Causal Data Augmentation: Causality to serve Machine Learning

An introduction

Audrey Poinsot, Ph.D. student
audrey.poinsot@ekimetrics.com

Young Statisticians & Probabilists (YSP 12), January 2024

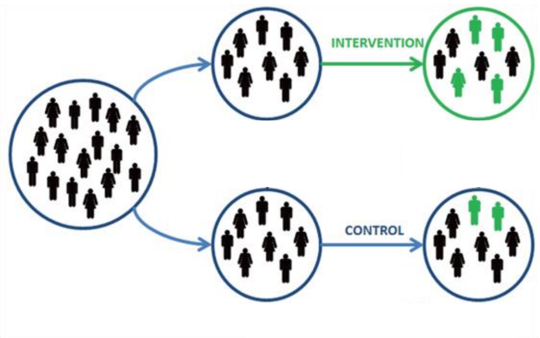
Context, from experimentation to observation

Experimental data → L_2

Randomized controlled trial

Clinical trial

A/B testing



Smoking? → Unfair

Major product? → Too expensive

Observational data → L_1

Statistics + Causal hypothesis

[Ibeling and Icard, 2020][1]

[Bareinboim et al. 2022][2]

L_2-quantities:

- Average Treatment Effect (ATE)
- Heterogeneous Treatment Effect (HTE)
- Conditional Average Treatment Effect (CATE)
- Individual Treatment Effect (ITE)

Ekimetrics.

[1] Duligur Ibeling and Thomas Icard Probabilistic reasoning across the causal hierarchy. In AAAI Conference on Artificial Intelligence, 2020.
[2] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's Hierarchy and the Foundations of Causal Inference. 2022.

Causality & Machine Learning

Machine Learning for Causality

Goal

Estimate causal quantities

ATE, ITE, CATE, Counterfactual, …

Causality role

Setting up a mathematical framework (i.e., hypotheses) under which correlation is causation

Machine Learning role

Estimate the “causal” correlations

↓

Double Machine Learning

Causal Machine Learning

Causality for Machine Learning

Goal

Improve Machine Learning models performances

Robustness, Generalization, Disentanglement, Data efficiency, …

Causality role

Incorporating causal knowledge (e.g., invariance rules, monotone effect, …) in the ML model

Machine Learning role

Performing a predictive task

↓

Causal regularization

Causal Data Augmentation

CausalDA, an approach to break down irrelevant correlations

Definition. DAG-constrained Causal Data Augmentation

Given:

- a set of variables $\mathbf{X} = (X_1, \ldots, X_d)$ distributed according to P_{obs},
- a DAG G encoding the causal dependencies that the variables must follow,
- a set of interventions I_{spl} applied to $\boldsymbol{X}_{\boldsymbol{int}} \subset \mathbf{X}$,

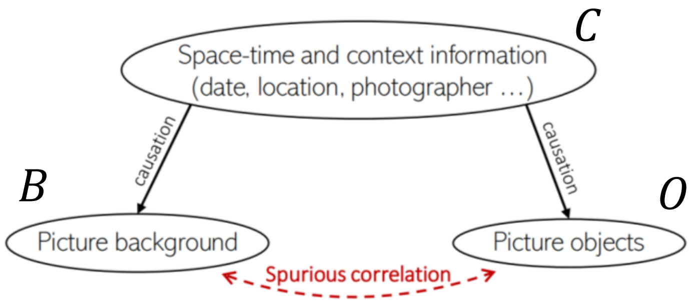
Causal Data Augmentation consists in sampling N data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G and the set of interventions $I_{spl} = \{P_{int}(X_k | PA_{int}(X_k)) \mid X_k \in \boldsymbol{X}_{\boldsymbol{int}}\}$.

$$P_{spl}(X_1, \ldots, X_d) = \underbrace{\prod_{X_i \notin \boldsymbol{X}_{\boldsymbol{int}}} P_{obs}(X_i | PA_G(X_i))}_{\text{In domain}} \underbrace{\prod_{X_k \in \boldsymbol{X}_{\boldsymbol{int}}} P_{int}(X_k | PA_{int}(X_k))}_{\text{Out-of-domain}}$$

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94



$B \not\perp\!\!\!\perp O$ and $B \perp\!\!\!\perp O \mid C$

S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in ECCV, 2018, pp. 456–473.

CausalDA, an approach to break down irrelevant correlations

Definition. DAG-constrained Causal Data Augmentation

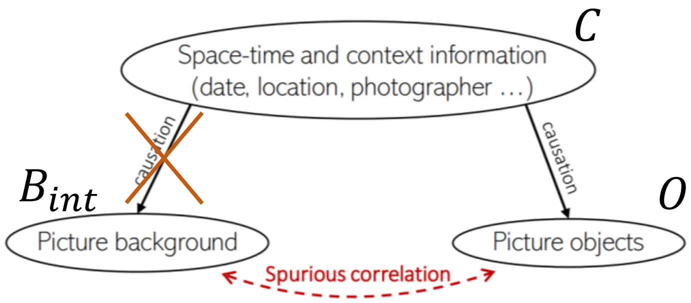
Given:

- a set of variables $\mathbf{X} = (X_1, \ldots, X_d)$ distributed according to P_{obs},
- a DAG G encoding the causal dependencies that the variables must follow,
- a set of interventions I_{spl} applied to $\boldsymbol{X_{int}} \subset \mathbf{X}$,

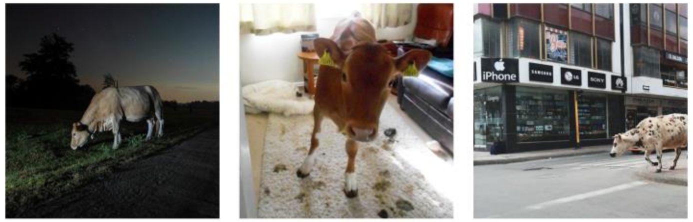
Causal Data Augmentation consists in sampling N data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G and the set of interventions $I_{spl} = \{P_{int}(X_k | PA_{int}(X_k)) \mid X_k \in \boldsymbol{X_{int}}\}$.

$$P_{spl}(X_1, \ldots, X_d) = \underbrace{\prod_{X_i \notin \boldsymbol{X_{int}}} P_{obs}(X_i | PA_G(X_i))}_{\text{In domain}} \underbrace{\prod_{X_k \in \boldsymbol{X_{int}}} P_{int}(X_k | PA_{int}(X_k))}_{\text{Out-of-domain}}$$

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98 (B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97 (C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94



$B_{int} \perp\!\!\!\perp O$ and $B_{int} \perp\!\!\!\perp C$



(A) A cow by night (B) A cow in a house (C) A cow in city center

Ekimetrics.

S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in ECCV, 2018, pp. 456–473.

Questions

References

Takeshi Teshima and Masashi Sugiyama. Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation. In Uncertainty in Artificial Intelligence, pp. 86–96, 2021

Audrey Poinsot and Alessandro Leite. A Guide for Practical Use of ADMG Causal Data Augmentation. In ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023. https://openreview.net/forum?id=kBcAZcKypug

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal Discovery Toolbox: uncovering causal relationships in Python. The Journal of Machine Learning Research, 21(1):1406–1410, 2020. https://jmlr.org/papers/v21/19-187.html

Duligur Ibeling and Thomas Icard Probabilistic reasoning across the causal hierarchy. In AAAI Conference on Artificial Intelligence, 2020.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's Hierarchy and the Foundations of Causal Inference. 2022.

Netflix Technology Blog. A survey of Causal Inference Applications at Netflix. Netflix TechBlog. 2021. https://netflixtechblog.com/a-survey-of-causal-inference-applications-at-netflix-b62d25175e6f

Christina Katsimerou. There's more to experimentation than A/B. Booking.com Data Science. 2020. https://booking.ai/theres-more-to-experimentation-than-a-b-223fba846876

S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in ECCV, 2018, pp. 456–473.

CausalDA, a promising approach to use with caution

Statistical KPIs matching business dynamics

Build a causal graph

Data reveal human biases
Experts alert on data issues

Apply Causal Data Augmentation

In domain:
- ADMGDA is a possible solution [Poinsot and Leite, 2023][1]
- Any other conditional density estimator might work

Out-of-domain:
- Causal Graphical Models such as Causal Bayesian Networks

Analyze the new dataset

Use the whole dataset to fit the models
Compute Marketing KPIs on observed data only

Adopt another perspective

Question your hypotheses

Ekimetrics.

[1] Audrey Poinsot and Alessandro Leite, A Guide for practical use of ADMG Causal Data Augmentation, In ICLR 2023 Workshop on Trustworthy ML, 2023.

Hybrid Causal Discovery to mitigate data and human biases

Data-driven Causal Discovery

Measurement error
Unobserved variables
Selection bias
Small data

Data reveal human biases

Expert-driven Causal Discovery

Wrong knowledge
Non-instantaneous reasoning
Human biases
Personal interest

Experts alert on data issues

Hybrid Causal Discovery

Inputs: data + knowledge
Output: DAG aligned with data & experts

Data

Business sense

1. **Causal modelization** — Build the causal graph through intuitions - business sense
2. **Estimands identification** — Is it possible to estimate the causal quantity we want?
3. **Confrontation with real data** — Conditional Independence Tests, Robustness Tests (DoWhy)
4. **Graph validation** — Validation or rejection of the tests results