Shattering Inequalities for Learning Optimal Decision Trees

Justin J. Boutilier, Carla Michini, and Zachary Zhou

University of Wisconsin-Madison

Background

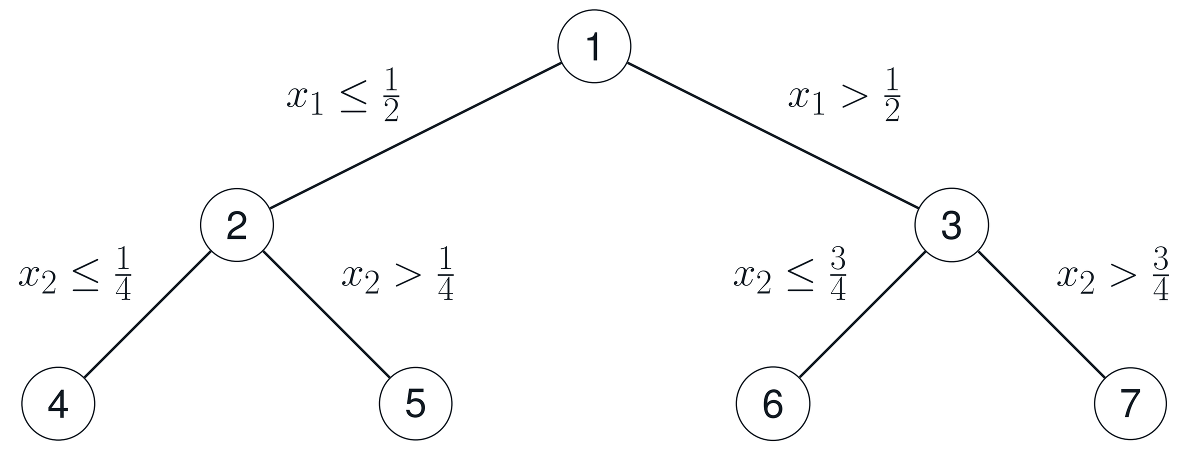

Fig. 1: An example of a univariate decision tree.

- Decision trees are among the most popular techniques for interpretable machine learning
- Observations begin at the root node and are guided down the tree via tests at each branch node until they reach a leaf node where they are classified
- The problem of learning an *optimal* decision tree is NP-hard, where optimality criteria may include accuracy, size of the tree, etc.; it is the subject of recent literature, both within and outside of the MIP community
- Many formulations and techniques now exist for learning optimal *univariate* decision trees, which perform tests involving only a single feature at branch nodes
- Considerably less work has been done pertaining to *multivariate* decision trees, which perform tests involving multiple features
- Although they may seem less interpretable than univariate splits, multivariate splits allow the decision tree to capture hyperplane boundaries more succinctly and accurately

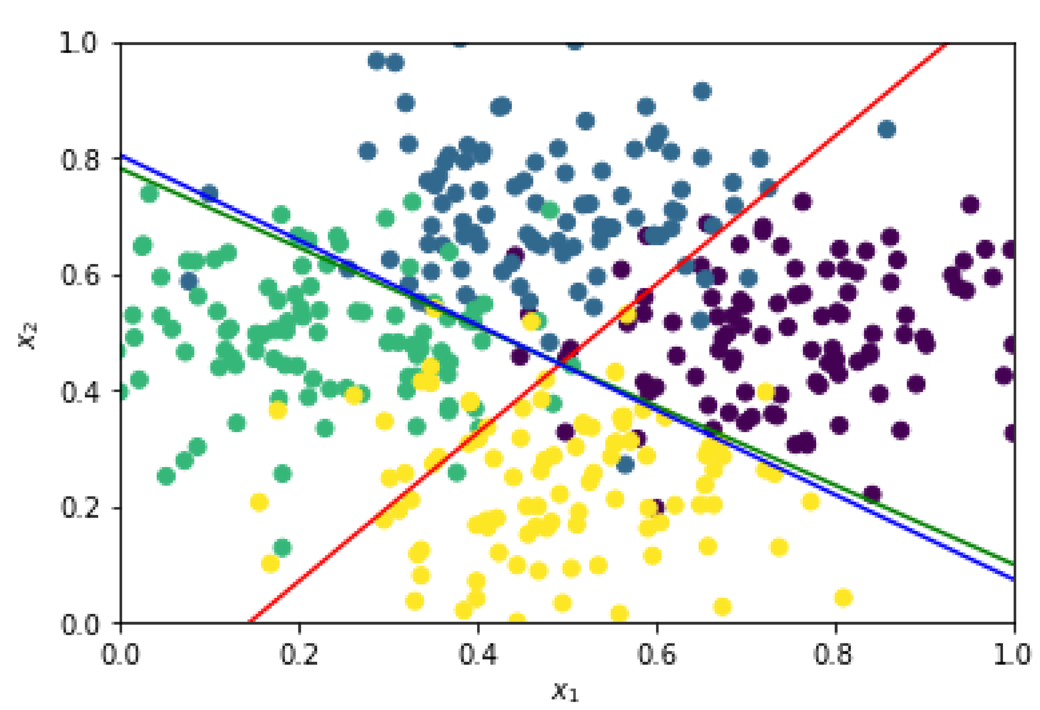

Fig. 2: A 4-class dataset in $\mathbb{R}^2$ that demonstrates the effectiveness of multivariate splits. A multivariate decision tree of depth 2 is sufficient to learn optimal decision boundaries, which are the diagonal lines. In contrast, a univariate tree is unable to capture these diagonal boundaries and thus generalizes poorly.

Our Contribution

Our goal is to efficiently compute optimal multivariate decision trees using MIP techniques. We propose a MIP model, and provide a class of valid inequalities for learning optimal multivariate decision trees. We show that our model can be solved using a Benders-like decomposition, where our valid inequalities can be used as feasibility cuts.

Notation

- Data:
  - Training set: N observations, p numerical features, K classes:
  $$\left\{(\mathbf{x}^i, y^i) \in [0,1]^p \times [K]\right\}_{i=1}^{N}$$
  - Formulation defined over full binary tree of depth $D \in \mathbb{N}$:
    - Branch nodes $\mathcal{B} = \{1, \dots, 2^D - 1\}$; $\forall t \in \mathcal{B}$, learn parameters $(\mathbf{a}_t, b) \in \mathbb{R}^p \times \mathbb{R}$:
      - If $\mathbf{a}_t^\top \mathbf{x} \le b_t$, then observation $\mathbf{x}$ is sent to t's left child $2t$
      - Otherwise, $\mathbf{x}$ is sent to t's right child $2t+1$
    - Leaf nodes $\mathcal{L} = \{2^D, \dots, 2^{D+1} - 1\}$; $\forall t \in \mathcal{L}$, assign a class $k \in [K]$
- Decision variables:
  - $c_{kt} \in \{0,1\},\ \forall k \in [K],\ t \in \mathcal{L}$: equals 1 iff leaf node t assigned class label k
  - $d_t \in \{0,1\},\ \forall t \in \mathcal{B}$: equals 1 iff branch node t applies a split
  - $w_{it} \in \{0,1\},\ \forall i \in [N],\ t \in \mathcal{B} \cup \mathcal{L}$: equals 1 iff observation i reaches node t
  - $z_{it} \in \{0,1\},\ \forall i \in [N],\ t \in \mathcal{L}$: equals 1 iff observation i is sent to leaf t and is correctly classified as y^i
  - $(\mathbf{a}_t, b_t) \in \mathbb{R}^p \times \mathbb{R},\ \forall t \in \mathcal{B}$: the hyperplane defining the multivariate split at branch node t

Formulation

Let $\alpha \ge 0$ be a complexity parameter in the objective to deter the model from using all branch nodes to split data. Our model, which we call S-OCT, is

$$\begin{aligned}
\underset{\mathbf{c},\mathbf{d},\mathbf{w},\mathbf{z},\mathbf{a},\mathbf{b}}{\text{minimize}} \quad & \frac{1}{N}\left(N - \sum_{i=1}^{N}\sum_{t\in\mathcal{L}} z_{it}\right) + \alpha \sum_{t\in\mathcal{B}} d_t && && \text{(1a)}\\
\text{subject to} \quad & \sum_{t\in\mathcal{L}} w_{it} = 1 && \forall i \in [N], && \text{(1b)}\\
& w_{it} = w_{i,2t} + w_{i,2t+1} && \forall i \in [N],\ t \in \mathcal{B}, && \text{(1c)}\\
& w_{i,2t+1} \le d_t && \forall i \in [N],\ t \in \mathcal{B}, && \text{(1d)}\\
& \sum_{k=1}^{K} c_{kt} = 1 && \forall t \in \mathcal{L}, && \text{(1e)}\\
& z_{it} \le w_{it} && \forall i \in [N],\ t \in \mathcal{L}, && \text{(1f)}\\
& z_{it} \le c_{y^i,t} && \forall i \in [N],\ t \in \mathcal{L}, && \text{(1g)}\\
& c_{kt} \in \{0,1\} && \forall k \in [K],\ t \in \mathcal{L}, && \text{(1h)}\\
& d_t \in \{0,1\} && \forall t \in \mathcal{B}, && \text{(1i)}\\
& w_{it} \in \{0,1\} && \forall i \in [N],\ t \in \mathcal{B} \cup \mathcal{L}, && \text{(1j)}\\
& z_{it} \in \{0,1\} && \forall i \in [N],\ t \in \mathcal{L}, && \text{(1k)}\\
& (\mathbf{a}_t, b_t) \in \mathcal{H}_t(\mathbf{w}) && \forall t \in \mathcal{B}, && \text{(1l)}
\end{aligned}$$

where, $\forall t \in \mathcal{B},\ \mathbf{w} \in \left\{\{0,1\}^{N \times (\mathcal{B} \cup \mathcal{L})} : \text{(1b)} - \text{(1d)}\right\}$,

$$\mathcal{H}_t(\mathbf{w}) = \{(\mathbf{a}_t, b_t) \in \mathbb{R}^p \times \mathbb{R} : \mathbf{a}_t^\top \mathbf{x}^i + 1 \le b_t \quad \forall\, i \in [N] : w_{i,2t} = 1, \quad \mathbf{a}_t^\top \mathbf{x}^i - 1 \ge b_t \quad \forall\, i \in [N] : w_{i,2t+1} = 1\}. \tag{2}$$

- Master problem (1a)-(1k) routes observations to leaves to minimize error rate plus regularization term
- LP feasibility subproblem enforces (1l) by checking existence of $(\mathbf{a}_t, b_t) \in \mathcal{H}_t(\mathbf{w})\ \forall t \in \mathcal{B}$, ensuring a multivariate decision tree can fulfill master problem's routing; if not, then must add feasibility cuts on the $\mathbf{w}$ variables

Shattering Inequalities

- Points $\{x^i\}$ can be *shattered* by a linear classifier if for any partition $\{x^i\} = X_1 \dot\cup X_2$ there exists a hyperplane separating X_1 and X_2
- Let $\mathcal{I} = \left\{I \subseteq [N] : \{x^i\}_{i\in I} \text{ cannot be shattered by linear classifiers}\right\}$. For $I \in \mathcal{I}$, let $\Lambda(I) \subset \{-1,1\}^I$ be assignments of binary labels so that points in I cannot be separated. The following *shattering inequalities* enforce (1l):
$$\sum_{i\in I:\lambda_i=-1} w_{i,2t} + \sum_{i\in I:\lambda_i=+1} w_{i,2t+1} \le |I| - 1 \quad \forall I \in \mathcal{I},\ \lambda \in \Lambda(I),\ t \in \mathcal{B}. \tag{3}$$
- Only need to consider *minimal* subsets $I' \in \mathcal{I}$ in (3). Suppose $\mathcal{H}_t(\mathbf{w}) = \varnothing$ for some integral $\mathbf{w},\ t \in \mathcal{B}$; let $I(t') = \{i \in [N] : w_{it'} = 1\}\ \forall t' \in \mathcal{B} \cup \mathcal{L}$:
$$\sum_{i\in I'\cap I(2t)} w_{i,2t} + \sum_{i\in I'\cap I(2t+1)} w_{i,2t+1} \le |I'| - 1. \tag{4}$$
I' indexes an *Irreducible Infeasible Subsystem (IIS)* of the constraints in (2), and can be found efficiently
- Shattering inequalities for minimal $I' \in \mathcal{I}$ always involve $\le p+2$ variables
- In Figure 3, let $x^i = (0,0), (0,1), (1,0), (1,1)$ be indexed by $i = 1,2,3,4$ resp. One shattering ineqaulity is $w_{1,2t} + w_{4,2t} + w_{2,2t+1} + w_{3,2t+1} \le 3$

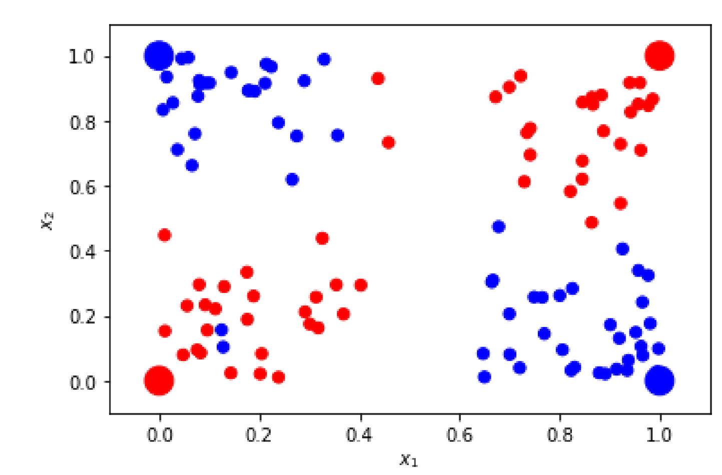

Fig. 3: Example of finding a shattering inequality in $\mathbb{R}^2$. The master problem proposes sending the red points to the left child and the blue points to the right child of some branch node t.

Experimental Results

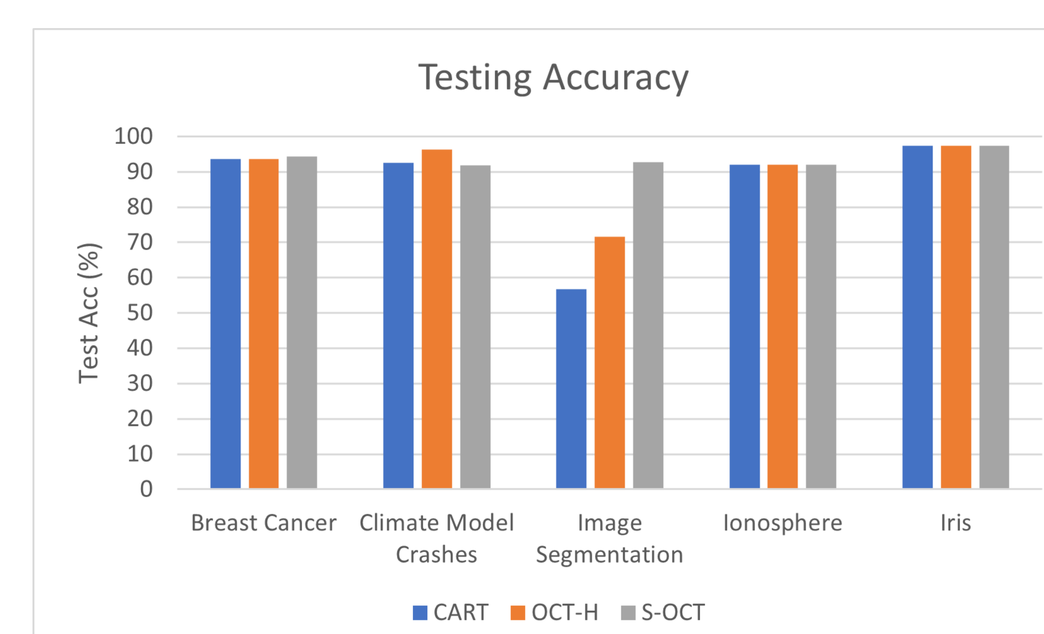

Fig. 4: Testing accuracy comparison between CART, OCT-H, and S-OCT on select datasets, $D = 3$.

We compare against CART and OCT-H across 10 different datasets and for max depths $D = 2, 3, 4$. We set a time limit of 10 minutes for all models. Overall, S-OCT reduces training time by 59.5% w.r.t. to OCT-H.