ML : Interpretability

Sanjit Paliwal

Agenda

- Interpretability : Definition and Advantages
- Interpretable Models so far
- Optimal Decision Trees :
  - Intro
  - Four Flavors of Trees
  - Result Comparison
- CORELS
  - Stop Explaining Black Box Models by Prof Cynthia Rudin, Duke University
  - CORELS working
  - COMPAS vs CORELS

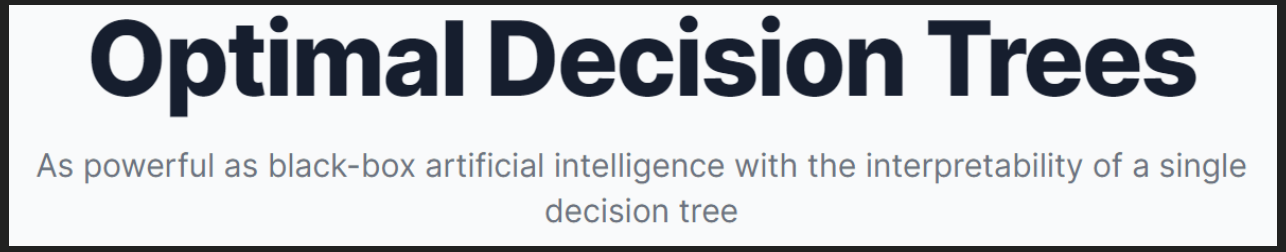
Interpretability

- **Interpretability is the degree to which a human can understand the cause of a decision.**[1]
- **ML Interpretability helps in checking following**[2] :
  - **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against underrepresented groups.
  - **Reliability or Robustness**: Ensuring that small changes in the input do not lead to large changes in the prediction.
  - **Causality**: Check that only causal relationships are picked up.
  - **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

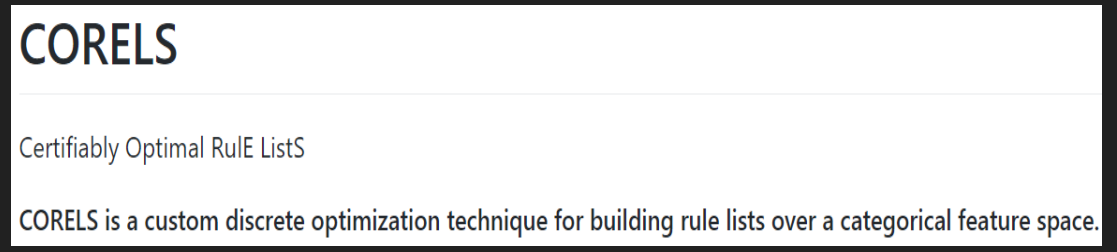
1 : Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).
2 : Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017)

Interpretable Models so far

- Some notable examples are Linear Regression, Logistic Regression and Decision Trees
- Interpretable models are not at par with some of the black box models like deep neural networks
- Most of the research focused on explaining these black box models
- But now some researchers/professors are trying to develop ML algorithms that are interpretable and provide state of the art performance



Source: https://www.interpretable.ai/products/optimal-trees/

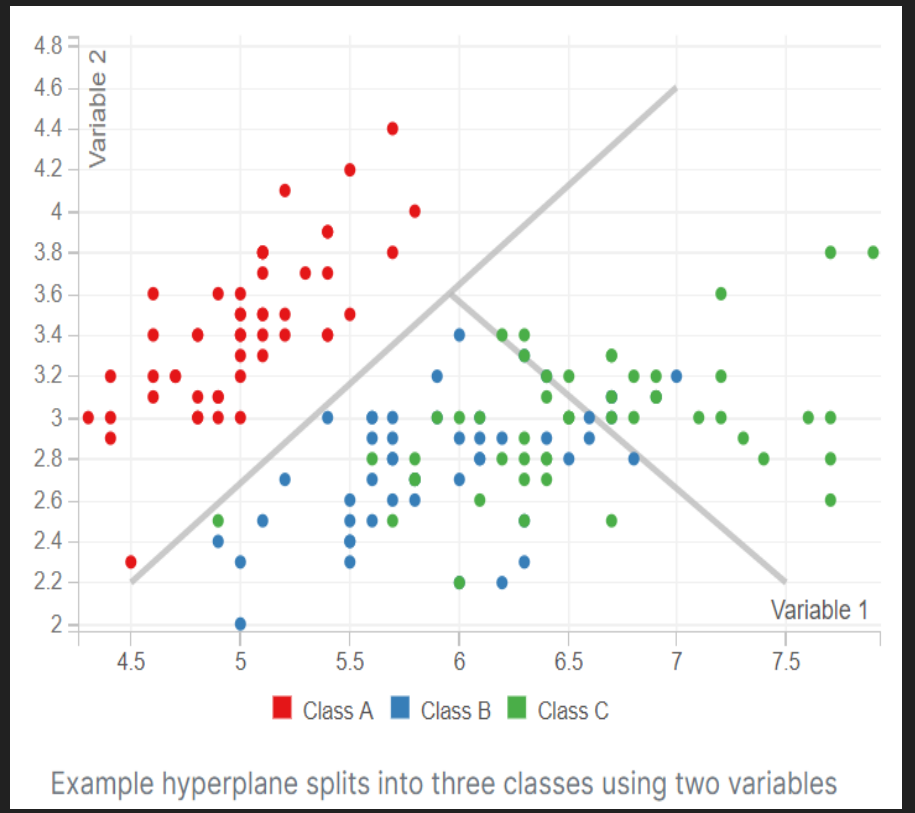


Source: https://corels.eecs.harvard.edu/corels/

Optimal Decision Trees

Optimal Decision Trees : Prof Bertsimas et al. (2017), MIT

- Limitation of CART and others[3]
  - Traditional Decision tree methods like CART, C4.5 and ID3 use greedy heuristics to form a tree one split at a time.
  - This leads to local optimal decision making but does not guarantee final tree to be optimal
- Optimal Decision Trees
  - Using Mixed Integer Optimization to construct the entire decision tree at once, finding the optimal tree that best fits the data
  - Unique to Optimal Decision Trees, hyperplane splits permit use of more than one feature at a time, enabling more expressive modeling and better performance

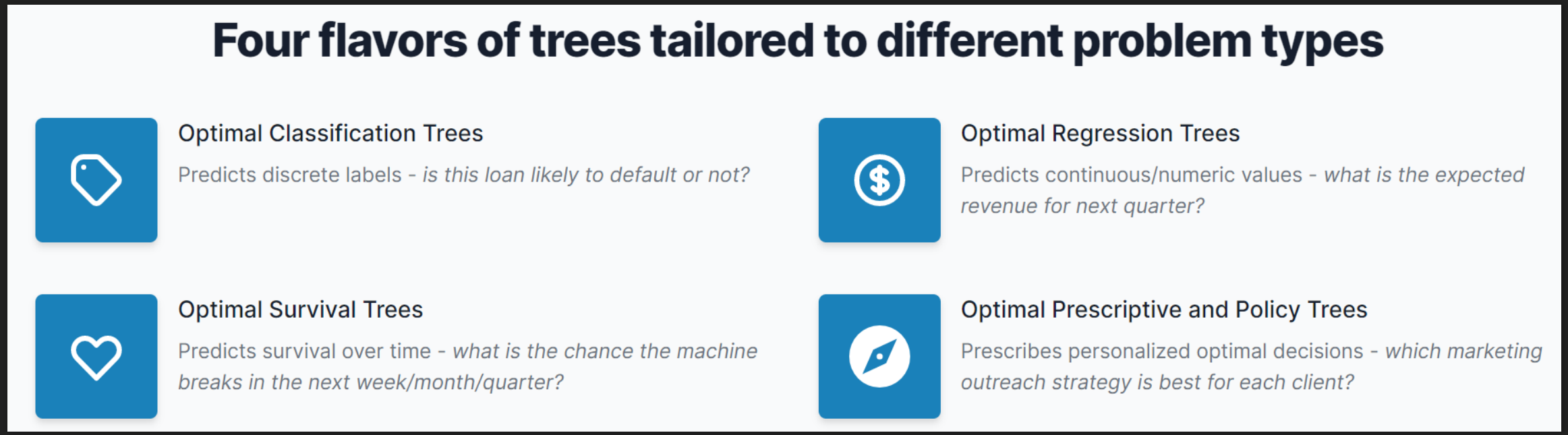


Example hyperplane splits into three classes using two variables

3 : Bertsimas, D., Dunn, J. Optimal Classification trees. Mach Learn 106, 1039–1082 (2017). https://doi.org/10.1007/s10994-017-5633-9

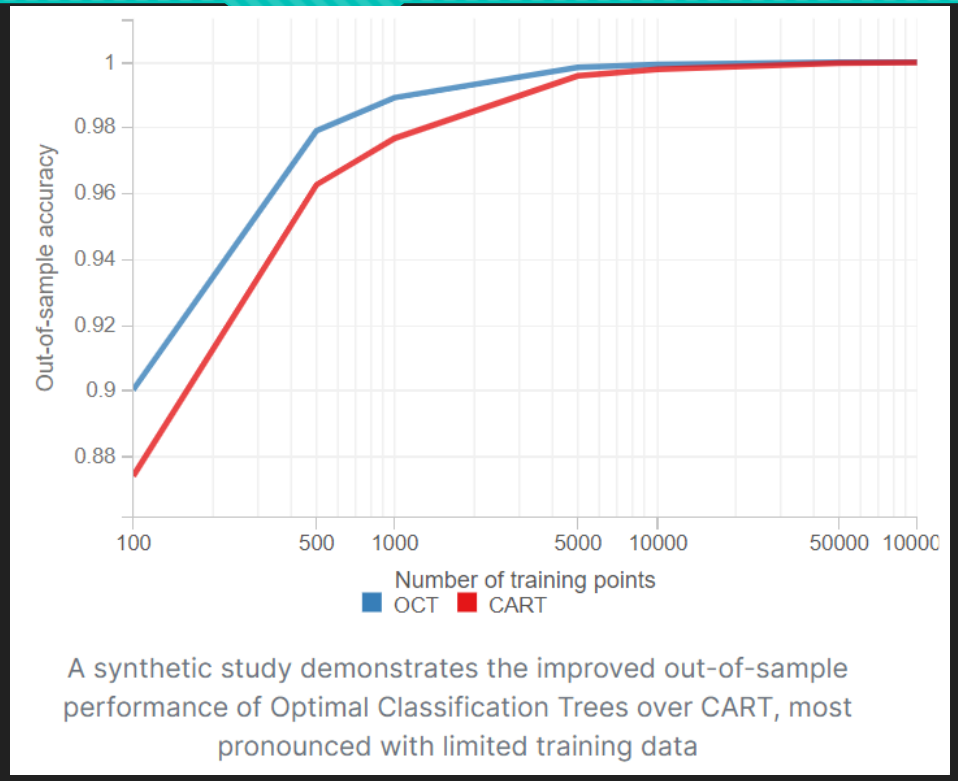
Source: https://www.interpretable.ai/products/optimal-trees/

Four Flavors of trees

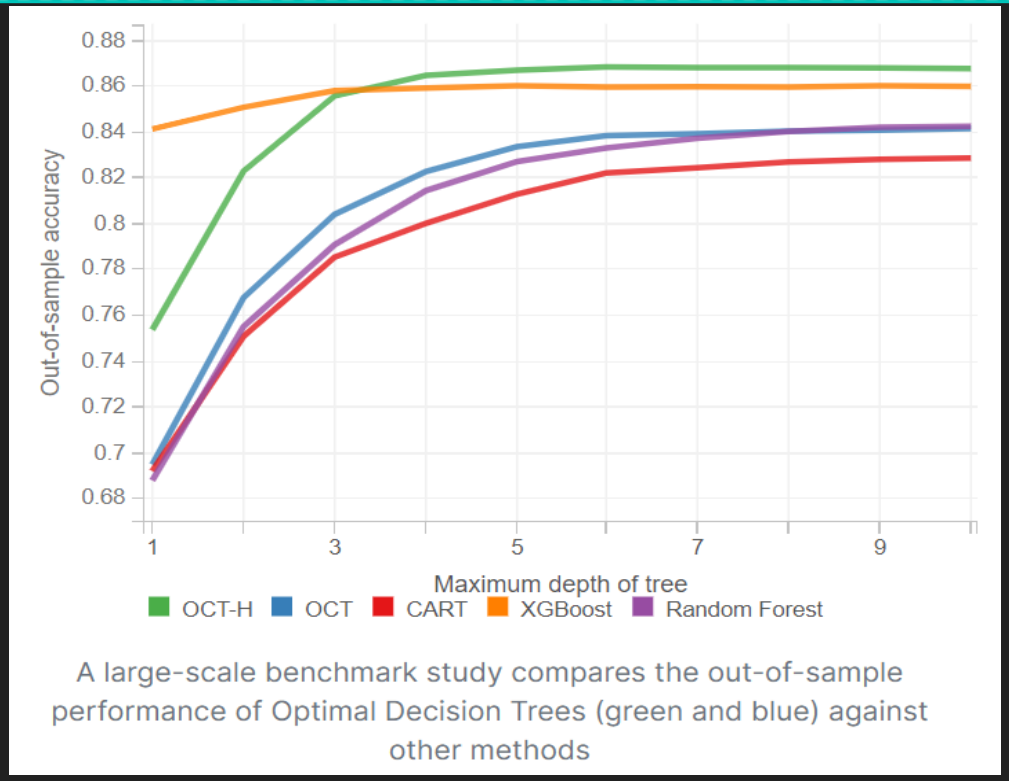


Source: https://www.interpretable.ai/products/optimal-trees/

Results Comparison



A synthetic study demonstrates the improved out-of-sample performance of Optimal Classification Trees over CART, most pronounced with limited training data



A large-scale benchmark study compares the out-of-sample performance of Optimal Decision Trees (green and blue) against other methods

Source: https://www.interpretable.ai/products/optimal-trees/

Compared to existing decision tree methods, Optimal Decision Trees deliver both higher predictive power and smaller trees, by making efficient use of data.

Optimal Decision Trees (OCT and OCT-H) deliver performance comparable to black-box methods while maintaining the interpretability of a single decision tree.

CORELS

Stop Explaining Black Box Models: Prof. Cynthia Rudin, Duke University (2019)

- A black box model could be either [5]
  - a function that is too complicated for any human to comprehend
  - or a function that is proprietary
- It is a myth that there is necessarily a trade-off between accuracy and interpretability
  - When considering problems that have structured data with meaningful features, there is often no significant difference in performance.
- Explanations often do not make sense

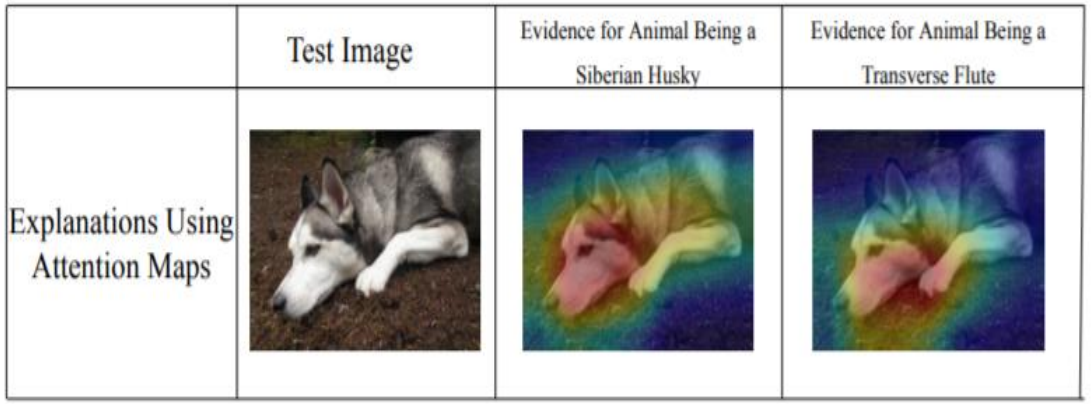


Figure 2: Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Figure credit: Chaofan Chen and [28].

5: Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 206–215 (2019).

CORELS : Certifiable Optimal RulE ListS

- Rule List/Decision List
  - A rule is simply an if-then statement.
  - A rule list is simply a group of rules in a particular order, followed by a default rule. Rule/Decision List is a one-sided decision tree.
- CORELS
  - Discrete optimization technique for building rule lists over a categorical feature space[7].
  - By leveraging algorithmic bounds, efficient data structures, and computational reuse, it achieves several orders of magnitude speedup in time and a massive reduction of memory consumption.

```
If age > 25 then predict likes sports = false

Else If Lives in Eastern US = true then predict likes sports = false

Else If wears glasses = true then predict likes sports = false

Else predict likes sports = true
```

Source: https://corels.eecs.harvard.edu/corels/whatarerulelists.html

7: Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning Certifiably Optimal Rule Lists for Categorical Data. JMLR, 2018

COMPAS vs CORELS

COMPAS proprietary recidivism risk prediction tool that is in widespread use in the U.S. Justice System for predicting the probability that someone will be arrested after their release[5]

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Figure 3: This is a machine learning model from the Certifiably Optimal Rule Lists (CORELS) algorithm [32]. This model is the minimizer of a special case of Equation 1 discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	full model is in Figure 3 only age, priors, (optional) gender no other information free, transparent

Table 1: Comparison of COMPAS and CORELS models. Both models have similar true and false positive rates and true and false negative rates on data from Broward County, Florida.

References

1. Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).
2. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017)
3. Bertsimas, D., Dunn, J. Optimal Classification trees. Mach Learn 106, 1039–1082 (2017). https://doi.org/10.1007/s10994-017-5633-9
4. Interpretable AI by Dimitris Bertsimas, MIT plus Opening of OR62 Conference. https://www.youtube.com/watch?v=gAZ4YRngEj0
5. Interpretable AI by Dimitris Bertsimas, MIT plus Opening of OR62 Conference. https://www.youtube.com/watch?v=gAZ4YRngEj0
6. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019).
7. The Problem with Black Boxes with Cynthia Rudin - TWIML Talk #290: https://www.youtube.com/watch?v=n_mwYWfI_sI
8. Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning Certifiably Optimal Rule Lists for Categorical Data. JMLR, 2018
9. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable by Christoph Molnar


Q
&
A


THANK
YOU


That's all Folks!