Abstract

# Original

Interpreting state-of-the-art machine learning algorithms can be difficult. For example, why does a complex ensemble predict a particular class?

Existing approaches to interpretable machine learning tend to be either local in their explanations, apply only to a particular algorithm, or overly complex in their global explanations.

In this work, we propose a global model extraction method which uses multi-objective genetic programming to construct accurate, simplistic and model-agnostic representations of complex black-box estimators.

We found the resulting representations are far simpler than existing approaches while providing comparable reconstructive performance.

This is demonstrated on a range of datasets, by approximating the knowledge of complex black-box models such as 200 layer neural networks and ensembles of 500 trees, with a single tree.

# Condensed

In this work, we present a new approach to interpretable machine learning; using multi-objective genetic programming to construct simple representations of complex black-box Machine learning models.

Our results are far simpler with comparable reconstructive performance than that of existing state-of-the-art methods. Demonstrated on a range of datasets our resulting interpretable model (a single decision tree) can approximate the knowledge of 200 layer neural networks and ensembles of 500 trees.