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 formalize a setting in which performativity arises using the language of causal modeling. We show that an equilibrium which closes the feedback loop is achieved when models use features that form a Markov blanket of the target variable. Thus, a predictive model that takes the causal features does not require any further adaptation after deployment even if it changes the 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). Apart from 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 which form a Markov blankets of 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 causality-aware 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. Background

2.1. Prediction in Causal Models

We consider a prediction task in which we want to predict values of a real-valued random *target variable* Y with range $\mathbb{Y} \subseteq \mathbb{R}$ from a set of m *covariate (feature) variables* X with joint range $\mathbb{X} \subseteq \mathbb{R}^m$. A *predictive model* $g(x_I)$ uses a subset of covariates $X_I \subseteq X$ to output a prediction, with $x_I \in \mathbb{X}_I = \mathbb{R}^{|X_I|}$ being a realization of variables in X_I.

We assume there exists a structural causal model (SCM) describing the relationships between X and Y (Pearl, 2009). We say that a set of features X_S is *Markov* if they form the Markov boundary of the target variable Y, which we define as the set of Y's parents, children, and the children's parents (spouses). We refer to, e.g., (Pearl, 2009) for a detailed treatment of the topic.

2.2. Classical Optimal Prediction

Let us introduce some notation and definitions regarding predictive models which are optimal in standard, non-performative settings.

In a non-performative setting we minimize the standard expected loss over some distribution $\mathcal{D}$:

$$R(g) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(g(x), y)], \tag{1}$$

where $\ell(\hat{y}, y) \geq 0$ is a loss function. A minimizer of the loss is called a *Bayes-optimal* model:

$$g^{\mathsf{bayes}} \in \arg\min_{g\colon \mathbb{X}\to\mathbb{Y}} R(g)$$

For the square loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, it is well-known that the Bayes-optimal model is the conditional expectation:

$$g^{\mathsf{bayes}}(x) = \mathbb{E}[Y \mid X = x].$$

For the 0-1 loss function $\ell(\hat{y}, y) = \mathbf{1}[\hat{y} = y]$, the optimal model is the maximum a posteriori rule:

$$g^{\mathsf{bayes}}(x) = \arg\max_{y\in\mathbb{Y}} \Pr[Y = y \mid X = x].$$

Prediction from Different Feature Sets We also consider optimal prediction of the target variable based on a given subset of features $X_I \subseteq X$:

$$R_{X_I}(g) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(g(x_I), y)]$$

with the respective Bayes-optimal model:

$$g^{\text{bayes}}_{X_I} \in \arg\min_{g:\, \mathbb{X}_I \to \mathbb{Y}} R_{X_I}(g)$$

2.3. Performative 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, in which a model minimizes performative loss:

$$PR(g) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(g(x), y)], \tag{2}$$

where $\mathcal{D}_g$ is the data distribution of (X, Y) induced by the deployment of a model g into its environment. The optimal model which minimizes **??** is called *performatively optimal*:

$$g^{\text{po}} \in \arg\min_{g:\, \mathbb{X} \to \mathbb{Y}} PR(g)$$

Performative optimality in general does not stop the feedback loop of needing to adapt the model to induced performative shifts. To do so, a model g^{ps} needs to satisfy the property called *performative stability*:

$$g^{\text{ps}} \in \arg\min_{g:\, \mathbb{X} \to \mathbb{Y}} PR(g; g^{\text{ps}}) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}_{g^{\text{ps}}}}[\ell(g(x), y)] \tag{3}$$

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

3. Performative Stability in the Soft-Intervention Model

To formalize the setting of performativity within the framework of causality, we extend the SCM in Section 2.1 by introducing a new root random variable M which represents the deployed model. The effects of the variable M on the data variables X and Y reflect the mechanism through which the predictive model affects the data distribution, as illustrated in Figure 1. We represent the deployment of a given predictive model as an intervention (Pearl, 2009) on the M variable. Thus, the performative distribution $\mathcal{D}_g$ is the distribution of (X, Y) under an intervention $M := g$, which is simply $P(X, Y \mid M = g)$ as M is a root variable. This is also known as a "soft intervention" on (X, Y) (Eberhardt & Scheines, 2007). We use a special value $M = \varnothing$ to denote the initial data distribution $\mathcal{D}_\varnothing = P(X, Y \mid M = \varnothing)$ prior to any deployment.

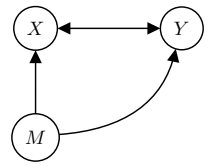

Figure 1. General structural causal model of performative prediction. The deployment of a predictive model M affects the distribution of data variables X, Y.

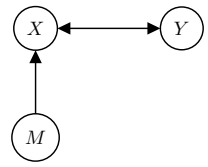

Figure 2. Structural causal model of performative prediction under Assumption 3.2. The deployment of a predictive model M affects only the covariates X.

We make the following assumptions about the setting:

Assumption 3.1 (Markov and Faithfulness (Pearl, 2009))**.** The conditional independencies of the joint distribution of variables are expressed in the graph, and vice versa.

Assumption 3.2. The performativity shift directly affects only the covariates: $M \to X$, yet $M \not\to Y$. Formally, the set of covariates X d-separate Y from M (Pearl, 2009). 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. Assumption 3.2 is equivalent to a covariate shift caused by the deployment of the model (Bühlmann, 2020). Unlike the standard setting of covariate shift, however, performative shift can be controlled by the model's deployer.

Next, we show that under these assumptions, surprisingly, performative stability can be achieved without any strategic training by using the Markov features.

Theorem 3.3. *Suppose that Assumption 3.1 and Assumption 3.2 hold. Then, in the case of square loss and 0-1 loss, the non-strategically optimal model which uses the Markov features X_S achieves performative stability across all models over the full set of covariates X:*

$$g^{\text{bayes}}_{X_S} \in \arg\min_{g:\, \mathbb{X} \to \mathbb{Y}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{g^{\text{bayes}}_{X_S}}}[\ell(g(x), y)] \tag{4}$$

Proof. Let us denote as $X_N = X \setminus X_S$ the set of all non-

Markov features. By Assumption 3.1, the Markov features X_S provide the conditional independence guarantee:

$$P[Y \mid X_S, X_N] = P[Y \mid X_S].$$

Moreover, from Assumption 3.2 we additionally have:

$$P[Y \mid X, M] = P[Y \mid X].$$

Therefore, the conditional distribution of $Y \mid X$ is constant for any performative shift $\mathcal{D}_g$:

$$P[Y \mid X, M = \varnothing] = P[Y \mid X, M = g^*(\cdot)] = P[Y \mid X]$$

Combining these, we have:

$$P[Y \mid X_S, X_N, M] = P[Y \mid X_S]. \tag{5}$$

Let us consider the case of the square loss. Then, the optimal non-performative model is $g^{\mathsf{bayes}}_{X_S}(x) = \mathbb{E}[Y \mid X_S = x_S]$. Observe that one performatively stable model, which minimizes $PR(\,\cdot\,;\, g^{\mathsf{bayes}}_{X_S})$, is the Bayes-optimal model over the distribution induced by $g^{\mathsf{bayes}}_{X_S}$. We know its closed form: it is the conditional expectation $\mathbb{E}[Y \mid X = x, M = g^{\mathsf{bayes}}_{X_S}]$. By Equation (5), it equals $g^{\mathsf{bayes}}_{X_S}(x)$.

An analogous argument holds for 0-1 loss. □

4. Conclusions

In this short paper, we provided a formal argument that causal prediction—using features on the Markov boundary of the target—can induce performative stability: remove the need to adapt to those changes of the environment which are due to the model itself under the assumption that performative shifts can be formalized as soft interventions on the data distribution. Future work could investigate other formalizations of performativity such as individual-level hard interventions, more appropriate to strategic classification settings (Miller et al., 2020).

References

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

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.