Robust Optimal Classification Trees against Adversarial Examples

Daniël Vos and Sicco Verwer

Delft University of Technology, Delft, The Netherlands

Decision trees are a popular choice of explainable model, but just like neural networks, they suffer from adversarial examples. Existing algorithms for fitting decision trees robust against adversarial examples are greedy heuristics and lack approximation guarantees. For example the robust heuristics TREANT [2] and GROOT [3] fail on XOR shaped data (see Figure 1) because they cannot identify a single good split. In this work, we aim to train decision trees with maximal training accuracy against adversarial attacks by taking inspiration from the field of optimal decision trees [1].

We propose ROCT, a collection of methods to train decision trees that are optimally robust against user-specified attack models. We show that the min-max optimization problem that arises in adversarial learning can be solved using a single minimization formulation for decision trees with 0-1 loss. We propose such formulations in Mixed-Integer Linear Programming and Maximum Satisfiability, which widely available solvers can optimize. To improve the efficiency of ROCT we can warm start the solver with a heuristic decision tree such as one produced by GROOT. This speeds up the process of finding the optimal tree and proving its optimality.

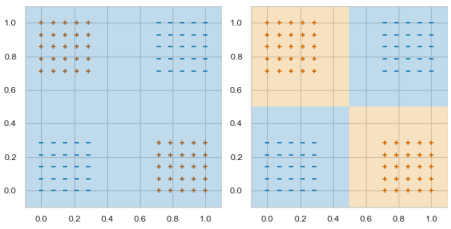

Fig. 1: Existing methods (left) greedily optimize one split at a time and cannot find a good tree to fit the XOR-shaped data. ROCT (right) optimizes the entire tree at once and finds the optimal model.

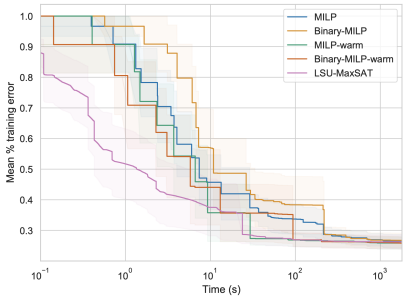

Fig. 2: Mean accuracy against adversarial attacks on 8 datasets when using ROCT with different solvers. The solvers can be stopped at any time and improve robustness until they have proven optimality.

We also present a method that uses bipartite matching to determine any model's upper bound on adversarial accuracy. This upper bound can be computed by leveraging the fact that if any two samples with different labels can be perturbed to an identical point in space, then these two samples cannot be

both correctly predicted against adversarial attacks. This fact can also generate redundant constraints to ROCT to improve solve time.

We compared the performance of regular decision trees, GROOT, TREANT and ROCT on 8 datasets with 3 different perturbation sizes each. Our results against optimal adversarial attacks in Table 1 show that using ROCT with Mixed-Integer Linear Programming warm starts clearly outperforms the existing heuristics. Surprisingly, the Maximum Satisfiability solver LSU also improved on existing techniques without requiring warm starts. Since ROCT can prove optimality given enough run time we have also compared how close GROOT and TREANT actually get to optimal results. We empirically demonstrate that the heuristics often score within 95% of the optimal score.

Algorithm	Mean adv. accuracy	Mean rank	Wins
Decision Tree	.388 ± .055	8.917 ± .083	0
TREANT	.692 ± .013	5.167 ± .604	7
Bin.-MILP	.714 ± .013	3.958 ± .576	10
MILP	.720 ± .015	2.917 ± .454	12
RC2-MaxSAT	.724 ± .014	2.667 ± .393	10
GROOT	.726 ± .015	2.375 ± .450	16
Bin.-MILP-w.	.726 ± .015	2.083 ± .399	16
LSU-MaxSAT	.729 ± .014	2.125 ± .303	13
MILP-warm	**.735** ± .015	**1.583** ± .225	**17**

Table 1: Aggregated test scores over eight datasets, means are shown with standard error. All methods were trained for 30 minutes and selected their depth using 3-fold cross-validation, methods using ROCT are highlighted.

To conclude, the existing heuristic GROOT performs close to optimally but its performance can be improved by optimizing the entire decision tree at once. We hope that future work will extend ROCT's formulation to improve solve time for larger datasets and decision trees. ROCT can also be used to understand the limits of adversarial robustness of decision trees as the method computes an upper bound on adversarial accuracy. Like existing work, ROCT has assumed that attackers perturb samples within a user-defined norm, however, this assumption can be unrealistic. Further work has to be done on how to train robust models against more realistic notions of adversarial attacks.

References

1. Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.
2. Stefano Calzavara, Claudio Lucchese, Gabriele Tolomei, Seyum Assefa Abebe, and Salvatore Orlando. Treant: training evasion-aware decision trees. *Data Mining and Knowledge Discovery*, 34(5):1390–1420, 2020.
3. Daniël Vos and Sicco Verwer. Efficient training of robust decision trees against adversarial examples. In *International Conference on Machine Learning*, pages 10586–10595. PMLR, 2021.