2.3)$ **then** PREDICT 1
2: **if** $(x_7 + 0.3x_{10} > -0.1)$ & $(x_1 + 9.8x_{20} \leq 2.3)$ **then** PREDICT 0
3: **if** $(x_7 + 0.3x_{10} \leq -0.1)$ & $(x_{20} + 1.2x_{27} \leq 0.7)$ **then** PREDICT 0
4: **if** $(x_7 + 0.3x_{10} \leq -0.1)$ & $(x_{20} + 1.2x_{27} > 0.7)$ & $(x_0 + 3.9x_1 > 1.0)$ **then** PREDICT 1
5: **else** PREDICT 0

Figure: Best univariate CART tree and bivariate TAO tree with their sets of rules (Breast Cancer dataset). Decision nodes of bivariate tree are annotated with their meaning.

4 The size of the pruned tree