Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210109

Automated Reasoning in Explainable AI

João MARQUES-SILVA [1]
IRIT, CNRS, Toulouse, France

Keywords. Decision systems, Machine Learning, Explainable AI

The envisioned applications of machine learning (ML) in high-risk and safety-critical applications hinge on systems that are robust in their operation and that can be trusted. Automated reasoning offers the solution to ensure robustness and to guarantee trust. This talk overviews recent efforts on applying automated reasoning tools in explaining black-box (and so non-interpretable) ML models [6], and relates such efforts with past work on reasoning about inconsistent logic formulas [11]. Moreover, the talk details the computation of rigorous explanations of black-box models, and how these serve for assessing the quality of widely used heuristic explanation approaches. The talk also covers important properties of rigorous explanations, including duality relationships between different kinds of explanations [7,5,4]. Finally, the talk briefly overviews ongoing work on mapping practical efficient [8,3] but also tractable explainability [9,10,2,1].

References

[1] Martinc C. Cooper and Joao Marques-Silva. On the tractability of explaining decisions of classifiers. In *CP*, October 2021.
[2] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On efficiently explaining graph-based classifiers. In *KR*, November 2021.
[3] Alexey Ignatiev and Joao Marques-Silva. SAT-based rigorous explanations for decision lists. In *SAT*, 2021.
[4] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva. From contrastive to abductive explanations and back again. In *AI*IA*, pages 335–355, 2020.
[5] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In *KR*, November 2021.
[6] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, pages 1511–1519, 2019.
[7] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. On relating explanations and adversarial examples. In *NeurIPS*, 2019.
[8] Yacine Izza and Joao Marques-Silva. On explaining random forests with SAT. In *IJCAI*, August 2021.
[9] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. In *NeurIPS*, 2020.
[10] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In *ICML*, pages 7469–7479, 2021.
[11] João Marques-Silva and Carlos Mencía. Reasoning about inconsistent formulas. In *IJCAI*, pages 4899–4906, 2020.

[1]Corresponding Author: Joao Marques-Silva, `joao.marques-silva@irit.fr`