Resilience Verification of Tree-Based Classifiers

Stefano Calzavara, **Lorenzo Cazzaro**, Claudio Lucchese,
Federico Marcuzzi, Salvatore Orlando

Ca' Foscari University, Venice, Italy

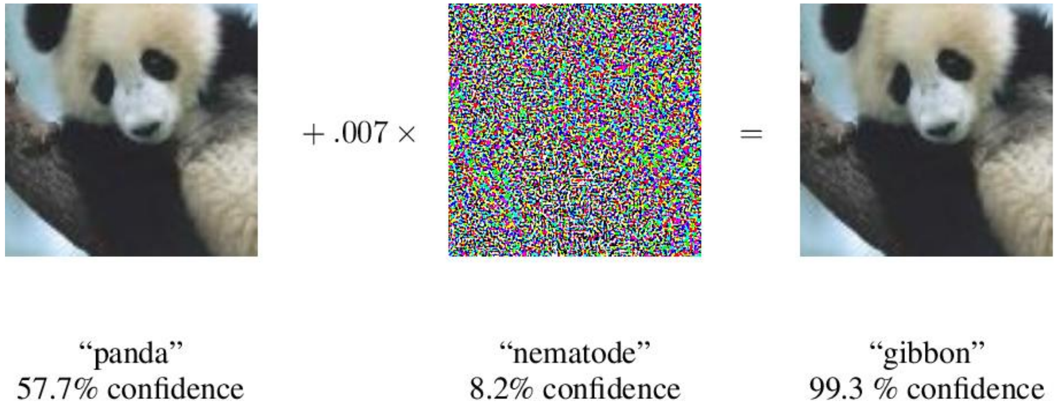
Università
Ca'Foscari
Venezia

Security of Classifiers

Machine Learning (ML) classifiers are vulnerable in adversarial scenarios ⟶ performance downgrade.

We focus on **evasion attacks**:

- (Imperceptible) Malicious manipulations of instances at test time.
- Objective: misprediction.
- Example: slight alteration of the pixels of an image.



Credits: *Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In ICLR. OpenReview.net*

Stability and Robustness

Consider:

- the classifier $g : \mathcal{X} \rightarrow \mathcal{Y}$.
- $A(\vec{x})$: the set of all the adversarial manipulations of the instance $\vec{x}$.

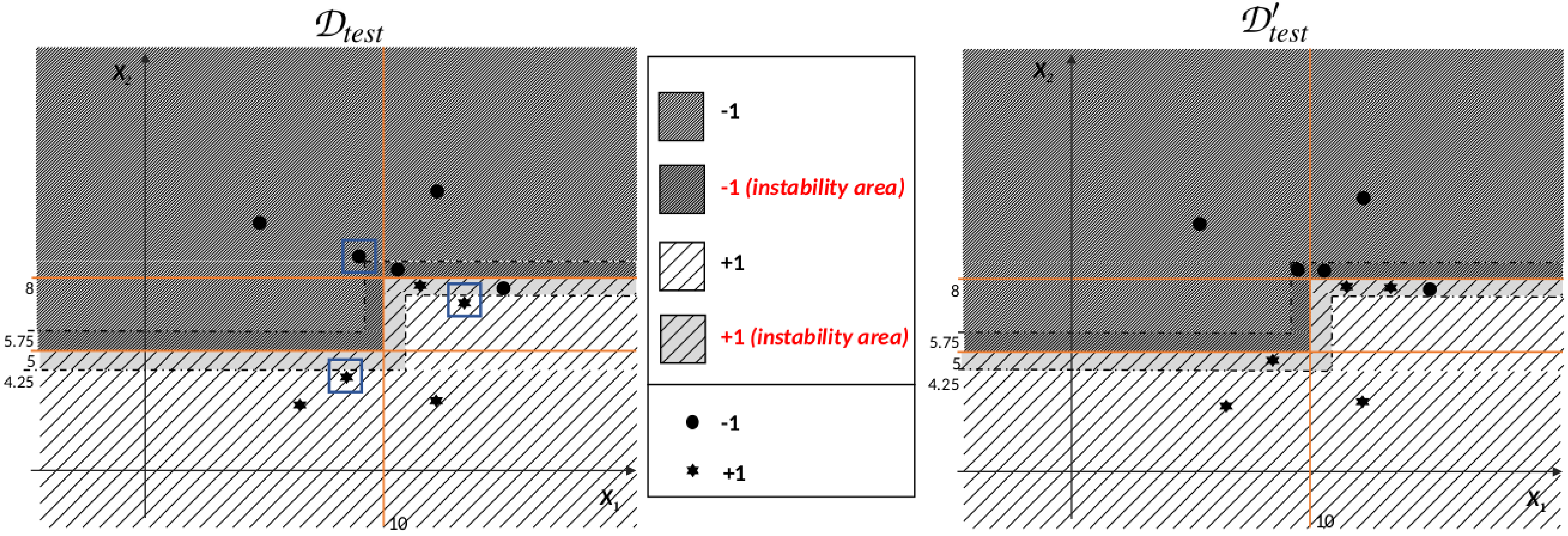
How to reason about the security of a classifier?

- **Stability**: the classifier g is **stable** on the instance $\vec{x}$ if and only if, for every adversarial manipulation $\vec{z} \in A(\vec{x})$, we have $g(\vec{x}) = g(\vec{z})$.

- **Robustness**: the classifier g is **robust** on the instance $\vec{x}$ if and only if $\vec{x}$ is correctly classified by g and g is stable on $\vec{x}$.

Shortcomings of Robustness

A key problem of robustness is its ***data-dependence***.
Tiny difference between two test sets ⟶ *quite different values* of robustness!



Accuracy = 9/10 = 0.9

Robustness = 7/10 = 0.7

Accuracy = 9/10 = 0.9

Robustness = 4/10 = 0.4

Contributions

1. Generalization of robustness beyond the test set: **resilience.**

2. How to verify resilience?
   - Robustness verification method + data-independent stability analysis (DISA) → DISA algorithm for decision trees and ensembles.

3. Experimental evaluation to motivate resilience and show the effectiveness of the proposed DISA.

Full paper* available on Arxiv.

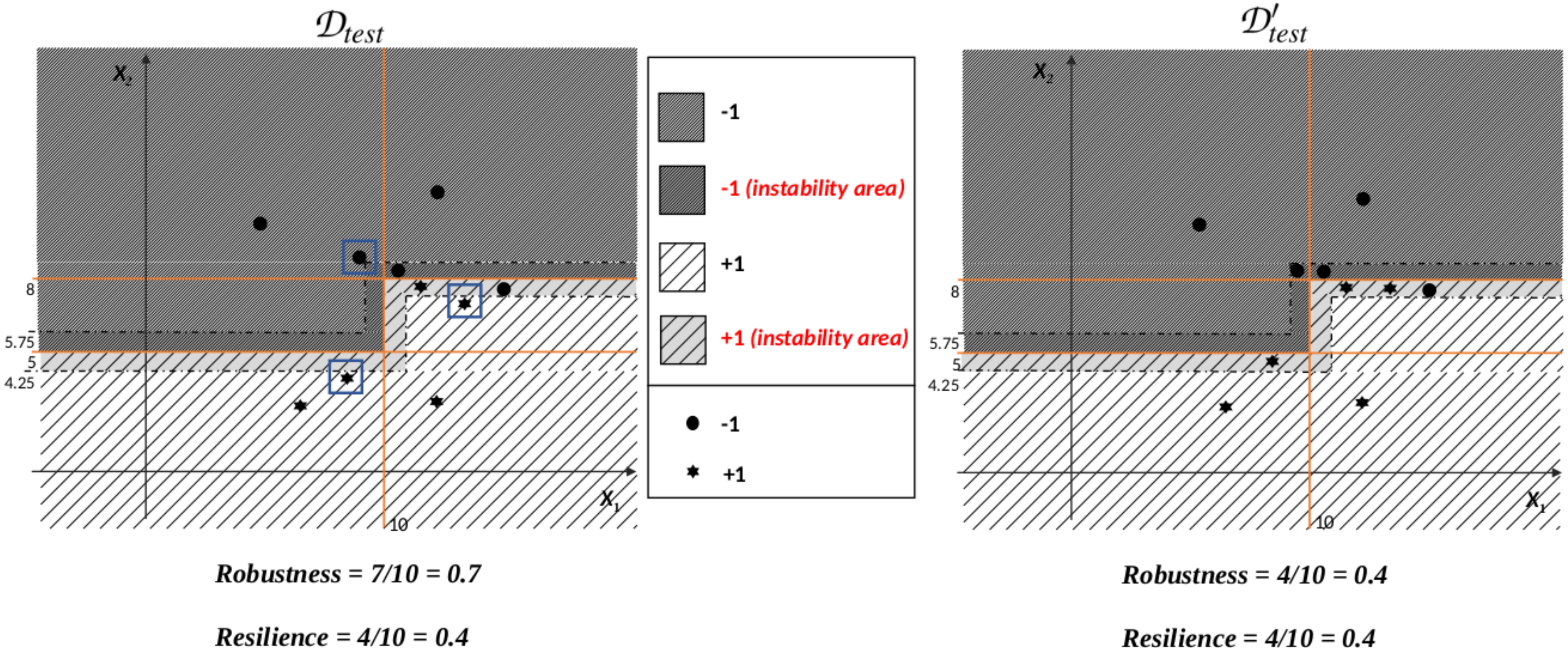
*https://arxiv.org/abs/2112.02705

Resilience

Resilience

$N(\vec{x})$ is the set of neighbours of $\vec{x}$, instances that could have been sampled in place of $\vec{x}$ $\longrightarrow$ it helps to generalize robustness beyond the test-set.

Resilience: a classifier g is **resilient** on the instance $\vec{x}$ if and only if g is robust on $\vec{x}$ and g is stable on all the instances $\vec{z} \in N(\vec{x})$.



Resilience Verification

Combine:

- Existing robustness verification methods.
- Data-independent stability analysis (DISA), that returns $X_s = \{\vec{x} \in \mathcal{X} \mid g \text{ is stable on } \vec{x}\}$.

Is g resilient on the instance $\vec{x}$?

1. Use DISA to obtain X_s (not trivial!).
2. Is g robust on $\vec{x}$ (use existiting methods or X_s)? If yes, go to step 3, otherwise g is not resilient on the instance.
3. $N(\vec{x}) \subseteq X_s$? If yes, g is resilient on $\vec{x}$, otherwise not.

Data-Independent Stability Analysis

Stability Analysis for Decision Trees/Forests

We designed a DISA algorithm for decision trees and forests. It's based on three steps:

1. **Annotate Leaf**
2. **Analyze Tree (proved sound)**
3. Analyze Ensemble (proved sound)

We provide an example of the analysis. See the full paper for the formalization of the three steps.

DISA - Annotate Leaf – Symbolic attack

Each node of the decision tree is annotated by a *symbolic attack* (SA) $\longrightarrow$ set of instances that can reach the node along with their *relevant* adversarial manipulations.

$$\underbrace{\langle(-\infty, 10], (4, 5]\rangle}_{\text{pre image}} \rhd \underbrace{\langle(-\infty, 10], (5, 6]\rangle}_{\text{post image}}{}_{\underset{\text{cost}}{\uparrow}1}$$

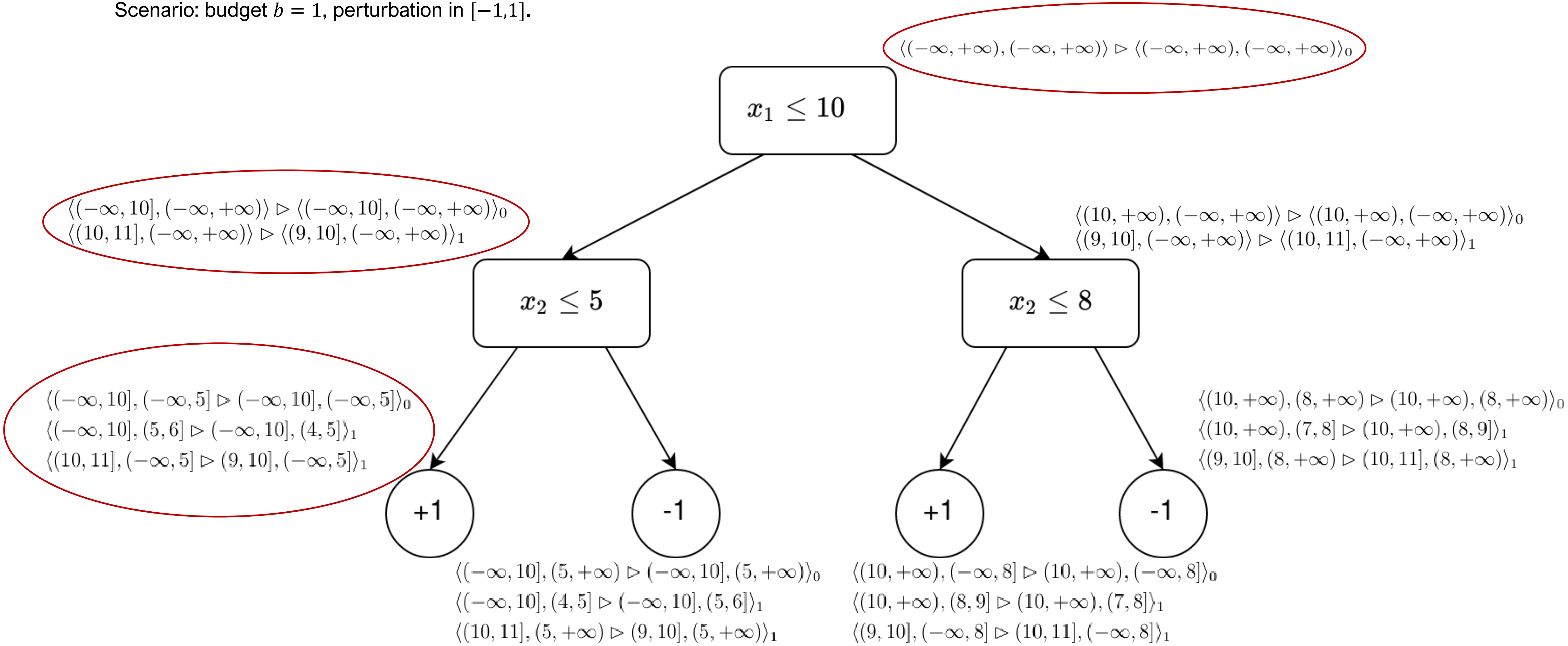
Components:

- Pre image: values before attack.
- Post image: values after attack.
- Cost: budget paid by the attacker.

Pre and post image are *hyperrectangles* (with as many intervals as the number of features).

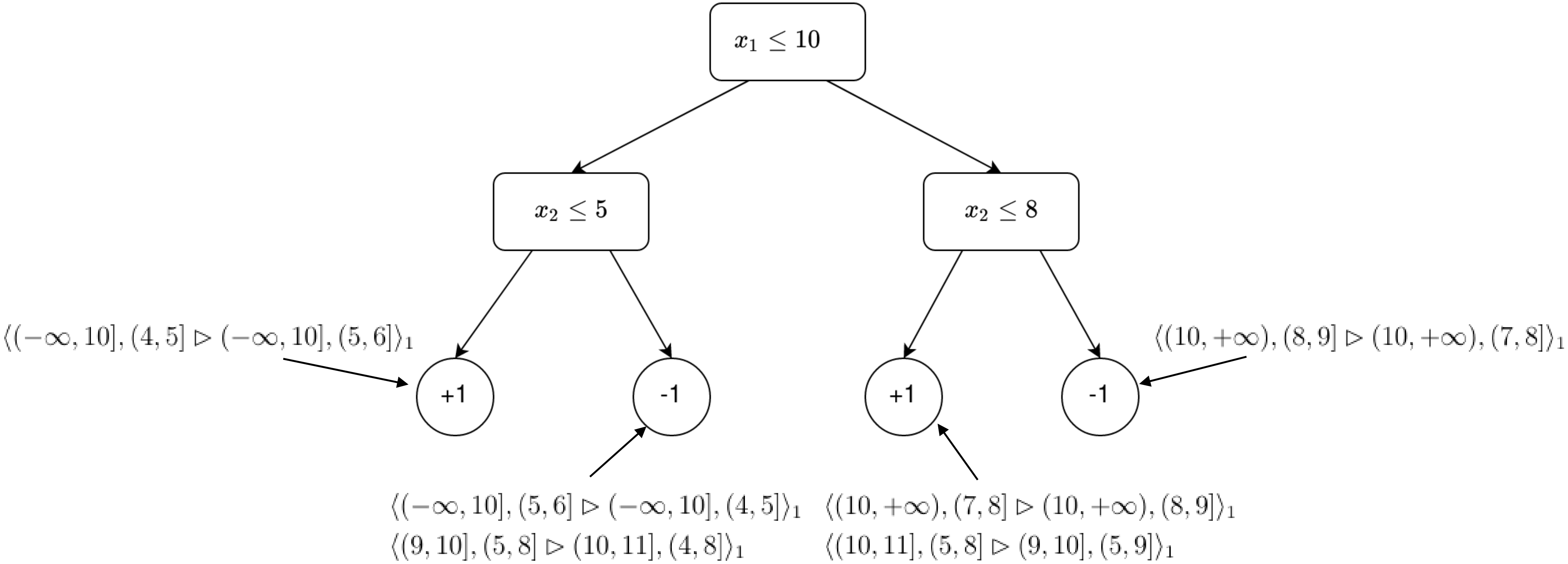
DISA - Annotate Leaf - Example

Scenario: budget $b = 1$, perturbation in $[-1,1]$.



DISA – Analyze Tree

Analyze Tree computes for each leaf the set of unstable SAs $U \longrightarrow$ SAs for which the attacker can force the decision tree to change its prediction.



Experimental Evaluation

Setup

Datasets: Breast Cancer, Cod-RNA, Diabetes (also new experiments with Sensorless).

ML models: standard and robust (TREANT*) decision trees and forests.

Attack scenario:

- Budget $b = 1$.
- The neighbourhood is $N(\vec{x}) = \{\vec{z} \in \mathcal{X} \mid \|\vec{z} - \vec{x}\|_\infty \leq \varepsilon\}$
- γ specifies the perturbation of the adversarial attacks.

Metrics: we use the test-set to compute the accuracy a, robustness r, its under-approximation $\hat{r}$ (using the result of the DISA), the under approximation of the resilience $\widehat{R}$ (using the result of the DISA).

* Stefano Calzavara, Claudio Lucchese, Gabriele Tolomei, Seyum Assefa Abebe, and Salvatore Orlando. Treant: training evasion-aware decision trees. Data Min. Knowl. Discov., 34(5):1390–1420, 2020.

Effectiveness of Resilience Verification - 1

Goals:

- Show that our estimate $\hat{R}$ is an **accurate under-approximation** of the actual resilience R.
- Show that **resilience significantly mitigates the shortcomings of robustness**.

Two experiments:

1. Use the similarity between r and $\hat{r}$ as a proxy of the precision of the stability analysis.
2. Compute $\bar{r}$, the robustness on the "most unlucky" sampling in the neighborhood of the original test set. If $\bar{r}$ is close $\hat{R}$, then most instances on which the classifier is not considered resilient by our analysis are indeed insecure.

Effectiveness of Resilience Verification - 2

Results:

- $\hat{r}$ **is a rather precise under-approximation** of the actual robustness $r \longrightarrow \hat{R}$ is a reasonably accurate estimate of R.
- **The gap between** r **and** $\hat{R}$ **may be quite significant** $\longrightarrow R$ provides a much more realistic security assessment than r.

				Standard Models					Robust Models				
Dataset	ε	# Trees	Depth	a	r	$\hat{r}$	$\overline{r}$	$\hat{R}$	a	r	$\hat{r}$	$\overline{r}$	$\hat{R}$
diabetes	0.01	5	3	0.708	0.662	0.643	0.656	0.636	0.727	**0.714**	0.701	**0.675**	**0.662**
		7	3	0.714	0.649	0.630	0.636	0.623	0.727	**0.714**	0.708	**0.675**	**0.662**
		9	3	0.747	0.656	0.630	0.623	0.617	0.753	**0.740**	0.727	**0.695**	**0.688**
cod-rna	0.01	5	3	0.775	**0.686**	0.672	**0.639**	**0.621**	0.752	0.715	0.707	0.698	0.691
		7	3	0.775	**0.686**	0.666	**0.640**	**0.612**	0.750	0.714	0.713	0.698	0.697
		9	3	0.769	**0.677**	0.663	**0.625**	**0.605**	0.750	0.714	0.713	0.698	0.697

Conclusion

Conclusion

1. Experimental results show that **robustness may give a false sense of security**.

2. **Resilience is useful in practice**, since it gives a more conservative account of the security of classifiers.

3. **Our data-independent stability analysis is precise and feasible.**

See the full paper for the formalization of the algorithms, the soundess theorems and proofs and additional experiments about scalability.