Fairness, Accountability and Transparency in Artificial Intelligence: A Case Study of Logical Predictive Models

Kacper Sokol
K.Sokol@bristol.ac.uk
Department of Computer Science, University of Bristol
Bristol, United Kingdom

ABSTRACT

Machine learning – the part of artificial intelligence aimed at eliciting knowledge from data and automated decision making without explicit instructions – is making great strides, with new algorithms being invented every day. These algorithms find myriads of applications, but their ubiquity often comes at the expense of limited interpretability, hidden biases and unexpected vulnerabilities. Whenever one of these factors is a priority, the learning algorithm of choice is often a method considered to be inherently interpretable, e.g. logical models such as decision trees. In my research I challenge this assumption and highlight (quite common) cases when the assumed interpretability fails to deliver. To restore interpretability of logical machine learning models (decision trees and their ensembles in particular) I propose to explain them with class-contrastive counterfactual statements, which are a very common type of explanation in human interactions, well-grounded in social science research. To evaluate transparency of such models I collate explainability desiderata that can be used to systematically assess and compare such methods as an addition to user studies. Given contrastive explanations, I investigate their influence on the model's security, in particular gaming and stealing the model. Finally, I evaluate model fairness, where I am interested in choosing the most fair model among all the models with equal performance.

CCS CONCEPTS

• **General and reference** → **General literature**; • **Theory of computation** → Machine learning theory; • **Computing methodologies** → Supervised learning.

KEYWORDS

fairness, accountability, transparency, logical models

ACM Reference Format:
Kacper Sokol. 2019. Fairness, Accountability and Transparency in Artificial Intelligence: A Case Study of Logical Predictive Models. In *2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306618.3314316

1 SOCIAL ARTIFICIAL INTELLIGENCE

Whether driven by a curiosity or legal requirements, explaining artificial intelligence (AI) systems – with a particular focus on predictive modelling – is an important first step to democratise the power of automated decision making. Therefore, in my research I focus on interpreting and explaining predictions of logical predictive models, decision trees in particular, given their simplicity, predictive power and popularity in real-life applications. This is a very interesting research problem since these models are often considered inherently interpretable. I challenge this assumption by arguing that simply being able to inspect a series of logical conditions leading to a particular decision is not enough to endorse its interpretability – e.g. the difficulty of attributing the result to a subset of these conditions – and logical reasoning is required on top of that to foster understanding. Consider a decision tree trained on large quantities of data with weak regularisation leading to good predictive performance at the expense of the tree being wide and deep. Hence any explanation native to the structure of the model is a long conjunction of logical conditions applied to the data features, which may be hard to interpret even for domain experts.

2 RESEARCH QUESTIONS

In my research, I develop approaches to explain such automated decisions in a sparse, interactive and sound way, suitable for both domain experts and a lay audience. To this end, I use class-contrastive counterfactual statements that usually conform to the following template:

> "The prediction is <prediction>. Had a small subset of
> features been different <foil>, the prediction would
> have been <counterfactual prediction> instead."

Despite them being a novelty in interpretable AI, counterfactual explanations are well grounded in social sciences, where they have been researched for decades – see a review by Miller et al. [4]. These studies have shown that they are very common in the process of human explanation, hence they are a good candidate for explaining automated decisions since they feel natural for human explainees. Moreover, counterfactual explanations have been deemed to satisfy the legal requirements of automated decisions explainability imposed in May 2018 by the General Data Protection Regulation[1] (GDPR) in the European Union [7].

Explanations of predictive systems are notoriously difficult to evaluate given their social nature. Doshi-Velez and Kim [1] have reviewed approaches to evaluating them with user studies, however there are many aspects of explanations that user studies cannot

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01.
https://doi.org/10.1145/3306618.3314316

[1] https://publications.europa.eu/s/inbX.

capture unless explicitly addressed. This includes *usability requirements* of an explanation such as its soundness, completeness and chronology, to name a few. *Functional aspects* (e.g. the problem supervision level, computational complexity of the explanation or applicable model class), *operational aspects* (used explanatory medium, system interaction model or explanation target audience) and *safety requirements* (information leakage, explanation misuse or explanation robustness) of an explanation are other important aspects ignored by user studies that should always be considered but rarely ever are. One reason behind this under-specification of explainable methods may be a lack of a coherent list of explainability desiderata for AI systems [2]. This observation has led me to review relevant computer science and social sciences literature on explanations and build a comprehensive list of desiderata for explainable systems spanning all these dimensions.

Functional and operational aspects of an explanation can be assessed analytically, but its accountability (security, privacy and robustness) requires a designated suite of tests. This is especially important for logical models as they use precise splits on data features in their decision boundaries that can be easily revealed with counterfactual explanations. Therefore, I want to investigate the extent of this phenomenon and techniques to mitigate compromising the accountability of predictive models when explaining them. In particular, I will investigate the lower bound on the number of explanations that allow stealing a model and techniques to prevent such malicious practices by providing fuzzy thresholds or replacing them with quantitative adjectives.

Fairness is another important social aspect of predictive modelling, as biases in data can be easily transferred to a predictive model during its training phase. A wealth of literature tackles fairness from disparate treatment and disparate impact perspectives [8]; however, to the best of my knowledge, there is no work that targets fairness of the model selection process itself. Given the stochastic nature of predictive modelling, it is usually possible to train multiple models with a similar overall predictive performance with predictions differing only for a small subset of the data set – predictive trade-offs between a number of individuals. I would like to investigate this phenomenon for logical predictive models to come up with "fairness bounds" for each individual in a data set with respect to a group of models with the same or comparable predictive performance for the whole population. Such an approach can be understood as applying the "presumption of innocence" rule for AI systems as the objective is to find the most beneficial model, hence prediction, for an individual without jeopardising the population-wise performance.

3 WORK PROGRESS

To date, I have designed an approach to explain decision trees with counterfactual statements. I have implemented this approach and used it as a part of an interactive system that can explain classification outcomes of a decision tree model via a dialogue with a user (both text-based chat and voice-enabled conversation). Furthermore, I have collated explainable systems desiderata and used them to compare and contrast my method against theoretical capabilities of contrastive explanations [3] and their various algorithmic implementations [6, 7]. This allowed me to show that my method is better in preserving theoretical properties of counterfactual explanations – an advantage of a model-specific approach – whereas others lose many of them.

My next step is to carry out user studies to demonstrate the advantages of my method over the interpretability of logical models (conjunctions of logical conditions extracted from the model) and model-agnostic approaches such as LIME [5]. I will also generalise my approach to other logical models, e.g. rule lists and sets, and their ensembles, e.g. random forests. The latter objective poses an interesting challenge of utilising the power of counterfactuals for each decision tree in an ensemble in order to create a smarter than majority count voting strategy. I will conclude my research in this area by publishing an open source implementation of my method to democratise interpretability of logical models. The next stage of my research will be to consider the safety and security of a predictive model and its training data given all the information that is revealed by counterfactual explanations. In particular, I will seek to determine the difficulty of reverse-engineering the decision boundaries and training data with a bounded number of explanations (which should be feasible given theoretical guarantees of my approach) and how to mitigate these threats. Finally, I will investigate proposed earlier "fairness bounds" for individual data points given a set of logical models with a fixed performance for the whole population.

REFERENCES

[1] Finale Doshi-Velez and Been Kim. 2018. Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, Gewerbestrasse 11, 6330 Cham, Switzerland, 3–17.

[2] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, ACM, New York, NY, USA, 126–137.

[3] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[4] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 36.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, ACM, New York, NY, USA, 1135–1144.

[6] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, ACM, New York, NY, USA, 465–474.

[7] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887.

[8] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1171–1180.