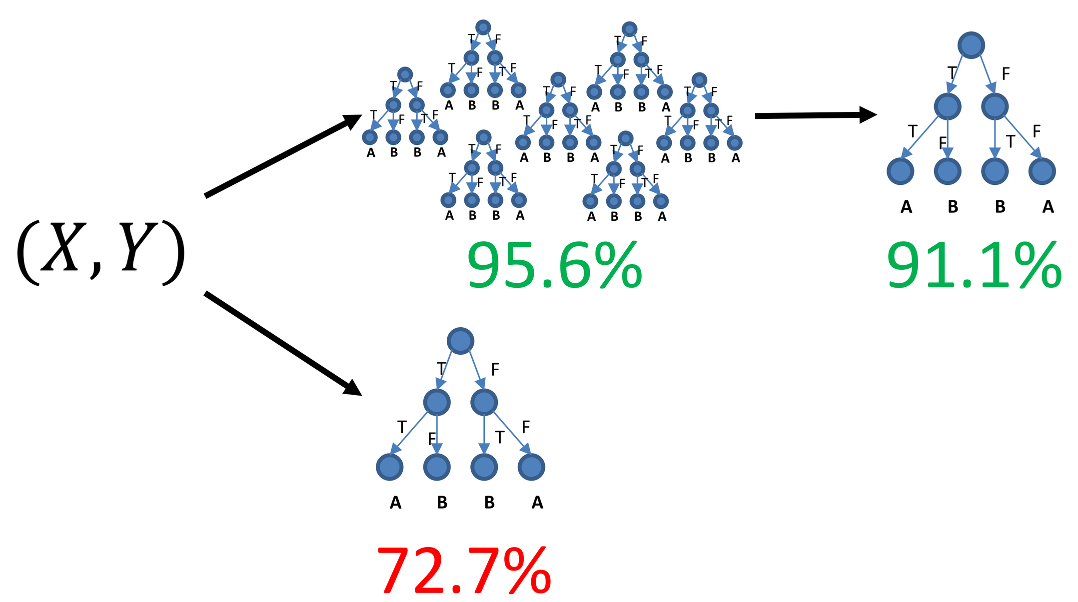
Interpreting Blackbox Models via Model Extraction

Osbert Bastani[1,4], Carolyn Kim[2], Hamsa Bastani[3,4]

[1]Massachusetts Institute of Technology, [2]Stanford University, [3]IBM Research, [4]University of Pennsylvania

Summary

- **Motivation**
  - Despite having high accuracy, blackbox machine learning models lack interpretability.
  - This is a concern when such models are used for consequential decisions, e.g., medical diagnosis.
- **Algorithm**
  - We propose interpreting blackbox models by extracting a decision tree that approximates the model.
  - We avoid overfitting by actively sampling new data points and labeling them using the model.



- **Related literature**
  - Directly learning interpretable models (Ustun-Rudin 2016)
  - Interpreting specific test points (Ribeiro et al., 2016)
  - Computing influence scores for features (Friedman 2001) or training points (Koh-Liang 2017)

Problem Formulation

- **Inputs**
  - Blackbox classifier $f: \mathcal{X} \to \mathcal{Y}$
  - Training set $(X, Y) \subseteq \mathcal{X} \times \mathcal{Y}$
  - Depth D of the decision tree to be extracted
- **Output**
  - An axis-aligned decision tree $T(X) \approx f(x)$
  - Use T to understand f

Exact Greedy Decision Tree

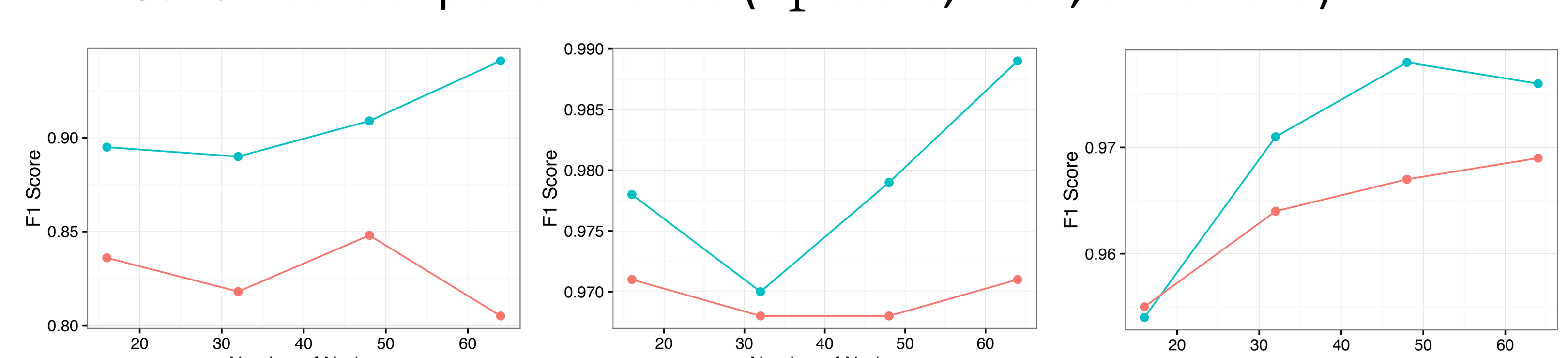
- **Estimate input distribution**
  - Fit a Gaussian mixture model P to X
  - Components of P are axis-aligned Gaussians
- **Iteratively construct tree**
  - **Initialization:** $T^* = \{N\}$ contains a single node
  - **Growth step:** Choose a leaf node N in T^*, and replace N with an internal node and two new leaf nodes
- **Single growth step**
  - For each node N, let $P_N = P \mid (x \text{ satisfies } C_N)$, i.e., P conditioned on x flowing to N in T^*
  - Choose N to be the node with highest gain (according to P_N) if replaced as described below
  - Choose an axis-aligned branch that maximizes the gain
  - Choose labels for new leaf nodes to be the majority labels

Estimated Greedy Decision Tree

- **Approximation**
  - Estimate gains above using m random samples $x \sim P_N$
  - To sample $x \sim P_N$, sample a component of P_N, and sample a point from that component (which is a truncated Gaussian)
  - Corresponding label is $y = f(x)$
- **Theorem:** As $m \to \infty$, the estimated tree converges to T^*

Comparison to CART

- **Datasets:** 6 UCI datasets and 3 classical control problems
- **Blackbox models:** random forest and neural net
- **Tree sizes:** ranging from 16 to 64 nodes
- **Metric:** test set performance (F_1 score, MSE, or reward)



Example Use Cases

- **Detect use of invalid features (e.g., response as a feature)**
  - We use a breast cancer dataset containing two response variables indicating recurrence. We trained a random forest where one response was incorrectly included as a feature for predicting the other. Then, we extract a decision tree.
  - The invalid feature occurred in every extracted tree, and as the top branch in 6 of the 10 trees.
- **Understand use of prejudiced features**
  - We use a student grade dataset where gender is a feature. We train a random forest to predict grade with gender as a feature, and extract decision trees.
  - Gender occurs at the fourth or fifth level in 7 of 10 trees.
  - Using the trees, we estimate that the gender variable has a large effect on 18.3% to 39.1% of students, with an effect size ranging from 0.44 to 0.77 grade points on this subgroup.
- **Comparing different models trained on the same dataset**
  - We train random forests and neural nets on a wine dataset.
  - Random forests achieved an F_1 score of at least 0.961, whereas neural nets were bimodal; 5 had F_1 score of at least 0.955, and the remaining had an F_1 score of at most 0.741.
  - In the extracted trees, the occurrence of the feature "chlorides" was highly correlated with poor performance.
- **Understanding a control policy**
  - The tree extracted from the Cartpole policy says to move the cart to the left exactly when

$$(\text{pole velocity} \le -0.286) \vee (\text{pole angle} \le -0.071)$$

  - In other words, move the cart to the left when the pole is already on the left, or when the pole is moving quickly towards the left.

References

Ustun &Rudin. Supersparse linear integer models for optimized medical scoring systems. Machine Learning, 2016.

Ribeiro, Singh, & Guestrin. Why should I trust you?: Explaining the predictions of any classifier. KDD, 2016.

Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001.

Koh & Liang. Understanding black-box predictions via influence functions. ICML, 2017