Intended for submission to the *Fifth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*

Dynamics in Algorithmic Recourse

Trustworthy Artificial Intelligence for Finance and Economics

Patrick Altmeyer

1 INTRODUCTION

Recent advances in artificial intelligence (AI) have propelled its adoption in domains outside of computer science including health care, bioinformatics and genetics. In finance, economics and other social sciences, applications of AI are still relatively limited. Decision-making in these fields has traditionally been guided by Generalized Linear Models (GLM), which are theoretically founded, interpretable and often sufficient to model relationships between variables. Model interpretability is crucial in the social sciences context, because inference is typically at least as important as predictive performance. Decision-makers in the social sciences are also typically required to explain their decisions to human stakeholders: central bankers, for example, are held accountable by the public for the policies they decide on. It is therefore not surprising that practitioners and academics in these fields are reluctant to adopt AI technologies that ultimately cannot be trusted. Deep learning models, for example, are generally considered as black boxes and therefore difficult to apply in a context that demands explanations. This PhD project is focused on exploring and developing methodologies that improve the trustworthiness of AI and thereby enable its application in Finance and Economics.

The remainder of this extended abstract is structured as follows: Section 2 presents one of the research questions I have investigated during the first months of my PhD: how do counterfactual explanations handle dynamics? Section 3 places this work in the broader context of Trustworthy AI for Finance and Economics.

©-Notice

2 DYNAMICS IN ALGORITHMIC RECOURSE

Counterfactual explanations (CE) explain how inputs into a model need to change for it to produce different outputs. They are intuitive, simple and intrinsically linked to the potential outcome framework for causal inference, which social scientists are familiar with. Counterfactual explanations that involve realistic and actionable changes can be used for the purpose of **Algorithmic Recourse** (AR) to help individuals facing adverse decisions. An example relevant to the Finance and Econonomics domain is consumer credit: in this context AR can be used to guide individuals to improve their credit worthiness, should they have previously been denied access to credit based on a black-box decision-making system .

Existing work on CE and AR has largely been limited to the static setting: given some classifier $M : \mathcal{X} \mapsto \mathcal{Y}$ we are interested in finding close (Wachter, Mittelstadt, and Russell 2017), actionable (Ustun, Spangher, and Liu 2019), plausible Schut et al. (2021), sparse (Schut et al. 2021), diverse (Mothilal, Sharma, and Tan 2020) and ideally causally founded counterfactual explanations (Karimi, Schölkopf, and Valera 2021) for some individual x. The ability of counterfactual explanations to handle dynamics like data and model shifts remains a largely unexplored research challenge at this point (Verma, Dickerson, and Hines 2020). Only one recent work considers the implications of **exogenous** domain shifts on the validity of recourse (Upadhyay, Joshi, and Lakkaraju 2021). The authors propose a simple minimax objective, that minimizes the counterfactual loss function for a maximal domain and model shift. They show that their approach yields more robust counterfactuals than existing approaches.

This project investigates **endogenous** domain and model shifts, that is shifts that occur when AR is actually impemented by a proportion of individuals and the classifier is updated in response. Figure 1 illustrates this idea for a binary problem involving a probabilistic classifier and a greedy counterfactual generator proposed by Schut et al. (2021): AR leads to a domain shift, which in turn causes a drastic model shift. As this game of implementing AR and updating the classifier is repeated, individuals who receive and implement algorithmic recourse end up forming a distinct

subgroup inside the target class, which may leave them vulnerable to discrimination. Through future experiments we want to investigate if this phenomenon is robust across different benchmark datasets and counterfactual generators.

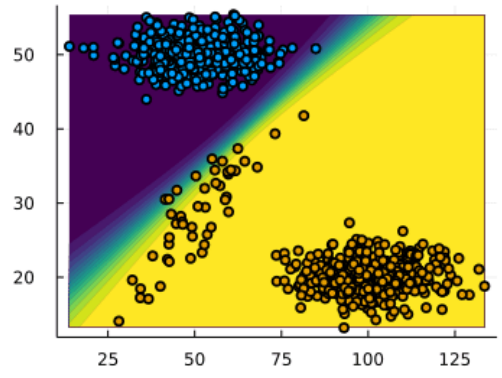

Figure 1: PLACEHOLDER: The dynamics of algorithmic recourse.

3 RELATED AND FUTURE WORK

3.1 Benchmarking CE in Julia

Alongside my research I have developed open-source implementations related to explainable AI. CounterfactualExplanations.jl is a Julia package that can be used to generate counterfactual explanations for models developed and trained not only in Julia, but also in other popular programming languages like Python and R. I have recently submitted the package along with a companion paper as a proposal for a main talk at JuliaCon. BayesLaplace.jl is a small Julia package that can be used to recover Bayesian representations of deep neural networks through Laplace approximation in a post-hoc manner. It is inspired by a recent paper (Daxberger et al. 2021) and has also been submitted to JuliaCon. Finally, deepvars is an R package that implements an approach towards vector autoregression that leverages deep learning. This was originally my master's thesis and later presented at the NeurIPS 2021 MLECON workshop. I have also published several blog posts on explainable AI and probabilisitic ML in an effort to make my research accessible to a broad audience.

3.2 Probabilistic methods for realistic counterfactual explanations

Probabilistic machine learning can be leveraged in this context and more generally facilitates inference and interpretability. It is also closely related to Bayesian statistics, which has played an important role in both finance and economics for many years.

To ensure that the generated explanations are realistic it is important to understand which input-output pairs are likely and which are not. To quantify their joint likelihood, previous work has either relied on generative models or restricted the analysis to probabilistic models that incorporate uncertainty in their predictions. While the former approach is more versatile since it is applicable to both deterministic and probabilistic models, the latter is computationally much more efficient. In my work I want to explore how recent advances in post-hoc uncertainty quantification can be leveraged to generate realistic and unambiguous counterfactual explanations for any model.

3.3 CE for time series

Data sets in finance and economics typically involve time series data. Therefore, I am naturally interested in the application of explainable AI to sequential data, an area which has so far not been explored extensively. In the future, I want to work on counterfactual explanations for time series models. I am also interested in seeing if and how Laplace approximation can be used for Bayesian deep learning with time series data. I hope that the findings from both of these projects can ultimately be used to build complex but interpretable time series models for classification and forecasting in finance and economics.

3.4 Explainable black-box models for time series

For example, I would like to leverage effortless Bayesian deep learning to make our proposed Deep Vector Autoregression model explainable.

REFERENCES

Antorán, Javier, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. 2020. "Getting a Clue: A Method for Explaining Uncertainty Estimates." *arXiv Preprint arXiv:2006.06848*.

Daxberger, Erik, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. "Laplace Redux-Effortless Bayesian Deep Learning." *Advances in Neural Information Processing Systems* 34.

Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems." *arXiv Preprint arXiv:1907.09615*.

Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. 2021. "Algorithmic Recourse: From Counterfactual Explanations to Interventions." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–62.

Mothilal, Ramaravind K, Amit Sharma, and Chenhao Tan. 2020. "Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–17.

Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. "Generating Interpretable Counterfactual Explanations by Implicit Minimisation of Epistemic and Aleatoric Uncertainties." In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.

Upadhyay, Sohini, Shalmali Joshi, and Himabindu

Lakkaraju. 2021. "Towards Robust and Reliable Algorithmic Recourse." *arXiv Preprint arXiv:2102.13620*.

Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. "Actionable Recourse in Linear Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.

Verma, Sahil, John Dickerson, and Keegan Hines. 2020. "Counterfactual Explanations for Machine Learning: A Review." *arXiv Preprint arXiv:2010.10596*.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harv. JL & Tech.* 31: 841.