IIT Bombay

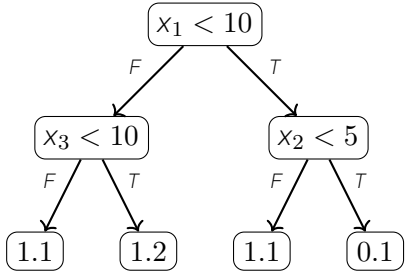
Sensitivity Analysis for Decision Tree Ensembles

Arhaan Ahmad, Tanay Tayal, Ashutosh Gupta, S. Akshay

Department of Computer Sciences and Engineering, Indian Institute of Technology Bombay, Mumbai, India

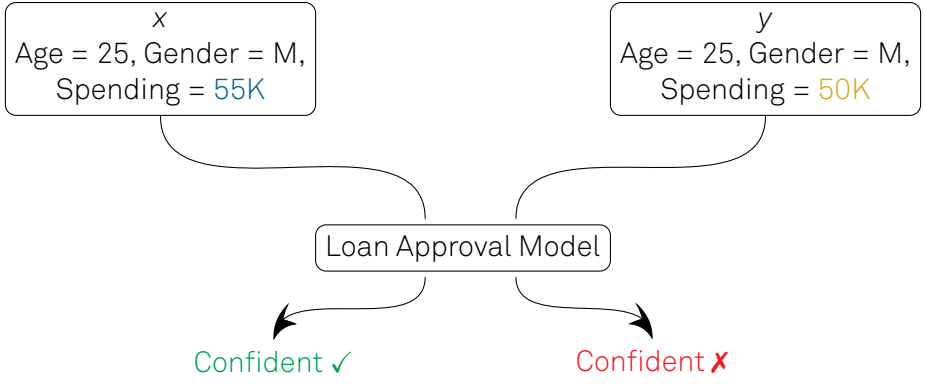
Decision Tree Ensembles

- Tree ensemble models, such as GBDTs and random forests, are widely popular models for a variety of machine learning tasks.
- Given an input $x = (x_1, x_2, x_3)$, a decision tree looks as follows:



- Multiple trees are aggregated by adding their outputs.
  - $c_{raw}(x) = \Sigma_{i=1}^{n} T_i(x)$
  - $c_{label}(x) = \sigma(c_{raw}(x))$
- If $c_{label}(x) > 0.5$ then it is classified as the positive class.

The Sensitivity Problem



Sensitivity Verification Problem

Defn. (p, F)-sensitivity

- Given binary classifier model $c : \mathbb{R}^n \to \{0, 1\}$, a confidence threshold $0 < p < 0.5$ and a subset of features F, do there exist inputs x and $y \in \mathbb{R}^n$ such that :
  - x and y have the same values for features not in F
  - $c_{label}(x) \geq 0.5 + p$
  - $c_{label}(y) \leq 0.5 - p$

Why this matters?

Sensitivity verification is crucial to building **trustworthy AI** systems.

- **Fairness**: Sensitive features (e.g., gender, race) should not unfairly influence decisions.
- **Interpretability**: Understanding which features drive decisions helps build trust in AI systems.
- **Security**: Manipulation of a small sensitive subset of features may enable adversarial attacks.

Contributions

1. We prove that the Sensitivity Verification problem is NP-Hard
   - The Sensitivity Verification problem, that is, checking whether a given tree ensemble classifier is sensitive to a feature set F, is NP-Hard for decision tree ensembles with trees of depth ≥ 3
2. We create an encoding for approximate representations of tree ensembles using Pseudo-Boolean constraints
3. We develop a tool, SensPB, to verify the sensitivity of tree ensembles, beating previous methods by a big margin

Pseudo-Boolean Constraints

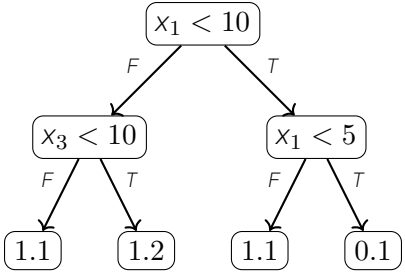
Inequalities involving Boolean variables.

$$\sum a_i x_i \geq A$$

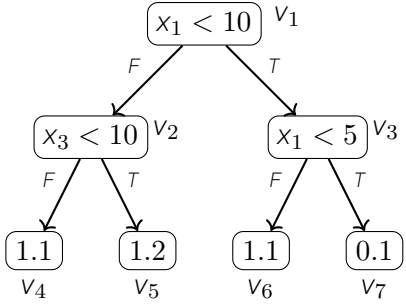
$$x_i \in \{0, 1\}$$

- State of the art solvers are relatively new (5 years or so)
- Stronger than SAT
- Natural for encoding tree ensembles
- Solvers require integer weights

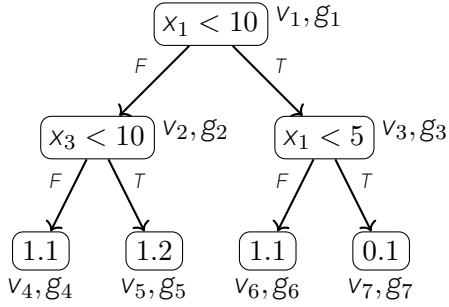
Encoding



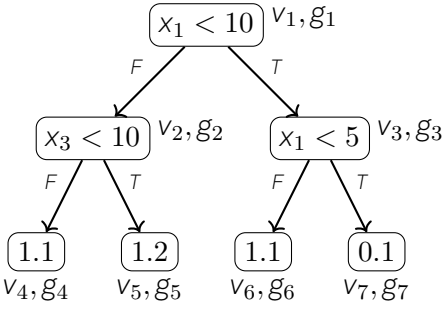
Encoding



Encoding

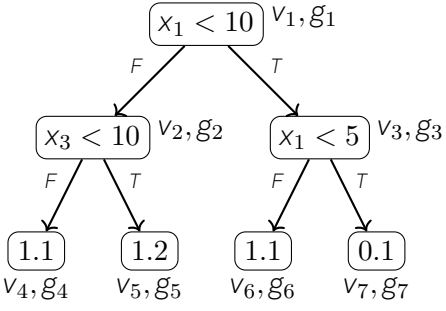

$x_1 < 10$
v_1, g_1
F
T
$x_3 < 10$
v_2, g_2
$x_1 < 5$
v_3, g_3
F
T
F
T
1.1
1.2
1.1
0.1
v_4, g_4
v_5, g_5
v_6, g_6
v_7, g_7

Encode Tree



- $v_1 \wedge g_1 \implies v_3$
- $v_1 \wedge \neg g_1 \implies v_2$
- $v_3 \wedge g_3 \implies v_7$
- ...
- $g_3 \implies g_1$

Encode Classification



SMT Encoding:

$$1.1v_4 + 1.2v_5 + 1.1v_6 + 0.1v_7 \leq -\delta$$

$$1.1v_4 + 1.2v_5 + 1.1v_6 + 0.1v_7 \geq \delta$$

SensPB Encoding: Pick a precision α

$$\lceil \alpha \cdot 1.1 \rceil v_4 + \lceil \alpha \cdot 1.2 \rceil v_5 + \lceil \alpha \cdot 1.1 \rceil v_6 + \lceil \alpha \cdot 0.1 \rceil v_7 \geq \lfloor \alpha \cdot \delta \rfloor$$

$$\lfloor \alpha \cdot 1.1 \rfloor v_4 + \lfloor \alpha \cdot 1.2 \rfloor v_5 + \lfloor \alpha \cdot 1.1 \rfloor v_6 + \lfloor \alpha \cdot 0.1 \rfloor v_7 \leq \lceil -\alpha \cdot \delta \rceil$$

where $\delta = \sigma^{-1}(0.5 + p)$.

Existing Work

- Veritas: Construct a graph using the leaf values and find max-cliques using approximate algorithms
- SMT-based approach: Encode the ensemble tree into an SMT formula and pass it to an SMT Solver

Results

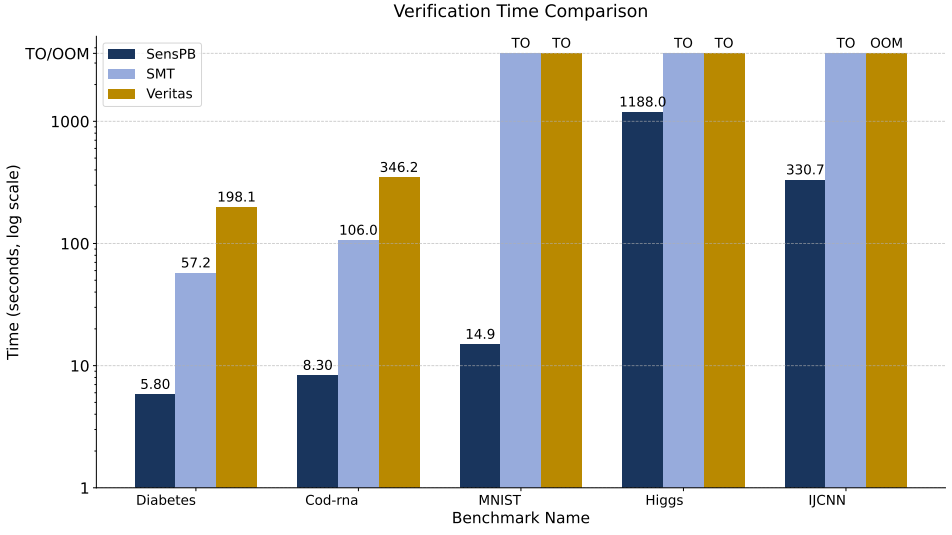


Figure: Time for running the three tools, on a log scale. TO = Time out, set to 3600 seconds. OOM = Out Of Memory

Benchmarks

Table: Tree Ensemble Sizes

Benchmark	#Trees	Depth	#Features
Diabetes	20	5	9
Cod-rna	80	4	8
Binary MNIST	50	6	784
Higgs	100	8	28
IJCNN	60	8	23

Future Work

- Quantifying Sensitivity of models
- Training Insensitive Models
- Further Scalability

Paper and Code

- Paper: https://openreview.net/pdf?id=h0vC0fm1q7
- Code: https://github.com/Arhaan/SensPB

Thanks for attending!

Acknowledgments

This work was supported by the SBI Foundation Hub for Data Science & Analytics, IIT Bombay.