Interpreting Blackbox Models via Model Extraction

Osbert Bastani\textsuperscript{1,4}, Carolyn Kim\textsuperscript{2}, Hamsa Bastani\textsuperscript{3,4}
\textsuperscript{1}Massachusetts Institute of Technology, \textsuperscript{2}Stanford University, \textsuperscript{3}IBM Research, \textsuperscript{4}University of Pennsylvania

\section*{Summary}
\begin{itemize}
  \item **Motivation**
    \begin{itemize}
      \item Despite having high accuracy, blackbox machine learning models lack interpretability.
      \item This is a concern when such models are used for consequential decisions, e.g., medical diagnosis.
    \end{itemize}
  \item **Algorithm**
    \begin{itemize}
      \item We propose interpreting blackbox models by extracting a decision tree that approximates the model.
      \item We avoid overfitting by actively sampling new data points and labeling them using the model.
    \end{itemize}
  \item **Related literature**
    \begin{itemize}
      \item Directly learning interpretable models (Ustun-Rudin 2016)
      \item Interpreting specific test points (Ribeiro et al., 2016)
      \item Computing influence scores for features (Friedman 2001) or training points (Koh-Liang 2017)
    \end{itemize}
\end{itemize}

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

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

\section*{Estimated Greedy Decision Tree}
\begin{itemize}
  \item **Approximation**
    \begin{itemize}
      \item Estimate gains above using \( m \) random samples \( x \sim P_N \)
      \item To sample \( x \sim P_N \), sample a component of \( P_N \), and sample a point from that component (which is a truncated Gaussian)
      \item Corresponding label is \( y = f(x) \)
    \end{itemize}
  \item **Theorem:** As \( m \rightarrow \infty \), the estimated tree converges to \( T^* \)
\end{itemize}

\section*{Example Use Cases}
\begin{itemize}
  \item **Detect use of invalid features (e.g., response as a feature)**
    \begin{itemize}
      \item 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.
      \item The invalid feature occurred in every extracted tree, and as the top branch in 6 of the 10 trees.
    \end{itemize}
  \item **Understand use of prejudiced features**
    \begin{itemize}
      \item 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.
      \item Gender occurs at the fourth or fifth level in 7 of 10 trees.
      \item 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.
    \end{itemize}
  \item **Comparing different models trained on the same dataset**
    \begin{itemize}
      \item We train random forests and neural nets on a wine dataset.
      \item 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.
      \item In the extracted trees, the occurrence of the feature “chlorides” was highly correlated with poor performance.
    \end{itemize}
  \item **Understanding a control policy**
    \begin{itemize}
      \item The tree extracted from the Cartpole policy says to move the cart to the left exactly when
      \begin{align*}
        \text{(pole velocity} & \leq -0.286) \lor (\text{pole angle} \leq -0.071) \\
      \end{align*}
      \item 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.
    \end{itemize}
\end{itemize}

\section*{Comparison to CART}
\begin{itemize}
  \item **Datasets:** 6 UCI datasets and 3 classical control problems
  \item **Blackbox models:** random forest and neural net
  \item **Tree sizes:** ranging from 16 to 64 nodes
  \item **Metric:** test set performance (\( F_1 \) score, MSE, or reward)
\end{itemize}

\section*{References}

Koh & Liang. Understanding blackbox predictions via influence functions. ICML, 2017