juliacon

Explaining black-box algorithms using CounterfactualExplanations.jl

Patrick Altmeyer[1] and Cynthia Liem[1]

[1]Delft University of Technology

ABSTRACT

Machine learning models like deep neural networks have become so complex and opaque over recent years that they are generally considered as black boxes. Nonetheless such models play a key role in modern automated decision-making systems. Counterfactual explanations (CE) can help programmers make sense of the systems they build: they explain how inputs into a system need to change for it to produce different decisions. Explanations that involve realistic and actionable changes can be used for the purpose of algorithmic recourse (AR): they offer individuals subject to algorithms a way to turn a negative decision into positive one. In this article we discuss the usefulness of counterfactual explanations for interpretable machine learning and demonstrate its implementation in Julia using the `CounterfactualExplanations` package.

Keywords

Julia, Interpretable Machine Learning, Counterfactual Explanations, Algorithmic Recourse

5. References

[1] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.