Manipulating SHAP via Adversarial Data Perturbations (Student Abstract)

Hubert Baniecki, Przemyslaw Biecek

Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland
{hubert.baniecki.stud, przemyslaw.biecek}@pw.edu.pl

Abstract

We introduce a model-agnostic algorithm for manipulating SHapley Additive exPlanations (SHAP) with perturbation of tabular data. It is evaluated on predictive tasks from healthcare and financial domains to illustrate how crucial is the context of data distribution in interpreting machine learning models. Our method supports checking the stability of the explanations used by various stakeholders apparent in the domain of responsible AI; moreover, the result highlights the explanations' vulnerability that can be exploited by an adversary.

Introduction

SHapley Additive exPlanations (Lundberg and Lee 2017) became a state-of-the-art explanation method for various machine learning use-cases like model debugging, trustworthy decision making, and knowledge discovery. However, a model-agnostic explanation is usually a function of model and data; therefore, both elements can be altered to manipulate SHAP. Manipulating explanations means a significant change of their numerical or visual representation, while the significance varies between use-cases and domains. Slack et al. (2020) provide a framework for manipulating SHAP via changing the model in question. It is an adversarial attack on a post-hoc explanation method where one constructs a globally biased (racist) classifier that produces safe explanations of the model's individual predictions. Correspondingly, it is possible to manipulate explanations by changing the reference data used to produce them. Ghorbani, Abid, and Zou (2019) utilize gradient-based optimization approaches, suited mainly for deep neural networks, to change feature importance via data perturbations, which is contradictory to altering the black-box model. Mishra et al. (2021) further discuss methods related to our contribution to the domain merging explainability with adversarial machine learning.

In contrast, this paper extends the work of Baniecki, Kretowicz, and Biecek (2021) by introducing a model-agnostic algorithm for manipulating SHAP via data perturbations. The main motivation is to provide a tool for model developers to assess the robustness of SHAP by giving a certificate of the explanations' stability to data shifts. Specifically, this result highlights the tabular explanations' shortcomings, which is alarming for auditors and prediction recipients.

Manipulation Method

We apply a genetic-based algorithm suited for any black-box machine learning predictive model trained on tabular data, which is as follows. The *population* consists of distinct datasets (individuals) initiated with the original reference data. The *mutation* operator adds white noise to the variables' values, *crossover* creates children by exchanging the variables' columns between the individuals, *evaluation* calculates the loss (fitness) value, and *selection* chooses proper individuals to remain in the population for the next epoch.

We acknowledge that this is only an exemplary implementation of the genetic operators, which are the main subject for extensions in future work. For example, mutation of categorical variables would involve substituting the variables' values from a given set. For simplicity, we omit to change the values of categorical variables in this work. The central part of the algorithm is a loss function — a weighted sum of two terms: (1) a distance between the values of explanations, e.g. L_1 distance; (2) a distance between the ordering of variables (ranking) in explanations, e.g. Kendall tau distance. The second term magnifies manipulating the interpretation of SHAP, as usually, it is only crucial for the stakeholders to recognize the most and least important variables.

This work considers utilizing genetic-based data perturbations to change SHAP importance for a given model (a global explanation of the model's reasoning) or change SHAP attributions for a given prediction (a local explanation for a single observation). We craft such explanations by minimizing the loss between the manipulated explanation and an arbitrarily chosen *target*. It is also possible to maximize the loss without a target to evaluate the explanations by investigating the occurring shift in data.

Experimental Setup

We provide illustrative scenarios based on two predictive tasks: `heart` disease classification and `apartment` price estimation (Poursabzi-Sangdeh et al. 2021). To both, we fit an XGBoost tree-ensemble model, which is one of the most popular machine learning algorithms suited for tabular data; moreover, very commonly explained with TreeSHAP (Lundberg et al. 2020). For additional crucial context, we measure the drift from the reference data with the mean of the variables' Jensen–Shannon divergence.

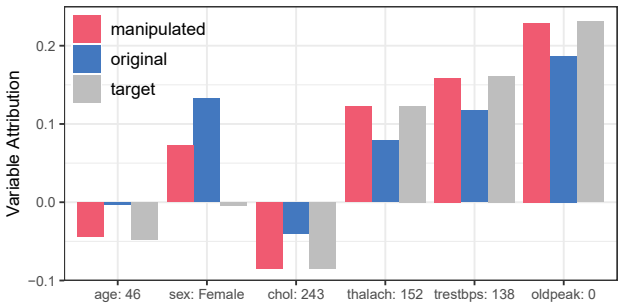

Figure 1: Original and manipulated SHAP attribution in the `heart` disease classification. The attribution of `sex` diminishes as other variables become more prevalent.

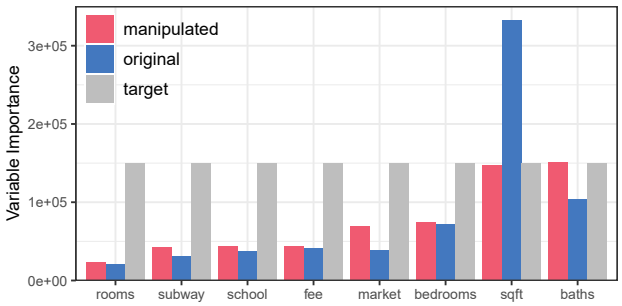

Figure 2: Original and manipulated SHAP importance in the `apartment` price estimation. The importance of variables becomes more evenly distributed and provides a less meaningful interpretation.

Results

Consider the following scenarios. In `heart`, one aims to explain the model to allow for trustworthy decision making. In this scenario, an adversary manipulates the local SHAP attribution explanation for a given observation to hide an impact of the variable `sex`, which refers to model fairness (Aivodji et al. 2019). Figure 1 illustrates the original SHAP where `sex` is the second most important variable attributing to the prediction. It becomes fifth in the manipulated explanation. This case is when an auditor has no access to the reference data, e.g. in healthcare, or is given an exemplary data subset, e.g. in the research review.

Alternatively, in the `apartment` price estimation, one aims to explain the model to allow for a reliable knowledge discovery process. In this scenario, a developer evaluates the global SHAP importance explanation to analyze its stability under the possible shift in data. Figure 2 illustrates the manipulated SHAP, in which the dominant variable `sqft` appears as less important. This case occurs when stakeholders admit that the reference data distribution might change or not be representative.

Supplementarily, Figure 3 shows the loss curves of a genetic-based manipulation algorithm in both cases. Table 1 reports performance measures: the distances between the original and manipulated explanation, and the divergence of data distribution. The algorithm's parameters can be tuned to improve the convergence and emphasize a given performance measure to better adapt to a specific scenario.

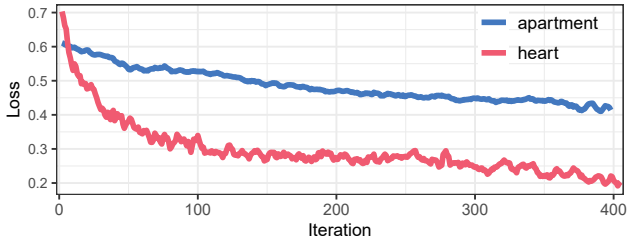

Figure 3: Convergence of the manipulation algorithm.

Measure \ Scenario	`heart`	`apartment`
L_1 norm	0.272	289 185
Kendall tau distance	0.6	0.86
Jensen–Shannon divergence	0.036	0.049

Table 1: Performance of the manipulation algorithm.

Conclusion

We deem it is necessary to carefully evaluate explanations in the context of reference data distribution. The introduced genetic-based algorithm for manipulating SHAP is useful as a sanity check for providing reliable interpretations of black-box models. We next intend to incorporate the on-manifold integer data perturbations to consider scenarios like COMPAS, Adult and German Credit. Code for this work is available at https://github.com/hbaniecki/manipulating-shap.

Acknowledgments

Work funded by NCN SONATA BIS 2019/34/E/ST6/00052.

References

Aivodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; and Tapp, A. 2019. Fairwashing: the risk of rationalization. In *ICML'19*.

Baniecki, H.; Kretowicz, W.; and Biecek, P. 2021. Fooling Partial Dependence via Data Poisoning. *arXiv preprint arXiv:2105.12837*.

Ghorbani, A.; Abid, A.; and Zou, J. 2019. Interpretation of Neural Networks Is Fragile. In *AAAI'19*.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; et al. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *NeurIPS'17*.

Mishra, S.; Dutta, S.; Long, J.; and Magazzeni, D. 2021. A Survey on the Robustness of Feature Importance and Counterfactual Explanations. *Workshop on Explainable AI in Finance*.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Wortman Vaughan, J. W.; and Wallach, H. 2021. Manipulating and Measuring Model Interpretability. In *CHI'21*.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *AIES'20*.