Causal Prediction Can Induce Performative Stability

Anonymous Authors[1]

Abstract

Predictive models affect the world through inducing a strategic response or reshaping the environment in which they are deployed—a property called performativity. This results in the need to constantly adapt and re-design the model. We show that prediction using only causal features—those that directly affect the prediction target, and not those that are otherwise correlated to the target—can achieve an equilibrium and close this feedback loop. Thus, a causal predictive model does not require any further adaptation after deployment even if it change its environment.

1. Introduction

Predictive models are often deployed from a position of power, e.g., to allocate a scarce resource, as in credit scoring or in welfare fraud prediction, or to issue other highly consequential high-stakes decisions, as in recidivism prediction.

In these scenarios, the models cause a change of their environment. This could be due to strategic responses of the population subjected to the model-mediated outcomes (Hardt et al., 2016), in which individuals attempt to game the algorithmic system, or are forced to adapt to its requirements (Miller et al., 2020). Other than this, the model's outcomes could change the environment in other ways such as changing observable features of individuals as a result of decisions such as denying bail. Modeling in these circumstances has been recently studied and formalized under the name of performative prediction (Perdomo et al., 2020).

Prediction using causal features—those that causally affect the prediction target—as opposed to features that are merely correlated with the target is known to have beneficial properties, such as robustness to certain distribution shifts (Rojas-Carulla et al., 2018), and to spurious changes to features (Veitch et al., 2021). In this short paper we show that causal prediction also can help avoid the feedback loops due to model performativity.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

2. Causal vs. Associative Prediction

We consider a prediction task in which we want to predict values of a real-valued random *target variable* Y with range $\mathbb{Y}$ from a set of *covariate variables* X (features) with joint range $\mathbb{X}$. A *predictive model* $g(x_S)$ uses a subset of covariates $X_S \subseteq X$ to output a prediction, with $x_S \in \mathsf{range}(X_S)$ being a realization of variables in X_S. We assume there exists a structural causal model (SCM) describing the relationships between X and Y (Pearl, 2009). We denote by $pa(T)$ the set of causal parents of a variable T in the SCM.

We say that a set of features X_C is *causal* if they are a subset of causal parents of the target variable: $X_C \subseteq pa(Y)$. Otherwise, we say that features are *associative*, or *non-causal*. Causal prediction and respective predictive models are those that use causal features.

3. Performativity and Causal Prediction

The concept of performative prediction (Perdomo et al., 2020) recognizes and formalizes the fact that predictive models affect the world once deployed. Within the probabilistic framework, this can be encoded as a distribution shift which occurs after the model's deployment. Such shifts are undesirable as they result in a constant drift away from the original distribution, turning the model that initially could have performed well into a suboptimal one. It requires the model to constantly adapt to its ever-changing environment.

A way to account for the future shift is strategic training:

$$\min_{g:\mathbb{X}\to\mathbb{Y}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{g(\cdot)}}[\ell(g(x),y)], \tag{1}$$

where $\mathcal{D}_{g(\cdot)}$ is the data distribution of (X,Y) induced by the deployment of a model $g(\cdot)$ into its environment, and $\ell(\hat{y},y)$ is a loss function. The optimal model from Equation (1) is called *strategically optimal*.

Such optimality, however, in general does not stop the feedback loop of needing to adapt the model to induced performative shifts. To do so, the model needs to achieve *performative stability*:

$$\min_{g:\mathbb{X}\to\mathbb{Y}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{g(\cdot)}}[\ell(g(x),y)], \tag{2}$$

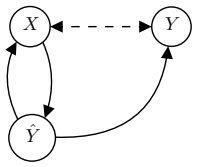

Figure 1. General cyclic structural causal model of performative prediction. Predictions $\hat{Y} = f(X)$ affect the data variables X, Y.

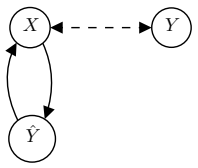

Figure 2. Cyclic structural causal model of performative prediction under Assumption 3.2. Predictions $\hat{Y} = f(X)$ affect only the covariates X.

A performatively stable predictive model is best-performing even after the shift its deployment induces.

3.1. Causal View of Performative Prediction

To formalize the setting of performative prediction within the framework of causality, we extend the SCM in Section 2 to include the predictions $\hat{Y} = f(X)$. We assume that $\hat{Y}$ can performatively affect X and Y after deployment, inducing the observed distribution shift. We model this effect using a cyclic connection $\hat{Y} \leftrightarrow X$ (Bongers et al., 2021), and $\hat{Y} \rightarrow Y$. See Figure 1 for the graph.

3.2. Causal Predictive Models Need not Cause Performative Shifts

We show that under mild assumptions, causal prediction avoids the issues of performativity by inducing a performatively stable algorithm. In particular, we make the following assumptions about the setting:

Assumption 3.1 (Faithfulness (Bongers et al., 2021))**.** The conditional independencies of the joint distribution of variables are expressed in the graph.

Assumption 3.2. The performativity shift directly affects only the covariates: $\hat{Y} \rightarrow X$, yet $\hat{Y} \not\rightarrow Y$. See Figure 2 for a graph that satisfies the assumption.

The Assumption 3.1 is a standard assumption for causal graphs that appropriately model the probability distribution. Let us see a case of a performative setting in which Assumption 3.2 holds. For example, in the setting of strategic classification (Hardt et al., 2016), the performativity shift occurs through individuals strategically adapting their features in response to the previous classification outcomes. This can be modeled as a response function $\Delta(x, f)$, so that the covariates after the responses satisfy $X = \Delta(X, \hat{Y})$. In this formalization, the predictions $\hat{Y}$ do not directly impact Y, only indirectly through the changed distribution of X.

Next, we show that under these assumptions, an optimal causal predictive model achieves performative stability using an example of regression:

Theorem 3.3. *Suppose that Assumption 3.1 and Assumption 3.2 hold. Moreover, suppose the following conditions hold about the modeling setup:*

1. *The variables X, Y are continuous with range $\mathbb{Y} \subseteq \mathbb{R}$, and $\mathbb{X} \subseteq \mathbb{R}^{|X|}$.*
2. *The loss function is squared loss: $\ell(\hat{y}, y) \triangleq (\hat{y} - y)^2$.*

Then the causal predictive model $g^*(x_C) \triangleq \mathbb{E}[Y \mid X_C = x_C]$ achieves performative stability:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{g^*(\cdot)}}[(g^*(x_C) - y)^2]. \tag{3}$$

Proof. We want to show that for any subset $X_N \subseteq X \setminus X_C$ of associative features, any other model $g'(x_C, x_N)$ incurs a higher performative loss:

$$\mathbb{E}_{X,Y}[(g'(X_C, X_N) - Y)^2] \geq \mathbb{E}_{X,Y}[(g^*(X_C) - Y)^2], \tag{4}$$

where (X, Y) are distributed according to the performative distribution $\mathcal{D}_{g^*(\cdot)}$.

Note that by the assumptions, we have $X_C \rightarrow Y \rightarrow X_N$, as Y is not directly affected by $\hat{Y}$ and so by X_N. Thus, $P(X_C, X_C, Y)$ decomposes as $P(Y \mid X_C)P(X_C)P(X_N)$.

The performative loss of $g'(\cdot)$ is equal to:

$$\begin{aligned} &\mathbb{E}_{X_C,X_N,Y}[(g'(X_C, X_N) - Y)^2] = \\ &= \int \underbrace{\int (g'(x_C, x_N) - y)^2 dP(x_C, y)}_{\text{Standard loss of } g'(\cdot) \text{ over } x_C, y} dP(x_N). \end{aligned} \tag{5}$$

The inner integral is the standard loss of $g'(x_C, x_N)$ as a function of x_C. Because $g^*(x_C)$ is Bayes-optimal with respect to squared loss over the distribution of X_C, Y, it incurs lower loss:

$$\mathbb{E}_{X_C,Y}(g'(X_C, X_N) - Y)^2] \geq \mathbb{E}_{X_C,Y}[(g^*(X_C) - Y)^2]$$

Combining this with Equation (5), we get the sought inequality in Equation (4). □

4. Conclusions

In this short paper, we provided a formal argument that causal prediction can induce performative stability: remove the need to adapt to those changes of the environment which

are due to the model itself. At the workshop, we hope to discuss the formal premise and the causal formalization of the performative prediction (e.g., other modeling options such as through interventions), other possible results of the causal view of the performative prediction (e.g., change of the causal structure due to performativity), and broader implications of the result.

References

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

Pearl, J. *Causality*. Cambridge university press, 2009.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.