Abstract
Learning a decision tree from data is a difficult optimization problem. The most widespread algorithm in practice, dating to the 1980s, is based on a greedy growth of the tree structure by recursively splitting nodes, and possibly pruning back the final tree. The parameters (function definition and internal node structure) of an internal node are approximately estimated by minimizing an impurity measure. We give an algorithm that, given an input tree (its structure and parameters), produces a new tree with the same or smaller structure but new parameter values that provably lower or at least do not increase the impurity measure. This can be applied to both axis-aligned and oblique trees and our experiments show it consistently outperforms several existing algorithms while being scalable to large datasets and trees. Further, this algorithm can handle a sparsity penalty, so it can also prune oblique trees, which is a structure that is not a subset of the original tree and few parameters. This combines the benefits of both the oblique trees: flexibility to modeled correlated data, low generalization error, fast inference and interpretable nodes that involve only a few features in their decision.

Work supported by NSF award 1432515.

2 Decision trees
The prediction model for a decision tree is given by a path from the root to a leaf consisting of a sequence of decisions, each forming a question of the type \( \text{"if } x_i \text{ is smaller than threshold } \theta_i \text{, then } \) (1), for a feature \( x_i \) bigger than threshold \( \theta_i \) for axis-aligned trees, or \( \text{if } \lambda_i x_i \geq \theta_i \) for oblique trees. This tree makes decision trees preferable over more accurate models such as neural nets in some applications, such as medical diagnosis or age analysis.

• Among the most widely used statistical models in practice.
• Easy to model nonlinear data.
• Very transparent.
• May be used with all input features to make a prediction.
• Interpretable: the path can be understood as a sequence of IF-THEN rules, which is familiar to humans. We can equivalently turn the tree into a database of rules.

This can make decision trees preferable over more accurate models such as neural nets in some applications, such as medical diagnosis or age analysis.

Traditional optimization of decision trees
Learning the tree from data is a very difficult optimization problem, involving a search over a large and complex space of trees. The most widespread algorithm in practice, dating to the 1980s, is based on a greedy growth of the tree structure by recursively splitting nodes, and possibly pruning back the final tree. The parameters (function definition and internal node structure) of an internal node are approximately estimated by minimizing an impurity measure. We give an algorithm that, given an input tree (its structure and parameters), produces a new tree with the same or smaller structure but new parameter values that provably lower or at least do not increase the impurity measure. This can be applied to both axis-aligned and oblique trees and our experiments show it consistently outperforms several existing algorithms while being scalable to large datasets and trees. Further, this algorithm can handle a sparsity penalty, so it can also prune oblique trees, which is a structure that is not a subset of the original tree and few parameters. This combines the benefits of both the oblique trees: flexibility to modeled correlated data, low generalization error, fast inference and interpretable nodes that involve only a few features in their decision.

Work supported by NSF award 1432515.