Desiderata for Interpretability: Explaining Decision Tree Predictions with Counterfactuals

Kacper Sokol, Peter Flach
Department of Computer Science, University of Bristol, Bristol, UK
{K.Sokol,Peter.Flach}@bristol.ac.uk

Abstract

Explanations in machine learning come in many forms, but a consensus regarding their desired properties is still emerging. In our work we collect and organise these explainability desiderata and discuss how they can be used to systematically evaluate properties and quality of an explainable system using the case of class-contrastive counterfactual statements. This leads us to propose a novel method for explaining predictions of a decision tree with counterfactuals. We show that our model-specific approach exploits all the theoretical advantages of counterfactual explanations, hence improves decision tree interpretability by decoupling the quality of the interpretation from the depth and width of the tree.

Introduction

Counterfactual explanations are becoming a *de facto* standard for explaining automated decisions (Wachter, Mittelstadt, and Russell 2018; Miller 2019; Tolomei et al. 2017). In their most popular form they follow this template:

> "The prediction is <prediction>. Had a small subset of features been different <foil>, the prediction would have been <different prediction> instead."

Such counterfactual explanations are deemed particularly useful for a lay audience since they do not presuppose any background in computer science or artificial intelligence. They particularly gained in popularity when the European Union's General Data Protection Regulation (GDPR) came into force in May 2018 requiring organisations that use algorithmic decision making to provide explanations on the client's request. To address this requirement Wachter, Mittelstadt, and Russell (2018) showed that counterfactuals are "user-friendly" and compliant with the GDPR.

Explainability and Its Desiderata

The recent surge in interpretability and explainability research in AI may suggest that this is a new research topic, but in fact it has been an active research area for much longer in the humanities – to the point that researchers have started to agree on a coherent list of *desiderata* for explainable systems (Kulesza et al. 2013; 2015). The seminal work of Miller (2019) shows that counterfactual explanations comply with most of these desiderata, although not necessarily all of them. For example, consider *completeness* – the extent to which the explanation covers the whole underlying system – of a counterfactual explanation. A counterfactual foil applies only to the queried data point and does not generalise to a broader class of similar data points; i.e., had the foil been true, *ceteris paribus*, the classification outcome would change. The possible incompleteness of a counterfactual explanation is often overlooked by the explainees as humans are known to overgeneralise. Therefore, for a counterfactual explanation to satisfy *completeness*, the explainer would have to provide the *necessary conditions* under which the foil, hence the explanation, holds (*contextfullness*). An example of such a generalised counterfactual would be "Had you earned £10,000 more, your loan application would be accepted *provided you do not change your mortgage and keep the same job*." Completeness and contextfullness are two of many explainability desiderata. Other important aspects of an explanation are: *soundness* – truthfulness of the explanation with respect to the predictive model; *interactiveness* – interactive explanations are better than static ones; *actionability* – explanations that give the user suggestions how to change the model's prediction are preferred; *chronology* – more recent causes of an event are preferred; *coherence* – explanations should agree with the user's mental model; *novelty* – the explanation should not repeat what the user already knows; *complexity* – the complexity of an explanation should be tuned to the user's ability and knowledge; and *parsimony* – shorter explanations are more comprehensive.

Systematically evaluating these properties of explainable techniques can be a useful precursor to user studies to show their capabilities and compliance with the best practices in the field. Despite theoretical guarantees of some desiderata for some explainability approaches, these guarantees can be lost in implementation. For example, model-agnostic approaches can render some desiderata difficult to achieve since they cannot take advantage of model-specific aspects of the predictive algorithms. What appears to be lacking in the literature is a connection between general desiderata and properties of specific methods. At best, some studies select a subset of desiderata and evaluate a selected approach for a particular task using them; for example Kulesza et al. (2013) evaluate soundness and completeness of interactive visualisations for music recommendation, and Kulesza et al. (2015) examine soundness, completeness, interactiveness, parsimony and actionability for interactive visualisations of a Naïve Bayes spam email classifier. A standardised list of explainability desiderata would provide a common ground for easy comparison of explainability approaches. As it stands, many implementations do not exploit the full potential of the selected explanatory technique, for example a method based on counterfactuals might not take advantage of their social and interactive aspects.

Counterfactuals for Decision Trees

Our contributions in this field are twofold. First, we collect and review desiderata for explainability techniques in AI. This allows a systematic comparison of explainability methods as an addition to user studies and uncover discrepancies between the theoretical capabilities and a specific implementation of a given technique. We demonstrate the usefulness of these desiderata using the case of counterfactual explanations and their two popular implementations from the literature: Tolomei et al. (2017) and Wachter, Mittelstadt, and Russell (2018).

Secondly, we propose a novel algorithm for composing counterfactual explanations of data points classified with a decision tree, and evaluate it against the full set of explainability desiderata. Here we advocate a model-specific approach in order to exploit all theoretical capabilities of counterfactual explanations in the implementation. The algorithm works with a decision tree model but can be easily generalised to the whole family of logical machine learning models and their ensembles. To guarantee fast and easy search, customisability and completeness of generated counterfactuals we map the internal tree structure to a leaf-to-leaf *counterfactual distance matrix* that describes how many (and what) changes are required for a given root-to-leaf path, hence a data point, to change its classification outcome. This distance matrix is built using a meta-feature set that is composed of logical conditions extracted form the tree's internal nodes. Each condition can either be true, false, or does not apply to a particular data point if a given feature is not used on a particular root-to-leaf path in the tree. Therefore, for every classified data point the algorithm can retrieve its smallest alteration that results in a different classification outcome.

Related Work

Recent explainable AI literature displays three main trends. The first one concerns studies discussing what is generally desired of explanations (Lipton 2018; Kulesza et al. 2013; 2015; Doshi-Velez and Kim 2017). Here the most related work to our first contribution is Doshi-Velez and Kim (2017), who provide desiderata for evaluating explainability techniques from the perspective of user studies, whereas our list of desiderata is much more comprehensive. The second trend investigates theoretical properties of selected explainable approaches, e.g. Miller (2019) for counterfactuals. The third trend involves papers discussing implementations and experimental results for selected methods. Here Wachter, Mittelstadt, and Russell (2018) study generating counterfactuals for differentiable models and Tolomei et al. (2017) do that for random forests, which are similar to our second contribution. Both methods are supported with an empirical evaluation, however neither of them takes full advantage of counterfactual explanations or explicitly compares capabilities of their implementation against explainability desiderata.

Conclusions and Future Work

Our work presents a collection of explainability desiderata and shows how they can be used to systematically evaluate and compare explainability approaches. We also propose an approach to explain decision trees (and other logical models) with counterfactual statements. Finally, we use the explainability desiderata to show that our method is compliant with all of them – an advantage of a model-specific approach – as opposed to other algorithms proposed in the literature.

In future, we will investigate a variety of distance functions that can be used with our approach and assess pros and cons of the resulting counterfactuals. Then, we will empirically evaluate the quality and effectiveness of our explanations with user studies where the participants will be shown two types of explanations for a classification outcome: a conjunction of logical conditions extracted form the underlying logical model and our counterfactuals for multiple distance functions. Finally, we will extend our method to ensembles of logical models, in particular random forests.

References

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; and Wong, W.-K. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, 3–10. IEEE.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. ACM.

Lipton, Z. C. 2018. The mythos of model interpretability. *Communications of the ACM* 16(3):30:31–30:57.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Tolomei, G.; Silvestri, F.; Haines, A.; and Lalmas, M. 2017. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 465–474. ACM.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31(2):841–887.