Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements

Kacper Sokol and **Peter Flach**
Department of Computer Science, University of Bristol
K.Sokol@bristol.ac.uk, Peter.Flach@bristol.ac.uk

Abstract

Machine learning models have become pervasive in our everyday life; they decide on important matters influencing our education, employment and judicial system. Many of these predictive systems are commercial products protected by trade secrets, hence their decision-making is opaque. Therefore, in our research we address interpretability and explainability of predictions made by machine learning models. Our work draws heavily on human explanation research in social sciences: contrastive and exemplar explanations provided through a dialogue. This user-centric design, focusing on a lay audience rather than domain experts, applied to machine learning allows explainees to drive the explanation to suit their needs instead of being served a precooked template.

1 Introduction

Over the past decade machine learning systems have become inseparable part of technology that we use every day. Since they are so pervasive and ubiquitous we enjoy their benefits as long as the algorithm operations align with our expectations. Machine learning pipeline usually includes collecting and preprocessing data, training a model, tweaking its parameters to minimise a loss function and deploying it in the wild. After that, especially in application driven machine learning, the model is rarely maintained unless it misbehaves, in which case more data are collected and more parameters tweaked to improve the performance. Such approach can be attributed to paying too much attention to optimising the solution and not enough to the task itself and the phenomenon underlying the data. Critically reflecting upon the problem and chosen solution, hence comprehending its inner logic, can improve one's understanding of the whole system and lead to fixing the cause rather than the symptoms.

2 Research Background

To better understand the behaviour of a machine learning system, we can design methods and tools capable of peeking inside these black-boxes. Explaining both: data used for the training of a machine learning model and the model itself is an active area of research, however our work focuses on interpreting and explaining predictions made by machine learning models i.e. providing the explainee with a rationale behind a particular decision outputted by a model.

Interpretability of the machine learning predictions is important for a variety of reasons. Scientists not interested in optimising accuracy of their models may be driven by a scientific curiosity hoping to use machine learning to elicit new knowledge from data. Moreover, in some applications, like medicine, the users have to trust the predictions, hence they have to understand the underlying decisive mechanisms [Lipton, 2017]. The case of AI making – sometimes legally binding [Kusner *et al.*, 2017] – decisions about humans without them knowing drawn attention of lawmakers and regulators leading to DARPA's Explainable AI (XAI) project[1] and European Union's General Data Protection Regulation (GDPR) coming into force in May 2018 [Wachter *et al.*, 2017].

3 Research Problem

Our research focuses on designing and implementing tools and techniques useful for explaining and interpreting predictions derived by machine learning algorithms. Interpretability approaches can be divided into two categories: *model dependent*, which have to be designed independently for each class of machine learning models; and *model agnostic*, which can work with any model, therefore are more universal.

Large volume of the machine learning explainability work, e.g. [Ribeiro *et al.*, 2016], uses concepts that explainees lacking technical background may struggle to understand. Our research addresses this shortcoming by focusing on solutions that are both: model agnostic and user-centric. This is achieved by explaining algorithmic predictions through a human-machine explanatory dialogue using *contrastive* statements and *exemplar-based* explanations [Miller *et al.*, 2017].

Counterfactual explanations can be applied to classification outcomes produced by any machine learning model, although their computation may differ for each model family. Furthermore, such approach is well grounded in human explanation research in social sciences [Miller, 2017]. Contrastive explanations are also legally sufficient and exceed GDPR expectations [Wachter *et al.*, 2017]. Moreover, an explanatory

[1] https://www.darpa.mil/program/explainable-artificial-intelligence

dialogue gives the explainee control over the explanation – it can be steered in a selected direction instead of simply being a one-size-fits-all template. The conversation with a machine learning model can be facilitated through a text-based chat or a voice-enabled off-the-shelf digital personal assistant.

Furthermore, a data point of interest together with counterfactuals generated by a user interaction with the system can be composed into a coherent story explaining this particular classification outcome. The generated story alongside model explanation in the neighbourhood of this data point can be used to compose a narrative describing the classifier's behaviour supported with examples, similar to ones found in government guidelines explaining selected regulations[2].

4 Contributions

> "Your loan application has been *declined*. If you were a *skilled employee* instead of an *unskilled – resident*, your loan application would be *accepted*."

is an example of a class-contrastive counterfactual explanation for a classification outcome outputted by a machine learning model trained on UCI German credit (Statlog) dataset[3]. In our work, to date, we have reviewed literature relevant to explainability and interpretability of the machine learning process: data, models and predictions. Furthermore, we have developed an approach to generate class-contrastive counterfactual statements for predictions outputted by decision trees, nevertheless the proposed method can be easily generalised to the whole family of logical machine learning models, e.g. rule lists or sets. The approach takes advantage of access to the internal structure of a decision tree to measure pairwise distance between all its leafs. The distance metric encodes how many changes in the feature space are necessary to change the classification prediction. It is based on L1 distance, therefore it favours sparsity in the adjustment of feature values, what results in the shortest possible counterfactual statement being generated.

Our work challenges a popular belief that decision trees, and logical models in general, are inherently explainable given that we can see a conjunction of logical conditions leading to a particular prediction. This might be a case with small datasets, however for big decision trees, explaining a prediction with a long conjunction of logical conditions is neither an explanation nor it helps to attribute the decision to a subset of the logical conditions in that list. Therefore, the quality of such explanations deteriorates with the increase in the dimensionality of a dataset. The method proposed in our research does not exhibit this drawback; explanations are always sparse, the user is able to interrogate them, hence explore and guide them to personalise the explanation such that it suits one's needs, which is especially desirable when the audience lacks topic-specific background knowledge.

Furthermore, our explainability approach can help to identify errors (e.g. decrease in income changes the predicted credit score from *bad* to *good*) and biases (e.g. a prediction depends on applicant's gender) in the underlying machine learning model. In our research, we are currently focusing on datasets with human-interpretable features as they facilitate easy evaluation of the quality of produced explanations.

5 Future Work

In the near future, we will investigate how our approach to generating counterfactuals can be improved, the distance metric in particular, to take into account human preferences, and feature actionability and importance. The next step will be focused on story generation from the extracted counterfactuals to facilitate describing classification outcomes supported with examples in an accessible format. We also plan to evaluate the explainability power of our approach by designing appropriate analytical tools and carrying out user studies among various communities. Adapting our approach to domains different than credit scoring is also a possible research direction.

Furthermore, we want to improve our system to handle human-incomprehensible features by expressing the counterfactuals in the same domain as the input space or by learning high-level concepts based on the underlying model's internal data representation [Kim *et al.*, 2017]. Finally, we will design and implement algorithms to generate counterfactual explanations for geometric and probabilistic models and investigate explainability and interpretability of other components of the machine learning process: data and models.

References

[Kim *et al.*, 2017] Been Kim, Justin Gilmer, Fernanda Viegas, Ulfar Erlingsson, and Martin Wattenberg. Tcav: Relative concept importance testing with linear concept activation vectors. *arXiv preprint arXiv:1711.11279*, 2017.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[Lipton, 2017] Zachary C Lipton. The doctor just won't accept that! In *Interpretable ML Symposium, 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[Miller *et al.*, 2017] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 36, 2017.

[Miller, 2017] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, (Forthcoming), 2017.

[2] https://www.gov.uk/expenses-if-youre-self-employed

[3] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)