Causal Machine Learning

Supervised Learning

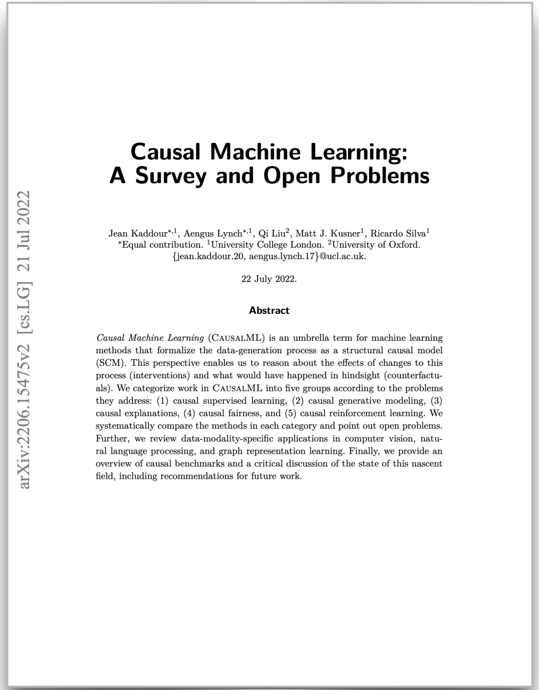
Jordi Vitrià
jordi.vitria@ub.edu

Universitat de Barcelona

Causal Machine Learning

Causal Machine Learning (CausalML) is an umbrella term for **machine learning methods** that are causally informed.

This perspective enables us to reason about the effects of changes in the data generation process (interventions) and what would have happened in hindsight (counterfactuals).



arXiv:2206.15475v2 [cs.LG] 21 Jul 2022

Causal Machine Learning:
A Survey and Open Problems

Jean Kaddour*[,1], Aengus Lynch*[,1], Qi Liu[2], Matt J. Kusner[1], Ricardo Silva[1]
*Equal contribution. [1]University College London. [2]University of Oxford.
{jean.kaddour.20, aengus.lynch.17}@ucl.ac.uk.

22 July 2022.

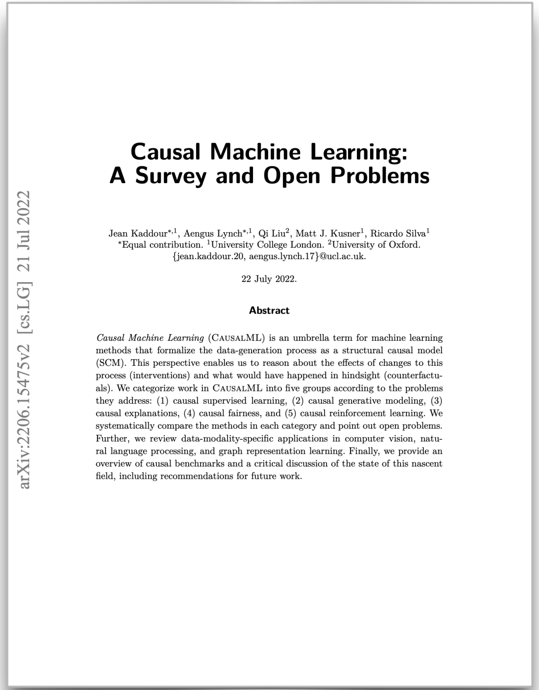
Abstract

Causal Machine Learning (CausalML) is an umbrella term for machine learning methods that formalize the data-generation process as a structural causal model (SCM). This perspective enables us to reason about the effects of changes to this process (interventions) and what would have happened in hindsight (counterfactuals). We categorize work in CausalML into five groups according to the problems they address: (1) causal supervised learning, (2) causal generative modeling, (3) causal explanations, (4) causal fairness, and (5) causal reinforcement learning. We systematically compare the methods in each category and point out open problems. Further, we review data-modality-specific applications in computer vision, natural language processing, and graph representation learning. Finally, we provide an overview of causal benchmarks and a critical discussion of the state of this nascent field, including recommendations for future work.

https://arxiv.org/pdf/2206.15475.pdf

Causal Machine Learning

We can categorize work in CausalML into five groups according to the problems they address: (1) **causal supervised learning**, (2) causal generative modeling, (3) causal explanations, (4) causal fairness, and (5) causal reinforcement learning.


arXiv:2206.15475v2 [cs.LG] 21 Jul 2022

Causal Machine Learning: A Survey and Open Problems

Jean Kaddour*[,1], Aengus Lynch*[,1], Qi Liu[2], Matt J. Kusner[1], Ricardo Silva[1]
*Equal contribution. [1]University College London. [2]University of Oxford.
{jean.kaddour.20, aengus.lynch.17}@ucl.ac.uk.

22 July 2022.

Abstract

Causal Machine Learning (CausalML) is an umbrella term for machine learning methods that formalize the data-generation process as a structural causal model (SCM). This perspective enables us to reason about the effects of changes to this process (interventions) and what would have happened in hindsight (counterfactuals). We categorize work in CausalML into five groups according to the problems they address: (1) causal supervised learning, (2) causal generative modeling, (3) causal explanations, (4) causal fairness, and (5) causal reinforcement learning. We systematically compare the methods in each category and point out open problems. Further, we review data-modality-specific applications in computer vision, natural language processing, and graph representation learning. Finally, we provide an overview of causal benchmarks and a critical discussion of the state of this nascent field, including recommendations for future work.

https://arxiv.org/pdf/2206.15475.pdf

Causal Supervised Learning

The goal of supervised learning is to learn the conditional distribution $P(Y|X)$, or more generally $\mathbb{E}(Y|X)$, by training on data of the form $D = \{(x_i, y_i)\}_{i=1}^N$, where X and Y denote covariates and label, respectively.

One of the most fundamental principles in supervised learning is to assume that our data D is independent and identically distributed (i.i.d.).

The validity of this assumption has been challenged; it has been famously called "*the big lie in machine learning*".

Causal Supervised Learning

As an alternative to the i.i.d. assumption, we can assume that our data is sampled from interventional distributions governed by a causal model.

For a given dataset generated across a set of environments ε, $\{(x_i^e, y_i^e)_{i=1}^N\}_{e \in \varepsilon}$, we view each environment $e \in \varepsilon$ as being sampled from a separate interventional distribution.

How can we estimate $P(Y \mid X)$ in a principled, robust manner?

Invariant Feature Learning

Invariant feature learning (IFL) is the task of identifying features of our data X, X_c, that are predictive of Y across a range of environments $\mathcal{E}$.

(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

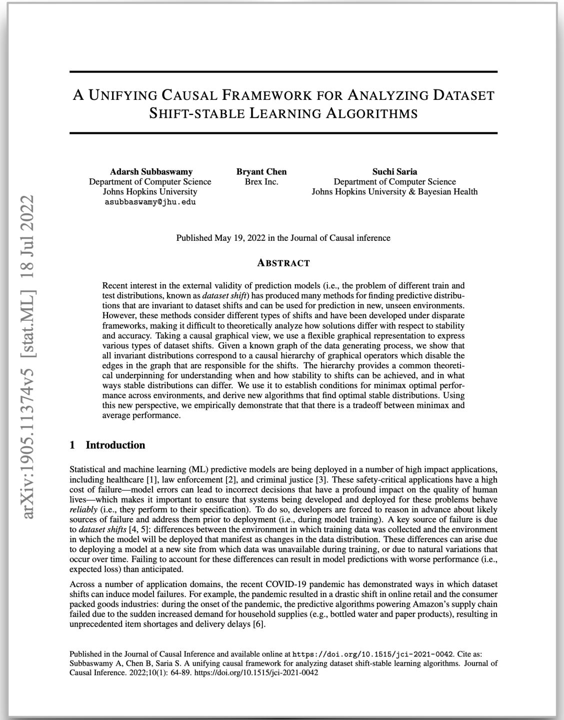
(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Distribution Shifts

In this paper, authors provide a unifying framework for **specifying dataset shifts** that can occur, analyzing model stability to these shifts, and determining conditions for achieving the lowest worst-case error across environments produced by these shifts.

This provides common ground so that we can begin to answer fundamental questions such as:

- **To what dataset shifts are the model's predictions stable vs unstable?** (Stability of the data generating model)
- **How will the model's performance be affected by these shifts?**



arXiv:1905.11374v5 [stat.ML] 18 Jul 2022

A UNIFYING CAUSAL FRAMEWORK FOR ANALYZING DATASET SHIFT-STABLE LEARNING ALGORITHMS

Adarsh Subbaswamy
Department of Computer Science
Johns Hopkins University
asubbaswamy@jhu.edu

Bryant Chen
Brex Inc.

Suchi Saria
Department of Computer Science
Johns Hopkins University & Bayesian Health

Published May 19, 2022 in the Journal of Causal inference

ABSTRACT

Recent interest in the external validity of prediction models (i.e., the problem of different train and test distributions, known as *dataset shift*) has produced many methods for finding predictive distributions that are invariant to dataset shifts and can be used for prediction in new, unseen environments. However, these methods consider different types of shifts and have been developed under disparate frameworks, making it difficult to theoretically analyze how solutions differ with respect to stability and accuracy. Taking a causal graphical view, we use a flexible graphical representation to express various types of dataset shifts. Given a known graph of the data generating process, we show that all invariant distributions correspond to a causal hierarchy of graphical operators which disable the edges in the graph that are responsible for the shifts. The hierarchy provides a common theoretical underpinning for understanding when and how stability to shifts can be achieved, and in what ways stable distributions can differ. We use it to establish conditions for minimax optimal performance across environments, and derive new algorithms that find optimal stable distributions. Using this new perspective, we empirically demonstrate that that there is a tradeoff between minimax and average performance.

1 Introduction

Statistical and machine learning (ML) predictive models are being deployed in a number of high impact applications, including healthcare [1], law enforcement [2], and criminal justice [3]. These safety-critical applications have a high cost of failure—model errors can lead to incorrect decisions that have a profound impact on the quality of human lives—which makes it important to ensure that systems being developed and deployed for these problems behave *reliably* (i.e., they perform to their specification). To do so, developers are forced to reason in advance about likely sources of failure and address them prior to deployment (i.e., during model training). A key source of failure is due to *dataset shifts* [4, 5]: differences between the environment in which training data was collected and the environment in which the model will be deployed that manifest as changes in the data distribution. These differences can arise due to deploying a model at a new site from which data was unavailable during training, or due to natural variations that occur over time. Failing to account for these differences can result in model predictions with worse performance (i.e., expected loss) than anticipated.

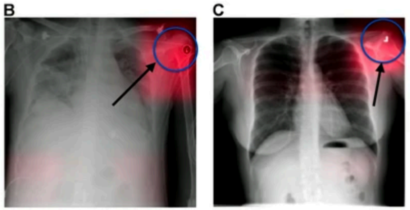
Across a number of application domains, the recent COVID-19 pandemic has demonstrated ways in which dataset shifts can induce model failures. For example, the pandemic resulted in a drastic shift in online retail and the consumer packed goods industries: during the onset of the pandemic, the predictive algorithms powering Amazon's supply chain failed due to the sudden increased demand for household supplies (e.g., bottled water and paper products), resulting in unprecedented item shortages and delivery delays [6].

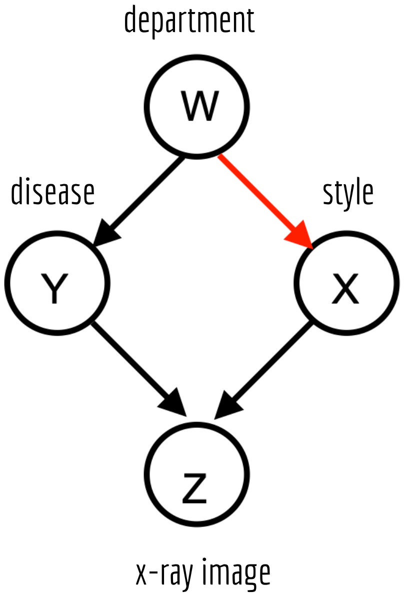
Published in the Journal of Causal Inference and available online at https://doi.org/10.1515/jci-2021-0042. Cite as: Subbaswamy A, Chen B, Saria S. A unifying causal framework for analyzing dataset shift-stable learning algorithms. Journal of Causal Inference. 2022;10(1): 64-89. https://doi.org/10.1515/jci-2021-0042

https://arxiv.org/pdf/1905.11374.pdf

Distribution Shifts

Example: The goal is to diagnose pneumonia Y from chest x-rays Z and stylistic features of the image X (i.e., orientation and coloring). The latent variable W represents the hospital department the patient visited.



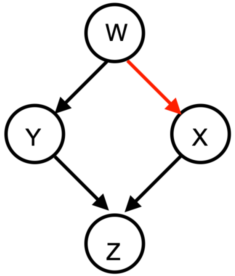


In the pneumonia example, each department has its own protocols and equipment, so the style preferences $P(X \mid W)$ vary across departments.

Distribution Shifts

Each environment is a different instantiation of that graph such that certain mechanisms differ.

Thus, the **factorization of the data distribution is the same in each environment**, but the **terms in the factorization corresponding to shifts will vary across environments**.



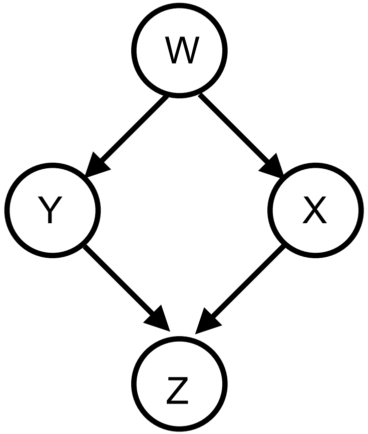
$$E = \{P(Z \mid Y, X)P(Y \mid W)P(X \mid W)P(W)\}$$

Distribution Shifts

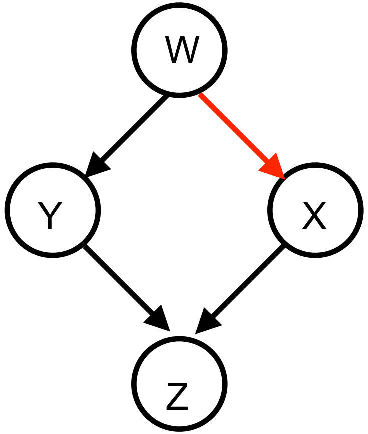
Key Result: Distribution shifts can be expressed in terms of edges.

Distribution Shifts

A graph and a set of edges which are marked as unstable defines an uncertainty set of environments whose distributions differ in the unstable factors.



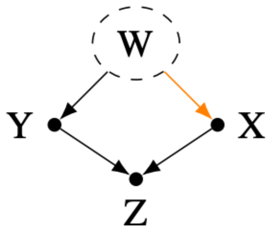
$$P(Z \mid Y, X) P(Y \mid W) P(X \mid W) P(W)$$



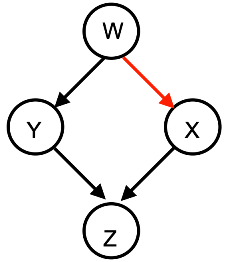
$$E = \{ P(Z \mid Y, X) P(Y \mid W) P(X \mid W) P(W) \}$$

Distribution Shifts

In this pneumonia example, if W is unobserved, a model of $P(Y \mid X, Z)$ will learn an association between Y and X through W. Thus, $P(Y \mid X, Z)$ contains an **unstable** path, and this distribution is **unstable** to shifts in the style mechanism. This means that $P(Y \mid X, Z)$ is different in each environment.



By contrast, if W were observed and we could condition on it, then $P(Y \mid X, Z, W)$ is **stable** to shifts in the style mechanism because all paths containing the unstable edge are blocked by W.
Thus, $P(Y \mid X, Z, W)$ is invariant across environments.
$P(Y \mid X, Z)$ is **unstable** because of the backdoor path.



Distribution Shifts

In order to achieve stable distributions to shifts we can

- find the maximal set of features to **condition** on so that the resulting model is stable with respect to the foreseen shifts,
- **intervene** $(do(\cdot))$ in variables with a shifted mechanism,
- compute **counterfactuals**.