Evaluating and Correcting Performative Effects of Decision Support Systems via Causal Domain Shift

University of Amsterdam

Booking.com

Philip Boeken   Onno Zoeter   Joris M. Mooij

Summary

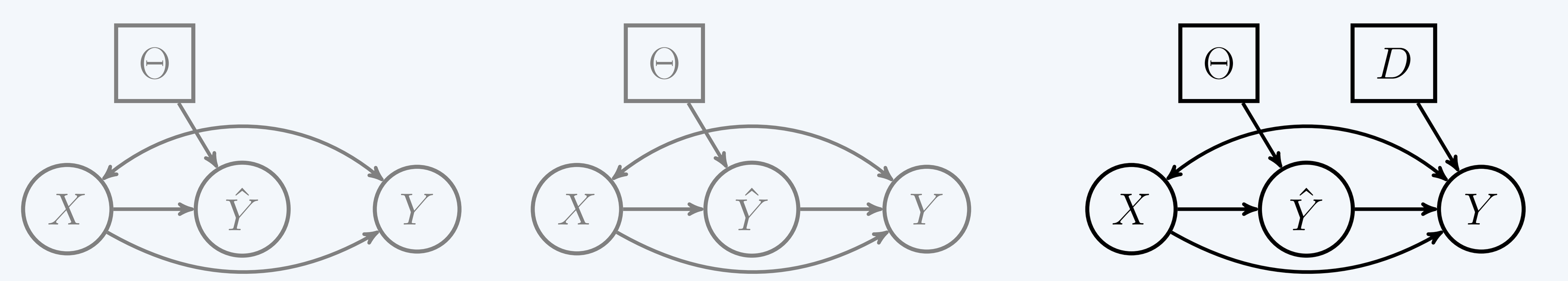
1. We model the deployment of a ***Decision Support System*** (DSS) with a causal model
2. which we use in two applications:
   A) **Evaluation**: we define the ***Deployment effect*** and ***Retraining effect*** (Def. 1, 2) as metrics to evaluate the effect of the deployment of a DSS.
   B) **Bias correction**: we specify a *baseline predictor* as suitable prediction model for the DSS, which corrects for ***performative bias*** (Def. 4) caused by a previous deployment of the DSS.

   Estimating these quantities constitutes three domain adaptation tasks (T1, T2, T3).
3. These tasks (T1, T2, T3) reduce to ***a single domain adaptation problem*** (Lemma 1), which cannot be solved without imposing extra assumptions (Prop. 1).
4. Our proposed solution is to consider a ***domain pivot*** (Def. 5) which facilitates domain adaptation (Prop. 2).

Causal Model of a Decision Support System

- Features X
- Outcome variable Y
- Prediction $\hat{Y}$ using X and parameters Θ
- Deployment indicator D

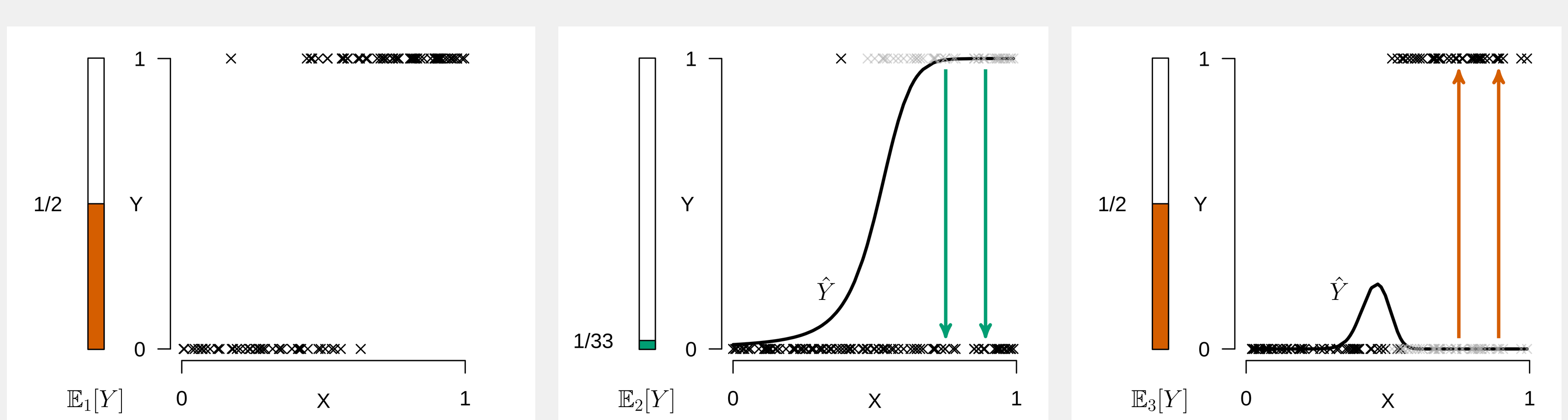
- We measure i.i.d. data from $\mathbb{P}(X, \hat{Y}, Y \mid \mathrm{do}(D, \Theta))$



(a) Graph for context $D = 0$ (b) Graph for context $D = 1$ (c) Graph G with indicator D

Performative prediction

- A prediction $\hat{Y}$ of Y is called *performative* if it affects Y.
- A numerical example: $Y = 1$ is to be prevented, $\hat{Y} = \mathbb{E}[Y \mid X]$ is a prediction of risk, and $\hat{Y} > 1/2$ instigates an action that effectively reduces the observed risk.



(a) $t = 1$, DSS off (b) $t = 2, \hat{Y} = \mathbb{E}_1(Y \mid X)$ (c) $t = 3, \hat{Y} = \mathbb{E}_2(Y \mid X)$

Application A: Evaluation (T1, T2)

1. a) A DSS with model parameters θ is proposed. Should it be deployed?
   b) A DSS with model parameters θ is in use. Should we switch it off?

   Definition 1 (Deployment effect)
   We define the *deployment effect* of a DSS with parameters θ as the average causal effect of the deployment of the DSS on the target variable, i.e.
   $$\tau(\theta) := \mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta)] - \mathbb{E}[Y \mid \mathrm{do}(D = 0)]. \tag{1}$$

2. A new model with parameters θ_{t+1} is proposed. Must they replace current parameters θ_t?

   Definition 2 (Retraining effect)
   We define the *retraining effect* as the average causal effect of the deployment of a retrained DSS on the target variable, i.e.
   $$\rho(\theta_{t+1}, \theta_t) := \mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta_{t+1})] - \mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta_t)]. \tag{2}$$

	Metric	Source domain	Target domain	Target quantity
T1.a	$\tau(\theta)$	$D = 0$	$D = 1, \Theta = \theta$	$\mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta)]$
T1.b	$\tau(\theta)$	$D = 1, \Theta = \theta$	$D = 0$	$\mathbb{E}[Y \mid \mathrm{do}(D = 0)]$
T2	$\rho(\theta_{t+1}, \theta_t)$	$D = 1, \Theta = \theta_t$	$D = 1, \Theta = \theta_{t+1}$	$\mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta_{t+1})]$

Table 1. Domain adaptation tasks for evaluation.

Application B: Bias correction (T3)

3. Let Y be an outcome whose expected value we want to minimize (e.g. a cost, negative utility/reward, etc.), and let $\hat{Y}$ be a prediction that can instigate an action that reduces the expected outcome below a known level. **A naively retrained model of $\hat{Y} = \mathbb{E}[Y \mid X]$ will underestimate the risk if the previous model was effective.**

   In certain settings the *baseline predictor*
   $$\hat{Y} := \mathbb{E}[Y \mid X, \mathrm{do}(D = 0)] \tag{3}$$
   is the optimal prediction model for preventing $Y = 1$.

   Definition 4 (Performative bias)
   When gathering data from the domain $D = 1, \Theta = \theta$, naive retraining will estimate $\mathbb{E}[Y \mid X, \mathrm{do}(D = 1, \Theta = \theta)]$ instead of $\mathbb{E}[Y \mid X, \mathrm{do}(D = 0)]$, yielding a *performative bias*:
   $$\mathbb{E}[Y \mid X, \mathrm{do}(D = 1, \Theta = \theta)] - \mathbb{E}[Y \mid X, \mathrm{do}(D = 0)]. \tag{4}$$

	Source domain	Target domain	Target quantity
T3	$D = 1, \Theta = \theta$	$D = 0$	$\mathbb{E}[Y \mid X, \mathrm{do}(D = 0)]$

Table 2. The domain adaptation task for performative bias correction.

Equivalence of T1–3, and non-identifiability

Lemma 1
Identifiability of the target quantities of the domain adaptation tasks T1, T2, T3 is equivalent to identifiability of the conditional expectation $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ from $\mathbb{P}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$ for $(d, \theta) \neq (d', \theta')$.

Proposition 1
In the class of SCMs with graph G, the target quantity $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is not identifiable from $\mathbb{P}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$ for $(d, \theta) \neq (d', \theta')$.

Problem: In high-stakes settings, performing an RCT (and thus measuring labels Y in the target domain) can be undesirable.

Solution: measure mediators of prediction and outcome

Definition 5 (Domain pivot)
A *domain pivot* for target variable Y and domain indicator (D, Θ) is a set of variables $\{X, Z\}$ such that $Y \perp\!\!\!\perp D, \Theta \mid X, Z$.

Consider the graph G' below. For solving tasks T1–T3, **we require measurements of the domain pivot $\{X, A, C\}$ with mediator A and confounder C in both the source- and target domain.** The outcome Y does *not* have to be measured in the target domain.

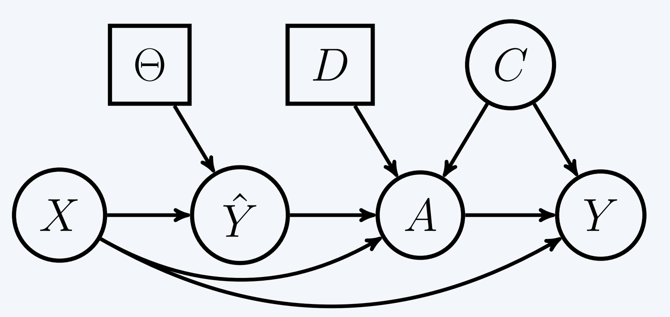


Figure 3. Graph G' with action A and confounder C, with $\{X, A, C\}$ as *domain pivot*.

Practical implementation: show the prediction $\hat{Y}$ to an agent, let them report their decision A and information C that influences this, but let another agent carry out their action of choice without having seen $\hat{Y}$.

Proposition 2
Under positivity assumptions, the target quantity $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is identifiable from
$$\{\ \mathbb{P}(X, Z, Y \mid \mathrm{do}(D = d', \Theta = \theta')),\ \ \mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta))\ \}$$
iff $Y \perp\!\!\!\perp D, \Theta \mid X, Z$, in which case
$$\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = \mathbb{E}[\ \mathbb{E}[Y \mid X, Z, \mathrm{do}(D = d', \Theta = \theta')]\ \mid X, \mathrm{do}(D = d, \Theta = \theta)].$$

Additional results in the paper:
- identifiability results when the data is subject to selection bias;
- the estimation of these quantities.

p.a.boeken@uva.nl