Explainable Machine Learning for Fairness: PDPs to Causal Dependence Plots

Joshua Loftus [1] Sakina Hansen [1] Lucius Bynum [2]

[1]Department of Statistics, London School of Economics and Political Science [2]Center for Data Science, New York University

Abstract

Machine learning and AI models are increasingly used to make impactful decisions about people, from lending to crime to education. Explaining the output and processes of these models is essential. For example, explanations may be needed to understand whether the model satisfies regulations about discrimination or privacy. Methods for model explanation/interpretation can be used to audit such models and inform or educate stakeholders, and potentially also to learn about relationships in the world that generated the training data. Audits often do not have access directly to the algorithm and so rely on model agnostic methods, such as the partial dependence plot (PDP). We investigate the efficacy of PDPs for fairness audits, followed by our development of causal dependence plots, addressing some of the limitations of PDPs.

1. Background

- Explainability: generating explanations that can help users and other stakeholders understand how an algorithm has come to a decision.
- Algorithmic Fairness: different individuals receiving different treatment from an algorithm based on features that have been systemically discriminated against.
- Audits: systematic reviews of algorithms, often having a particularly focus in mind, e.g. reviewing if an algorithm performs fairly.
- Partial Dependence Plot: A partial dependence plot (PDP) displays the marginal effect of a feature on the predicted outcome of a machine learning model, marginalized over a dataset. Where f is the predictor function, the PD function is defined as $\hat{f}_s(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)$, where x_S are the features for which the partial dependence function is plotted for [1].

Causal Modelling

We follow causal modelling as defined by [2].

Structural Causal Model (SCM). A (probabilistic) SCM $\mathcal{M}$ is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P_{\mathbf{U}} \rangle$ where $P_{\mathbf{U}}$ is the joint distribution of the exogeneous variables.

Interventions. For the SCM $\mathcal{M}$ an intervention I produces a modified SCM denoted $\mathcal{M}^{\text{do}(I)}$ which may have different structural equations $\mathbf{F}^I$. The DAG representation $\mathcal{G}^{\text{do}(I)}$ may also change, and the new interventional distribution is $P^{\mathcal{M};\text{do}(I)}$.

Counterfactuals. A counterfactual distribution is an interventional distribution defined over a specific dataset that lets us answer 'what if questions'. For variable V_j with observed values of its parents $\mathbf{PA}_j = v$, we may hold some or all of v fixed and vary $U_j := u$, passing these through $f_j(v, u)$. The counterfactuals $V_j(\tilde{v}, u)$ are values V_j would have taken if any of its observed and/or exogeneous parents had taken the different values $(\tilde{v}, u)$.

Individual counterfactuals and expected effects. Denote $\hat{f}(P^{\mathcal{M}})$ to denote using data from $\mathcal{M}$ as the input the the black-box function $\hat{f}$. Individual counterfactual curves are$(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)})$ with empirical average over the explanatory dataset: $\hat{E}\left[(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)})\right]$.

2. Analysis of PDP

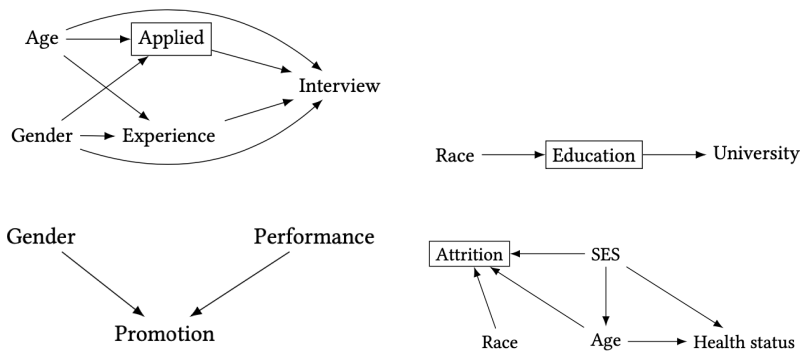

Figure 1. DAGs helpful for explaining PDP limitations for fairness. Left to right: Data Dependence, Unfairness via Mediators, Interaction, Attrition

Data Dependence: Unrepresentative data could lead to inaccurate conclusions about the world.

Unfairness via Mediators (Proxy Discrimination): If a black-box does not take a sensitive attribute as an input it can still perform proxy discrimination

Interaction: PDPs are most effective at showing model dependence on each predictor if the model is additive, but can hide dependence if there are interactions. This strong dependence on model structure complicates the interpretation of PDPs. However, an ICE plot may alleviate this.

Attrition and Counterfactual Fairness: Attrition by age has been studied related to unfairness to in law and health , where age interacts with other factors. In such examples, attrition can violate the backdoor criterion [3].

3. PDP Experiment: Hiring Interaction Simulation

Based on the first DAG in Figure 1.

- $\hat{f}_E$: only experience as a predictor.
- $\hat{f}_{EAG}$: experience, age, and gender, but as linear effects.
- $\hat{f}_{\text{int}}$: experience, age, and gender as predictors with interaction effects (correctly specified).

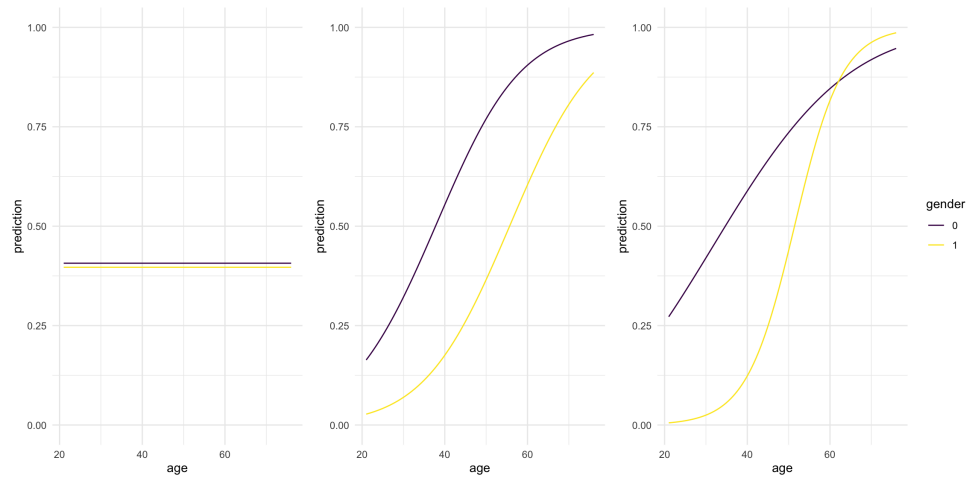

Figure 2. Conditional PDPs of age conditional on gender. The left plot shows a PDP of $\hat{f}_E$, the center plot $\hat{f}_{EAG}$, and the right plot $\hat{f}_{\text{int}}$.

4. Causal Dependence Plots

- Total Dependence Plot (TDP): The empirical average of $\text{TE}(I) = (P^{\mathcal{M}_X|=x;\text{do}(I)})$ estimates the total effect
- Partially Controlled Dependence Plot (PCDP): The empirical average of $\text{PCE}(I) = (P^{\mathcal{M}_X|=x;(I,C)})$, estimates a partially conditioned effect, where atomic intervention C holds constant other variables.
- Natural Direct Dependence Plot (NDDP): The empirical average of $\text{NDE}(I) = (P^{\mathcal{M}_X|=x;(I,J)})$ estimates the natural direct dependence where J is an intervention on all children of nodes that are intervened upon by I. **PDPs show this effect.**
- Natural Indirect Dependence Plot (NIDP): The empirical average of $\text{NIE}(I) = (P^{\mathcal{M}_X^{do(I)}|=x;(K)})$ estimates the natural indirect effect, where K removes outgoing edges from any nodes intervened on by I.

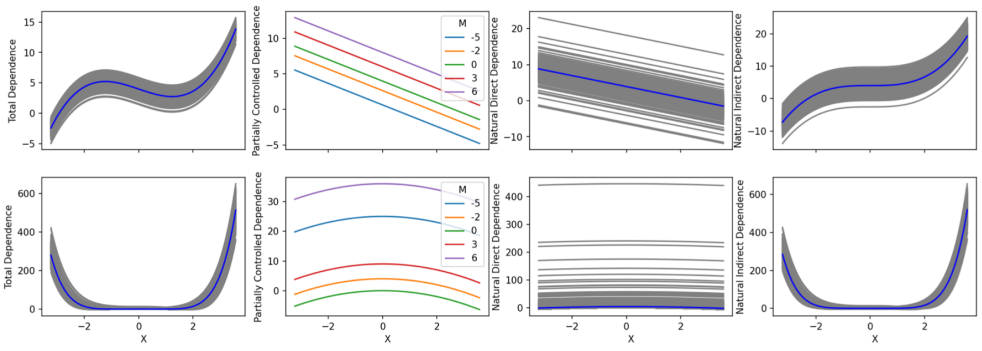

Figure 3. TDP, PCDP, NDDP, NIDP for a correctly specified black box function (bottom row), and an incorrectly specified black box function (top row), for the data: $X \sim \mathcal{N}(0, 1)$, Mediator $M = \frac{1}{2}X^3 + \mathcal{N}(0, 1)$, Outcome $Y = M^2 - \frac{1}{2}X^2 + \mathcal{N}(0, 1)$

5. Summary

- PDPs contain important limitations for auditing for fairness
- Incorporating causal knowledge can help alleviate the problems with PDP
- Causal dependence plots have other applications: scientific discovery with causal discovery methods, explanations under covariate shift, causal semi-supervised learning
- **Limitation**: Other model-agnostic methods not explored: SHAP and LIME explanations
- **Limitation**: Requires some causal knowledge about the ground truth data to be meaningful

References

[1] Christoph Molnar.
Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
https://christophm.github.io/interpretable-ml-book/, 2022.

[2] Judea Pearl et al.
Models, reasoning and inference.
Cambridge, UK: CambridgeUniversityPress, 19(2), 2000.

[3] Qingyuan Zhao and Trevor Hastie.
Causal interpretations of black-box models.
Journal of Business & Economic Statistics, 39(1):272–281, 2021.