Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals

Patrick Altmeyer[1], **Mojtaba Farmanbar**[2], **Arie van Deursen**[1], **Cynthia C. S. Liem** [1]

[1]Delft University of Technology,
[2]ING

Abstract

Counterfactual explanations offer an intuitive and straightforward way to explain black-box models and offer algorithmic recourse to individuals. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn how the input data is distributed. This effectively reallocates the task of learning realistic explanations for the data from the model itself to the surrogate. Consequently, the generated explanations may seem plausible to humans but need not necessarily describe the behaviour of the black-box model faithfully. We formalise this notion of faithfulness through the introduction of a tailored evaluation metric and propose a novel algorithmic framework for generating **E**nergy-**C**onstrained **C**onformal **Co**unterfactuals that are only as plausible as the model permits. Through extensive empirical studies, we demonstrate that *ECCCo* reconciles the need for faithfulness and plausibility. In particular, we show that for models with gradient access, it is possible to achieve state-of-the-art performance without the need for surrogate models. To do so, our framework relies solely on properties defining the black-box model itself by leveraging recent advances in energy-based modelling and conformal prediction. To our knowledge, this is the first venture in this direction for generating faithful counterfactual explanations. Thus, we anticipate that *ECCCo* can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

References