Robust Causal Domain Adaptation in a Simple Diagnostic Setting

Thijs van Ommen - Utrecht University, The Netherlands

Utrecht University

Overview

In statistics/machine learning, we want to learn from data, for example to make good decisions.

Domain adaptation is necessary when we have access to data from a source domain and want to make decisions in a target domain, while the data distributions are different in these two domains. In this paper, we look at the *proactive* version of this problem: we have no training data from the target domain.

Causal domain adaptation: the source and target distribution have something in common, namely a common causal graph (see right).

Robust causal domain adaptation: we make no assumptions about the part of the distribution that changes, but take a worst-case (robust Bayes) approach.

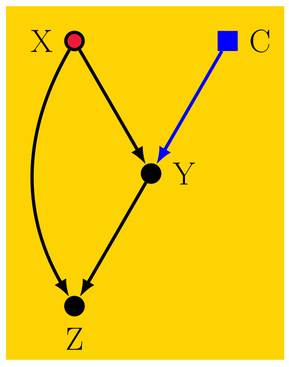


Figure 1: Causal graph for our 'simple diagnostic setting.' X is to be predicted given Y and Z, and C is a context variable which distinguishes the source from the target domain. All our data come from the source domain ($C = 0$), but we want to predict in the target domain ($C = 1$).

Motivating example

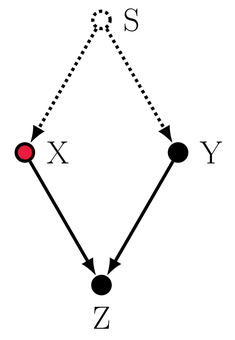
(from [SSS19])

- X: lung cancer — to be diagnosed
- S: smoking (unobserved variable)
- Y: aspirin — may be prescribed to smokers due to their risk of heart disease
- Z: chest pain

This model is trained in one hospital, where it is used succesfully. But in another hospital,

- the policy for prescribing aspirin Y to smokers S may be different
- the other probability tables remain the same
- no data is available

The causal graph we study in this work (Figure 1) is simplified to have no unobserved variables, but is equivalent for our purposes.



Numerical illustration

Example: all variables are binary, and we observe in the source domain that

$$P(X = 1) = \tfrac{1}{2};$$
$$P(Z = 1 \mid Y, X) = \begin{cases} \frac{1}{2} & \text{if } Y = X; \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

Introducing parameters α_0, α_1 to define the unknown $P(Y = x \mid X = x, C = 1) = \alpha_x$, we can compute expressions for any joint distribution P as given in Table 1.

Table 1: General form of $P(X, Y, Z \mid C = 1)$ consistent with (2).

$P(X, Y, Z \mid C = 1)$	$X = 0$	$X = 1$
$Y = 0,\ Z = 0$	$\alpha_0/4$	0
$Y = 0,\ Z = 1$	$\alpha_0/4$	$\frac{1}{2} - \alpha_1/2$
$Y = 1,\ Z = 0$	0	$\alpha_1/4$
$Y = 1,\ Z = 1$	$\frac{1}{2} - \alpha_0/2$	$\alpha_1/4$

Applying Theorem 1 to this example, we can compute the worst-case P^* for different loss functions (see Tables 2 and 3). This gives the decision maker the information they need.

Note that

- the optimal decisions depend on the loss function (even for two strictly proper scoring rules)
- finding such tables analytically becomes difficult quickly, but the answers can be found using convex optimization of (1)

Table 2: $P^*_{\text{Brier}}(X, Y, Z)$ optimal under Brier loss ($\alpha_0 = \alpha_1 = 2 - \sqrt{2}$). Also shown is $P^*_{\text{Brier}}(X \mid Y, Z, C = 1)$, the distribution that the decision maker is actually interested in.

$P^*_{\text{Brier}}(X, Y, Z \mid C = 1)$	$X = 0$	$X = 1$
$Y = 0,\ Z = 0$	0.146	0
$Y = 0,\ Z = 1$	0.146	0.207
$Y = 1,\ Z = 0$	0	0.146
$Y = 1,\ Z = 1$	0.207	0.146

$P^*_{\text{Brier}}(X \mid Y, Z, C = 1)$	$X = 0$	$X = 1$
$Y = 0,\ Z = 0$	1	0
$Y = 0,\ Z = 1$	0.414	0.586
$Y = 1,\ Z = 0$	0	1
$Y = 1,\ Z = 1$	0.586	0.414

Robust approach: Decision maker vs. adversary

Protocol

1. Simultaneously,
   - adversary chooses a distribution $P(X, Y, Z \mid C = 1)$ ($C = 1$ representing the target domain) that is consistent with the invariant parts of the source domain (see 'Numerical illustration' for an example)
   - decision maker chooses strategy $A : \mathcal{Y} \times \mathcal{Z} \to \mathcal{A}$
2. x, y, z are drawn from $P(X, Y, Z \mid C = 1)$
3. the decision maker suffers loss $L(x, A(y, z))$

The decision maker wants to minimize the expected loss, while the adversary works to maximize it.

Loss functions Two possible loss functions are: (both defined over action space $\mathcal{A} = \Delta_X$, the set of all distributions over $\mathcal{X}$)

$$L_{\text{Brier}}(x, Q) = \sum_{x' \in \mathcal{X}} \left(\mathbf{1}_{x'=x} - Q(x')\right)^2;$$
$$L_{\log}(x, Q) = -\log Q(x).$$

For any L, we can define the generalized entropy H_L as the expected loss when the decision maker, knowing P, chooses Q optimally:

$$H_L(P) := \inf_{Q \in \Delta_{\mathcal{X}}} \sum_x P(x) L(x, Q).$$

Adversary's perspective If the adversary chooses P and the decision maker knows P, the expected loss is

$$\sum_{\substack{(y,z) \in \mathcal{Y} \times \mathcal{Z}: \\ P(y,z) > 0}} P(y, z) H_L(P(\cdot \mid y, z)). \tag{1}$$

In a similar situation, [**vO**KFG16] show that for many L, the game has a Nash equilibrium, and an optimal strategy for the decision maker can easily be determined once the P maximizing (1) is known.

Theorem 1 (Existence and characterization of P^*)**.** *For H_L finite and continuous, a P maximizing (1) exists, and P^* is such a maximizer if and only if there exists a $\lambda^* \in \mathbf{R}^{\mathcal{X}}$ such that*

(i) for every $y \in \mathcal{Y}$ with $P^*(y) > 0$,

$$\sum_{\substack{z \in \mathcal{Z}: \\ P^*(y,z) > 0}} P^*(z \mid y) H_L(P^*(\cdot \mid y, z)) = \sum_x P^*(x \mid y) \lambda^*_x;$$

(ii) for every $y \in \mathcal{Y}$, for all $P' \in \Delta_{\mathcal{X}}$, let $P'(x, z \mid y) := P'(x) P(z \mid x, y)$, then

$$\sum_{\substack{z \in \mathcal{Z}: \\ P'(z \mid y) > 0}} P'(z \mid y) H_L(P'(\cdot \mid y, z)) \le \sum_x P'(x \mid y) \lambda^*_x.$$

Conclusion and future work

We have shown that it is possible (but not straightforward) to define robust decisions in a causal domain adaptation setting.

Topics for future work:

- The decision found by [SSS19] are not as 'robust' as ours (i.e. worst-case behaviour is worse), but their theory applies to many causal graphs. We would like to extend our theory to other graphs as well.
- [M**vO**CBVM18] consider causal domain adaptation where the causal graph is unknown. Their results might be improved using ideas from robust decision making.
- We used insights from [**vO**KFG16] to solve the numerical example, but did not yet show that their theory holds in our setting.

Table 3: $P^*_{\log}(X, Y, Z)$ optimal under logarithmic loss ($\alpha_0 = \alpha_1 = 1 - \frac{1}{5}\sqrt{5}$).

$P^*_{\log}(X, Y, Z \mid C = 1)$	$X = 0$	$X = 1$
$Y = 0,\ Z = 0$	0.138	0
$Y = 0,\ Z = 1$	0.138	0.224
$Y = 1,\ Z = 0$	0	0.138
$Y = 1,\ Z = 1$	0.224	0.138

$P^*_{\log}(X \mid Y, Z, C = 1)$	$X = 0$	$X = 1$
$Y = 0,\ Z = 0$	1	0
$Y = 0,\ Z = 1$	0.382	0.618
$Y = 1,\ Z = 0$	0	1
$Y = 1,\ Z = 1$	0.618	0.382

References

[M**vO**CBVM18] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NeurIPS*, 2018.

[SSS19] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *AISTATS*, 2019.

[**vO**KFG16] Thijs van Ommen, Wouter M. Koolen, Thijs E. Feenstra, and Peter D. Grünwald. Robust probability updating. *International Journal of Approximate Reasoning*, 74:30–57, 2016.