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

1. Introduction

In section Section 1

2. Methodological background

Counterfactual search happens in the feature space: we are interested in understanding how we need to change individual attributes in order to change the model output to a desired value or label [5]. Typically the underlying methodology is presented in the context of binary classification: $M : \mathcal{X} \mapsto y$ where and $y \in \{0, 1\}$. Let $t = 1$ be the target class and let $\overline{x}$ denote the factual feature vector of some individual outside of the target class, so $\overline{y} = M(\overline{x}) = 0$. Then the counterfactual search objective originally proposed by [12] is as follows

$$\min_{\underline{x} \in \mathcal{X}} h(\underline{x}) \quad \text{s. t.} \quad M(\underline{x}) = t \tag{1}$$

where $h(\cdot)$ quantifies how complex or costly it is to go from the factual $\overline{x}$ to the counterfactual $\underline{x}$. To simplify things we can restate this constrained objective (Equation 1) as the following unconstrained and differentiable problem:

$$\underline{x} = \arg\min_{\underline{x}} \ell(M(\underline{x}), t) + \lambda h(\underline{x}) \tag{2}$$

Here ℓ denotes some loss function targeting the deviation between the target label and the predicted label and λ governs the stength of the complexity penalty. Provided we have gradient access for the black-box model M the solution to this problem (Equation 2) can be found through gradient descent. This generic framework lays the foundation for most state-of-the-art approaches to counterfactual search and is also used as the baseline approach in `CounterfactualExplanations`.

That being said, numerous extensions of this simple approach have been developed since counterfactual explanations were first proposed in 2017 (see [11] and [2] for surveys). The various approaches largely differ in how they define the complexity penalty. In [12], for example, $h(\cdot)$ is defined in terms of the Manhattan distance between factual and counterfactual feature values. While this is an intuitive choice, it is too simple to address many of the desirable properties of effective counterfactual explanations that have been set out. These desiderata include: **closeness** -the average distance between factual and counterfactual features should be small ([12]); **actionability** - the proposed feature perturbation should actually be actionable ([10], [7]); **plausibility** - the counterfactual explanation should be plausible to a human ([1]); **unambiguity** - a human should have no trouble assigning a label to the counterfactual ([8]); **sparsity** - the counterfactual explanation should involve as few individual feature changes as possible; **robustness** - the counterfactual explanation should be robust to domain and model shifts ([9]); **diversity** - ideally multiple diverse counterfactual explanations should be provided ([6]); and **causality** - counterfactual explanations reflect the structual causal model underlying the data generating process ([4],[3]).

3. Using `CounterfactualExplanations`

3.1 Counterfactual generators

4. Empirical example

5. Related and future work

6. References

[1] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[2] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

[3] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Con-*

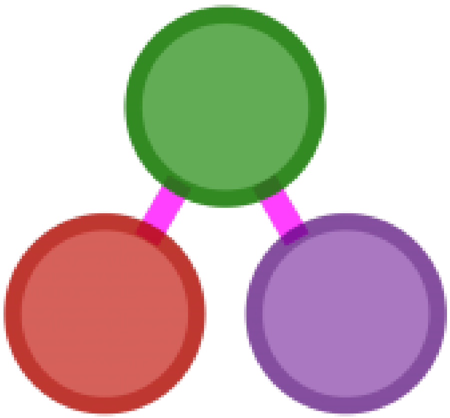

Fig. 1. Figure

ference on Fairness, Accountability, and Transparency, pages 353–362, 2021.

[4] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *arXiv preprint arXiv:2006.06831*, 2020.

[5] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[6] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[7] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[8] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR, 2021.

[9] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *arXiv preprint arXiv:2102.13620*, 2021.

[10] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[11] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[12] 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.