Causal Inference and Deep Learning

Alexandre Guimarães

Oct 22, 2021

Motivation

Joshua Angrist, 2021 Nobel Prize Winner in Economics

"Econometrics is the original data science."

Potential Outcomes Framework

(Rubin-Neyman Causal Model)

- Suppose we have a treatment T, and an outcome Y
- We also have *covariates*, which cause both Y and T and can *confound* the treatment effect, and they are called X (high-dimensional)
- For instance, Y is the academic performance of the student, and T is the school providing tablets for studying
- Suppose that, for any given individual, there are two *potential outcomes* $Y(t)$:
  - $Y(T=1) = Y(1) = Y_1$, the outcome *with* the treatment;
  - $Y(T=0) = Y(0) = Y_0$, the outcome *without* the treatment.

The causal inference question

What is the *Average Treatment Effect* (ATE) of T on Y? Mathematically,

$$\text{ATE} = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$$

We can also ask what the effect of the treatment will be in each individual. This is known as Individual Treatment Effect (ITE) or Conditional Average Treatment Effect (CATE), and defined by

$$\text{ITE} = \mathbb{E}[Y_1 - Y_0|X]$$

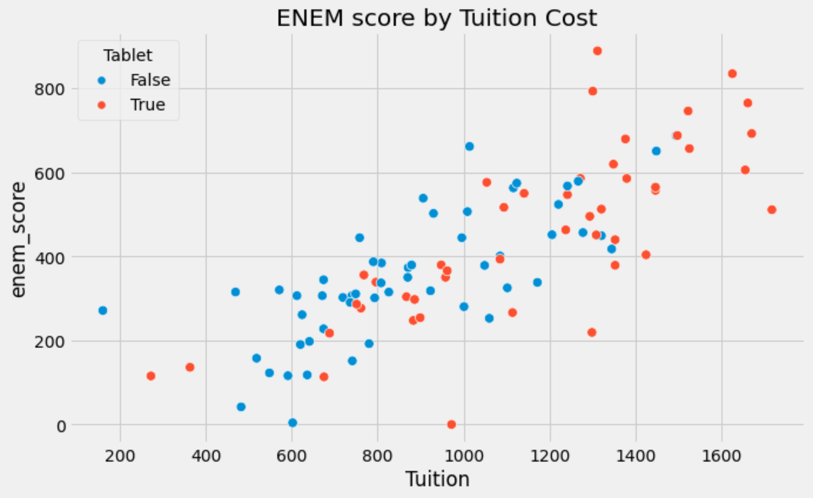
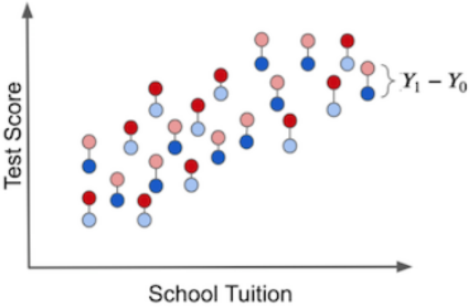


Figure 1: *

Causal Inference for The Brave and True

Why Causal Inference is hard

The fundamental problem of Causal Inference



We can only observe one potential outcome for each individual!

With observational data, we only know the *factual* outcome for each individual.

Observed factual outcome

For each individual x_i receiving treatment t_i,

$$y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$$

However, in order to do causal inference, we would need to also know the *counterfactuals*.

Unobserved counterfactual outcome

For each individual x_i receiving treatment t_i,

$$y_i = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$

Unlike in supervised learning, we don't have all the information!

Why Causal Inference is hard

Naïve estimation inconsistency

- Taking our tablet example, if someone says that schools that provide tablets tend to show better academic performance, we could reply that schools that give away tablets tend to be richer, and that may be causing the better performance, not the digital equipments.
- **Correlation is not causation!**

Mathematically, we can show that

$$\underbrace{\mathbb{E}[Y|T=1]-\mathbb{E}[Y|T=0]}_{\text{ATE}}=\underbrace{\mathbb{E}[Y_1-Y_0|T=1]}_{\text{ATET}}+\underbrace{\{\mathbb{E}[Y_0|T=1]-\mathbb{E}[Y_0|T=0]\}}_{\text{bias}}.$$

- The population that receives the treatment may be different than the one that doesn't, and that introduces a *bias*.
- If the populations are equal, the bias term will be zero, and correlation *will* be causation.

What we can do about it

- We can achieve zero bias with *randomization*, i.e., creating *control* and *treated* groups coming from the same population.
- However, randomization is expensive, and sometimes just can't be done.

Some thing we can do to learn from non-randomized observational data

- Covariate adjustment
- Propensity score re-weighting
- Doubly robust estimators
- Matching
- ...

Common assumptions

Ignorability

We will suppose that *all* the possible confounders are contained by X, i.e., there are no hidden confounders we don't know of (**talk to domain experts!**). This hypothesis is called $ignorability$ and can be formalized as

$$Y_0, Y_1 \perp\!\!\!\perp T | X,$$

i.e., the potential outcomes don't depend on T, given X. (The outcome Y does!)

Common support

Another assumption is that there is, to some extent, *overlap* between similar instances of X with different treatments, i.e., there is some kind of stochasticity. This hypothesis is called *common support*, and can be formalized as

$$0 < p(T = t | X = x) < 1 \ \ \forall t, x$$

We could explicitly model the outcome based on treatment and covariates (*outcome modeling*).
Let's say there is a function $h = h(x, T)$ and we would use an ML model, taking X and T as features, to fit it and predict Y, i.e.,

$$h(x, t) \approx \mathbb{E}[Y_t | T = t, x]$$

We could then estimate the ATE with

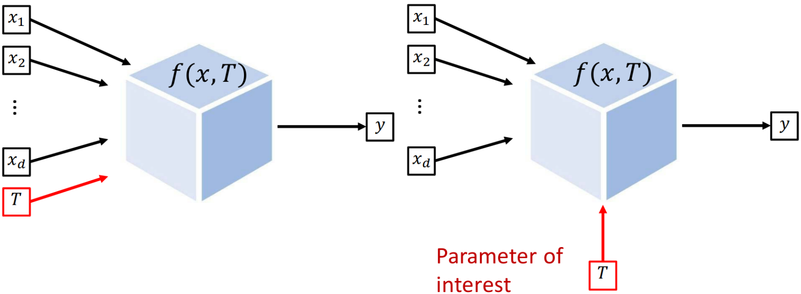
$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} [h(x_i, 1) - h(x_i, 0)]$$

and the ITE with

$$\widehat{\text{ITE}}(x_i) = h(x_i, 1) - h(x_i, 0)$$

The problem with the classical Machine Learning approach

- In classical Machine Learning, our measure of success will be the Y prediction accuracy.
- The problem is that there is no guarantee that the model will correctly learn the influence of T on Y.
- There may be features in X that predict Y much more strongly, and the model can simply learn to ignore T.



- The paradigm shift we have to make is that T is not just a feature like the others, but a *parameter of interest*.

Let's assume that our variables' relationship follows this linear model

$$Y_t(x) = \beta x + \gamma t + \epsilon_t,$$

where $\mathbb{E}[\epsilon_t] = 0$.
In this case, we can show that

$$\text{ATE} = \text{CATE} = \gamma$$

If our goal is *causal inference*, we should care about getting γ right, not Y.
Identification, not prediction!

What about model misspecification?

Now, assume we tried to fit the previous model to data that was generated from the process

$$Y_t(x) = \beta x + \gamma t + \delta x^2$$

We can show that ATE is still γ.
Our estimator would be

$$\hat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} t,$$

and our bias can be show to be

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]},$$

i.e., arbitrarily large or small (or negative) depending on δ. Misspecifying our model can simply lead us to a conclusion that is **opposite** from the truth.

We should not use models that make strong assumptions about our causal processes!

Some ML models that really work

- Random Forests
- Bayesian Trees
- Gaussian Processes
- **Neural Networks**

Why neural networks?

- They are almost **non-linear models** (but so are trees);
- Neural networks naturally extract relevant information through **representation learning**;
- Causal inference in quantitative data, text, images, and graphs.

The ideas behind neural networks for causal inference

Representation learning

The idea that neural networks can learn a function Φ that produces representations in which covariates are deconfounded from outcome and treatment, making downstream prediction of multiple outcomes (Y_1 and Y_0) possible.

Multi-task learning

In order to make sure the representations still do a good job in predicting the outcomes, we can use *multiple loss functions* and balance our fundamental tradeoff between prediction accuracy and deconfounded representations.

First concepts

Learning Representations for Counterfactual Inference

Fredrik D. Johansson* FREJOHK@CHALMERS.SE
CSE, Chalmers University of Technology, Göteborg, SE-412 96, Sweden

Uri Shalit* SHALIT@CS.NYU.EDU
David Sontag DSONTAG@CS.NYU.EDU
CIMS, New York University, 251 Mercer Street, New York, NY 10012 USA

* **Equal contribution**

Abstract

Observational studies are rising in importance due to the widespread accumulation of data in fields such as healthcare, education, employment and ecology. We consider the task of answering counterfactual questions such as, "Would this patient have lower blood sugar had she received a different medication?". We propose a new algorithmic framework for counterfactual inference which brings together ideas from domain adaptation and representation learning. In addition to a theoretical justification, we perform an empirical comparison with previous approaches to causal inference from observational data. Our deep learning algorithm significantly outperforms the previous state-of-the-art.

1. Introduction

Inferring causal relations is a fundamental problem in the sciences and commercial applications. The problem of causal inference is often framed in terms of *counterfactual* questions (Lewis, 1973; Rubin, 1974; Pearl, 2009) such as "Would this patient have lower blood sugar had she received a different medication?", or "Would the user have clicked on this ad had it been in a different color?". In this paper we propose a method to learn representations suited for counterfactual inference, and show its efficacy in both simulated and real world tasks.

We focus on counterfactual questions raised by what are known as *observational studies*. Observational studies are studies where interventions and outcomes have been recorded, along with appropriate context. For example, consider an electronic health record dataset collected over several years, where for each patient we have lab tests and past diagnoses, as well as data relating to their diabetic status, and the causal question of interest is which of two existing anti-diabetic medications A or B is better for a given patient. Observational studies are rising in importance due to the widespread accumulation of data in fields such as healthcare, education, employment and ecology. We believe machine learning will be called on more and more to help make better decisions in these fields, and that researchers should be careful to pay attention to the ways in which these studies differ from classic supervised learning, as explained in Section 2 below.

In this work we draw a connection between counterfactual inference and domain adaptation. We then introduce a form of regularization by enforcing similarity between the distributions of representations learned for populations with different interventions. For example, the representations for patients who received medication A versus those who received medication B. This reduces the variance from fitting a model on one distribution and applying it to another. In Section 3 we give several methods for learning such representations. In Section 4 we show our methods approximately minimizes an upper bound on a regret term in the counterfactual regime. The general method is outlined in Figure 1. Our work has commonalities with recent work on learning fair representations (Zemel et al., 2013; Louizos et al., 2015) and learning representations for transfer learning (Ben-David et al., 2007; Gani et al., 2015). In all these cases the learned representation has some invariance to specific aspects of the data: either an identity of a certain group such as racial minorities for fair representations, or the identity of the data source for domain adaptation, or, in the case of counterfactual learning, the type of intervention enacted in each population.

In machine learning, counterfactual questions typically arise in problems where there is a learning agent which performs actions, and receives feedback or reward for that

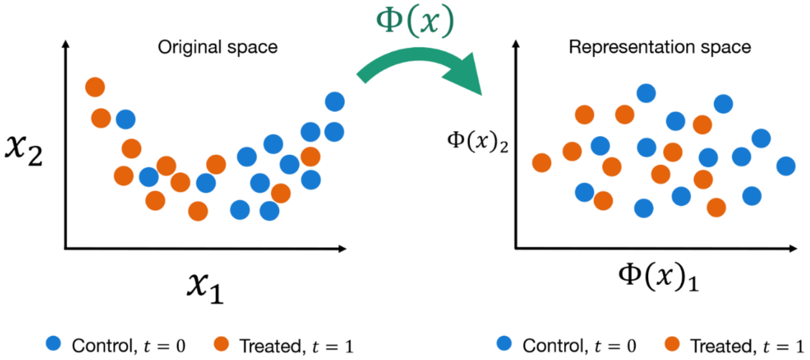
Proceedings of the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Learning Representations for Counterfactual Inference, 2016 (Fredrik D. Johansson, Uri Shalit, David Sontag.)

"We propose to perform counterfactual inference by amending the direct modeling approach, taking into account the fact that the learned estimator h must generalize from the factual distribution to the counterfactual distribution."

Balancing representation

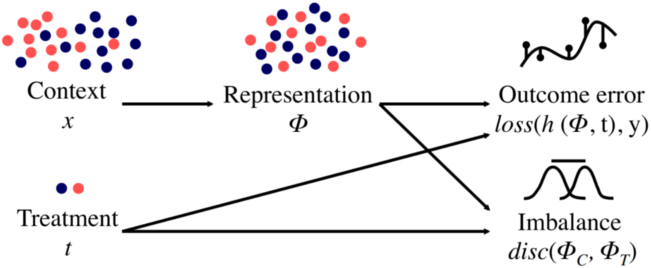
Instead of directly modeling $h(X, T)$, we learn a representation $\Phi(X)$ where the treated and control groups are similar or balanced, and only then apply $h(\Phi(X), T)$.



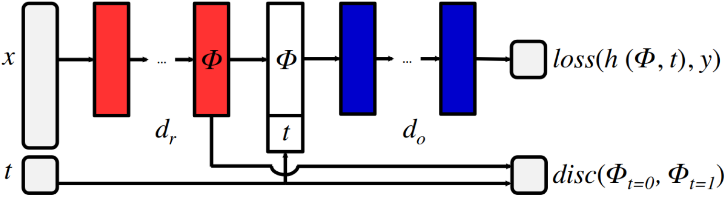
Balancing counterfactual regression

Our networks should trade off between 3 main objectives:

- low-error prediction of the observed outcomes;
- low-error prediction of unobserved outcomes;
- balance between treatment populations in representation space.



Neural network architecture



Our objective is to minimize

$$B_{\alpha,\gamma}(\Phi, h) = \overbrace{\frac{1}{n}\sum_{i=1}^{n}|h(\Phi(x_i), t_i) - y_i^{\mathrm{F}}|}^{\text{factual loss}} + \overbrace{\frac{\gamma}{n}\sum_{i=1}^{n}|h(\Phi(x_i), 1 - t_i) - y_{\mathrm{nn}(i)}^{\mathrm{F}}|}^{\text{nearest-neighbor ``counterfactual'' loss}} +$$

$$+ \underbrace{\alpha\,\mathrm{disc}(\hat{P}_{\Phi}^{\mathrm{F}} - \hat{P}_{\Phi}^{\mathrm{CF}})}_{\text{populations discrepancy}},$$

where α and γ are hyperparameters.

The seminal paper

Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit [*1] Fredrik D. Johansson [*2] David Sontag [2 3]

Abstract

There is intense interest in applying machine learning to problems of causal inference in fields such as healthcare, economics and education. In particular, individual-level causal inference has important applications such as precision medicine. We give a new theoretical analysis and family of algorithms for predicting individual treatment effect (ITE) from observational data, under the assumption known as strong ignorability. The algorithms learn a "balanced" representation such that the induced treated and control distributions look similar, and we give a novel and intuitive generalization-error bound showing the expected ITE estimation error of a representation is bounded by a sum of the standard generalization-error of that representation and the distance between the treated and control distributions induced by the representation. We use Integral Probability Metrics to measure distances between distributions, deriving explicit bounds for the Wasserstein and Maximum Mean Discrepancy (MMD) distances. Experiments on real and simulated data show the new algorithms match or outperform the state-of-the-art.

1. Introduction

Making predictions about causal effects of actions is a central problem in many domains. For example, a doctor deciding which medication will cause better outcomes for a patient; a government deciding who would benefit most from subsidized job training; or a teacher deciding which study program would most benefit a specific student. In this paper we focus on the problem of making these predictions based on *observational data*. Observational data is data which contains past actions, their outcomes, and possibly more context, but without direct access to the mechanism which gave rise to the action. For example we might have access to records of patients (context), their medications (actions), and outcomes, but we do not have complete knowledge of why a specific action was applied to a patient.

The hallmark of learning from observational data is that the actions observed in the data depend on variables which might also affect the outcome, resulting in *confounding*: For example, richer patients might better afford certain medications, and job training might only be given to those motivated enough to seek it. The challenge is how to untangle these confounding factors and make valid predictions. Specifically, we work under the common simplifying assumption of "no-hidden confounding", assuming that all the factors determining which actions were taken are observed. In the examples above, it would mean that we have measured a patient's wealth or an employee's motivation.

As a learning problem, estimating causal effects from observational data is different from classic learning in that in our training data we never see the individual-level effect. For each unit, we only see their response to one of the possible actions - the one they had actually received. This is close to what is known in the machine learning literature as "learning from logged bandit feedback" (Strehl et al., 2010; Swaminathan & Joachims, 2015), with the distinction that we do not have access to the model generating the action.

Our work differs from much work in causal inference in that we focus on the individual-level causal effect ("c-specific treatment effects" Shpitser & Pearl (2006); Pearl (2015)), rather than the average or population level. Our main contribution is to give what is, to the best of our knowledge, the first generalization-error[1] bound for estimating individual-level causal effect, where each individual is identified by its features x. The bound leads naturally to a new family of representation-learning based algorithms (Bengio et al., 2013), which we show to match or outperform state-of-the-art methods on several causal effect inference tasks.

[*]Equal contribution [1]CIMS, New York University, New York, NY 10003 [2]IMES, MIT, Cambridge, MA 02142 [3]CSAIL, MIT, Cambridge, MA 02139. Correspondence to: Uri Shalit <shalit@cs.nyu.edu>, Fredrik D. Johansson <fredrikj@mit.edu>, David Sontag <dsontag@csail.mit.edu>.

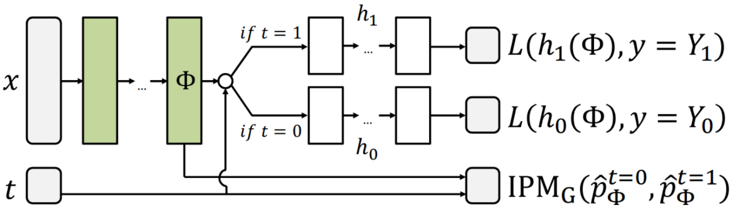
Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]Our use of the term generalization is different from its use in the study of *transportability*, where the goal is to generalize causal conclusion across distributions (Bareinboim & Pearl, 2016).

Estimating individual treatment effect: generalization bounds and algorithms, 2017 (Uri Shalit, Fredrik D. Johansson, David Sontag.)

"The bound we derive points the way to a family of algorithms based on the idea of representation learning (Bengio et al., 2013): Jointly learn hypotheses for both treated and control on top of a representation which minimizes a weighted sum of the factual loss (the standard supervised machine learning objective), and the IPM distance between the control and treated distributions induced by the representation."

Counterfactual Regression Net (CFRNet)



- Now our network has 2 heads, and also 2 regression functions (h_0 and h_1), one for each potential outcome.
- The Y_1 head is trained with samples from the treated group and the Y_0 head is trained with samples from the control group.
- The IPM is an *integral probability measure*, which measures the distance between the 2 distributions in the representation space.

Minimization objective and bounds

Our objective is now to find

$$\min_{h,\Phi} \frac{1}{n} \sum_{i=1}^{n} w_i \cdot L(h(\Phi(x_i), t_i), y_i) + \lambda \cdot \mathfrak{R}(h) + \alpha \cdot \mathrm{IPM_G}(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}),$$

where $w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}$, $u = \frac{1}{n}\sum_{i=1}^{n} t_i$, $\mathfrak{R}(h)$ is a model complexity term and α, γ are hyperparameters. Particularly, if $\alpha = 0$, the resulting architecture is called *Treatment Agnostic Representation Network* (TARNet).

Additionaly, this paper proves a bound on the expected error in estimating the ITE for a given representation. The expected *Precision in Estimation of Heterogeneous Effect* (PEHE), which is the MSE between predicted and true ITE, is bounded by

$$\epsilon_{\mathrm{PEHE}}(h, \Phi) \leq 2(\underbrace{\epsilon_{\mathrm{F}}^{t=0}(h, \Phi) + \epsilon_{\mathrm{F}}^{t=1}(h, \Phi)}_{\text{treatment and control losses}} + B_\Phi \underbrace{\mathrm{IPM_G}(p_\Phi^{t=1}, p_\Phi^{t=0})}_{\text{discrepancy}} - 2\sigma_Y^2)$$

Treatment modeling and targeted regularization

Adapting Neural Networks for the Estimation of Treatment Effects

Claudia Shi[1], David M. Blei[1,2], and Victor Veitch[2]

[1]Department of Computer Science, Columbia University
[2]Department of Statistics, Columbia University

Abstract

This paper addresses the use of neural networks for the estimation of treatment effects from observational data. Generally, estimation proceeds in two stages. First, we fit models for the expected outcome and the probability of treatment (propensity score) for each unit. Second, we plug these fitted models into a downstream estimator of the effect. Neural networks are a natural choice for the models in the first step. The question we address is: how can we adapt the design and training of the neural networks used in the first step in order to improve the quality of the final estimate of the treatment effect? We propose two adaptations based on insights from the statistical literature on the estimation of treatment effects. The first is a new architecture, the Dragonnet, that exploits the sufficiency of the propensity score for estimation adjustment. The second is a regularization procedure, targeted regularization, that induces a bias towards models that have non-parametrically optimal asymptotic properties 'out-of-the-box'. Studies on benchmark datasets for causal inference show these adaptations outperform existing methods. Code is available at github.com/claudiashi57/dragonnet.

1 Introduction

We consider the estimation of causal effects from observational data. Observational data is often readily available in situations where randomized control trials (RCT) are expensive or impossible. However, causal inference from observational data must address (possible) confounding factors that affect both treatment and outcome. Failure to adjust for confounders can lead to incorrect conclusions. To address this, a practitioner collects covariate information in addition to treatment and outcome status. The causal effect can be identified if the covariates contain all confounding variables. We will work in this 'no hidden confounding' setting throughout the paper. The task we consider is the estimation of the effect of a treatment T (e.g., a patient receives a drug) on an outcome Y (whether they recover) adjusting for covariates X (e.g., illness severity or socioeconomic status).

We consider how to use neural networks to estimate the treatment effect. The estimation of treatment effects proceeds in two stages. First, we fit models for the *conditional outcome* $Q(t, x) = \mathbb{E}[Y \mid t, x]$ and the *propensity score* $g(x) = P(T = 1|x)$. Then, we plug these fitted models into a downstream estimator. The strong predictive performance of neural networks motivates their use for effect estimation [e.g. SJS16; JSS16; Lou+17; AS17; AWS17; SLK18; YJS18; FLM18]. We will use neural networks as models for the conditional outcome and propensity score.

In principle, using neural networks for the conditional outcome and propensity score models is straightforward. We can use a standard net to predict the outcome Y from the treatment and covariates, and another to predict the treatment from the covariates. With a suitable choice of training objective, the trained models will yield consistent estimates of the conditional outcomes and propensity scores. However, neural network research has focused on *predictive* performance. What is

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Adapting Neural Networks for the Estimation of Treatment Effects, 2019 (Claudia Shi, David M. Blei, Victor Veitch.)

"We propose two adaptations based on insights from the statistical literature on the estimation of treatment effects. The first is a new architecture, the Dragonnet, that exploits the sufficiency of the propensity score for estimation adjustment. The second is a regularization procedure, targeted regularization, that induces a bias towards models that have non-parametrically optimal asymptotic properties 'out-of-the-box'."

Introducing propensity scores

Propensity Score

Measures the conditional distribution of T given X, and is given by $g(X) = p(T = 1|X)$. A famous result states that the propensity score is sufficient for estimating the ATE controlling X.

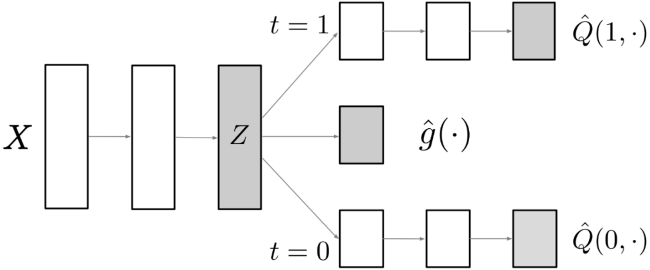
If the ATE is identifiable form the observational data, then

$$\begin{aligned} \text{ATE} &= \mathbb{E}\left[\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y|g(X), T = 1] - \mathbb{E}[Y|g(X), T = 0]\right] \end{aligned}$$

The intuition here is that, when controlling for X, the only part of X that is important is the one which confounds T. If it is not relevant for T, then it is not relevant for estimating treatment effect.

Using g in our estimation process means that we are not just modeling *outcomes*, but also the treatment itself.

Dragonnet architecture



Dragonnet is very similar to TARNet. The addition is the extra head g. This network tries to predict both *outcome* and *treatment*. The representation layer Z has a similar idea from the previous archtectures, but the reasoning here is leveraging the sufficiency of the propensity score and trading off *predictive accuracy* and *propensity score representation*. The minimization objective for Dragonnet is

$$\hat{\theta} = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} [\underbrace{(Q(t_i, x_i; \theta) - y_i)^2}_{\text{outcome loss}} + \alpha \underbrace{\text{CrossEntropy}(g(x_i; \theta), t_i)}_{\text{treatment loss}}]$$

Targeted Regularization

- Borrowing concepts from semi-parametric theory and Targeted Maximum Likelihood Estimation (TMLE), this paper introduces a form of regularization called Targeted Regularization (TarReg).
- TMLE attempts to optimize a likelihood function focusing on the aspects of the observed distribution that are most useful for estimating our target (in this case, treatment effects).
- It has several useful asymptotic properties, like consistency and confidence intervals.
- TarReg adapts the optimization process of the neural network by adding a parameter ϵ that forces the model to minimize the *Efficient Influence Curve* (EIC) from TMLE, and making the outcome model Q estimate the ATE more sharply.

Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects

Fredrik D. Johansson*[1], Uri Shalit[2], Nathan Kallus[3], David Sontag[4]

[1]Chalmers University of Technology
[2]Technion, Israel Institute of Technology
[3]Cornell Tech
[4]Massachusetts Institute of Technology

Abstract

Practitioners in diverse fields such as healthcare, economics and education are eager to apply machine learning to improve decision making. The cost and impracticality of performing experiments and a recent monumental increase in electronic record keeping has brought attention to the problem of evaluating decisions based on non-experimental observational data. This is the setting of this work. In particular, we study estimation of individual-level causal effects, such as a single patient's response to alternative medication, from recorded contexts, decisions and outcomes. We give generalization bounds on the error in estimated effects based on distance measures between groups receiving different treatments, allowing for sample re-weighting. We provide conditions under which our bound is tight and show how it relates to results for unsupervised domain adaptation. Led by our theoretical results, we devise representation learning algorithms that minimize our bound, by regularizing the representation's induced treatment group distance, and encourage sharing of information between treatment groups. We extend these algorithms to simultaneously learn a weighted representation to further reduce treatment group distances. Finally, an experimental evaluation on real and synthetic data shows the value of our proposed representation architecture and regularization scheme.

1 Introduction

Evaluating intervention decisions is a key question in many diverse fields including medicine, economics, and education. In medicine, an optimal choice of treatment for a patient in the intensive care unit may mean the difference between life and death. In public policy, job reforms have impact on the unemployment rate and the economy of a nation. To evaluate such interventions, we must study their *causal effect*—the difference in an outcome of interest under alternative choices of intervention. Since only one option may be carried out at a time, any data to support such evaluations only reveals the outcome of the action taken and never the outcome of the action not taken, which remains an unknown *counterfactual*.

*Correspondence to: fredrik.johansson@chalmers.se

Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects, 2021 (Fredrik D. Johansson, Uri Shalit, Nathan Kallus, David Sontag.)

"We give generalization bounds on the error in estimated effects based on distance measures between groups receiving different treatments, allowing for sample re-weighting. We provide conditions under which our bound is tight and show how it relates to results for unsupervised domain adaptation."

Hyperparameter selection for causal inference

- Selecting hyperparameters in supervised learning is fairly straightforward. We just have to experiment with different values and pick the combination that is best offers the best prediction on the validation set.
- In causal inference we have no access to our counterfactuals, which is what we are trying to predict, so how could we select the best hyperparameters for our observatoinal data?
- This paper proposes an experimental setup with baseline estimators as well as a hyperparameter selection method, based on nearest neighbors.

To substitute the ground-truth potential outcomes, we use pseudo-labels for the CATE. Suppose $j(i)$ is the nearest "*counterfacutal*" neighbor of the sample i in Euclidean distance, such that $t_{j(i)} \neq t_i$. The metric

$$\widehat{\text{MSE}}_{\text{nn}}(f) := \frac{1}{n} \sum_{i=1}^{n} \left[(1-2t)(y_{j(i)} - y_i) - (f(x_i, 1) - f(x_i, 0))\right]^2$$

is computed in the validation set for each hyperparameter combination, and the best combination is selected.

What is missing towards human-level AI?

- Models depend too much on training data and are not good at generalization;
- Machine Learning models assume that real-world data will have the same distribution as the one used in training;
- humans learn with no such assumption, by learning the underlying structure of reality, in other words, a *causal model*.

How can Machine Learning evolve?

Towards Causal Representation Learning

Bernhard Schölkopf †, Francesco Locatello †, Stefan Bauer ⋆, Nan Rosemary Ke ⋆, Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

***Abstract*—The two fields of machine learning and graphical causality arose and developed separately. However, there is now cross-pollination and increasing interest in both fields to benefit from the advances of the other. In the present paper, we review fundamental concepts of causal inference and relate them to crucial open problems of machine learning, including transfer and generalization, thereby assaying how causality can contribute to modern machine learning research. This also applies in the opposite direction: we note that most work in causality starts from the premise that the causal variables are given. A central problem for AI and causality is, thus, causal representation learning, the discovery of high-level causal variables from low-level observations. Finally, we delineate some implications of causality for machine learning and propose key research areas at the intersection of both communities.**

I. Introduction

If we compare what machine learning can do to what animals accomplish, we observe that the former is rather limited at some crucial feats where natural intelligence excels. These include transfer to new problems and any form of generalization that is not from one data point to the next (sampled from the same distribution), but rather from one problem to the next — both have been termed *generalization*, but the latter is a much harder form thereof, sometimes referred to as *horizontal*, *strong*, or *out-of-distribution* generalization. This shortcoming is not too surprising, given that machine learning often disregards information that animals use heavily: interventions in the world, domain shifts, temporal structure — by and large, we consider these factors a nuisance and try to engineer them away. In accordance with this, the majority of current successes of machine learning boil down to large scale pattern recognition on suitably collected *independent and identically distributed (i.i.d.)* data.

To illustrate the implications of this choice and its relation to causal models, we start by highlighting key research challenges.

a) Issue 1 – Robustness: With the widespread adoption of deep learning approaches in computer vision [101, 140], natural language processing [54], and speech recognition [85], a substantial body of literature explored the robustness of the prediction of state-of-the-art deep neural network architectures. The underlying motivation originates from the fact that in the real world there is often little control over the distribution from which the data comes from. In computer vision [75, 228], changes in the test distribution may, for instance, come from aberrations like camera blur, noise or compression quality [106, 129, 170, 206], or from shifts, rotations, or viewpoints [7, 11, 63, 282]. Motivated by this, new benchmarks were proposed to specifically test generalization of classification and detection methods with respect to simple algorithmically generated interventions like spatial shifts, blur, changes in brightness or contrast [106, 170], time consistency [94, 227], control over background and rotation [11], as well as images collected in multiple environments [19]. Studying the failure modes of deep neural networks from simple interventions has the potential to lead to insights into the inductive biases of state-of-the-art architectures. So far, there has been no definitive consensus on how to solve these problems, although progress has been made using data augmentation, pre-training, self-supervision, and architectures with suitable inductive biases w.r.t. a perturbation of interest [233, 59, 63, 137, 170, 206]. It has been argued [188] that such fixes may not be sufficient, and generalizing well outside the i.i.d. setting requires learning not mere statistical associations between variables, but an underlying *causal model*. The latter contains the mechanisms giving rise to the observed statistical dependences, and allows to model distribution shifts through the notion of interventions [183, 237, 218, 34, 188, 181].

b) Issue 2 – Learning Reusable Mechanisms: Infants' understanding of physics relies upon objects that can be tracked over time and behave consistently [52, 236]. Such a representation allows children to quickly learn new tasks as their knowledge and intuitive understanding of physics can be re-used [15, 52, 144, 250]. Similarly, intelligent agents that robustly solve real-world tasks need to re-use and re-purpose their knowledge and skills in novel scenarios. Machine learning models that incorporate or learn structural knowledge of an environment have been shown to be more efficient and generalize better [14, 10, 16, 84, 197, 212, 8, 274, 26, 76, 83, 141, 157, 177, 211, 245, 258, 272, 57, 182]. In a modular representation of the world where the modules correspond to physical causal mechanisms, many modules can be expected to behave similarly across different tasks and environments. An agent facing a new environment or task may thus only need to adapt a few modules in its internal representation of the world [220, 84]. When learning a causal model, one should thus require fewer examples to adapt as most knowledge, i.e.,

†: equal contribution.
⋆: equal contribution.
B. Schölkopf is at the Max-Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany, bs@tuebingen.mpg.de.
F. Locatello is at ETH Zurich, Computer Science Department and the Max Planck Institute for Intelligent Systems. Work partially done while interning at Google Research Amsterdam. francesco.locatello@gmail.com
S. Bauer is at the Max-Planck Institute for Intelligent Systems, stefan.bauer@tuebingen.mpg.de.
N. R. Ke is at Mila and the University of Montreal, *rosemary.nan.ke@gmail.com*.
N. Kalchbrenner is at Google Research Amsterdam, nalk@google.com.
A. Goyal is at Mila and the University of Montreal, anirudhgoyal9119@gmail.com
Y. Bengio is at Mila, the University of Montreal, CIFAR Senior Fellow *yoshua.bengio@mila.quebec*

Towards Causal Representation Learning, 2021 (Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio.)

"...causality, with its focus on representing structural knowledge about the data generating process that allows interventions and changes, can contribute towards understanding and resolving some limitations of current machine learning methods."

References

Papers

- Learning Representations for Counterfactual Inference
- Estimating individual treatment effect: generalization bounds and algorithms, 2017
- Adapting Neural Networks for the Estimation of Treatment Effects, 2019
- Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects, 2021
- Towards Causal Representation Learning, 2021

External Material

- Joshua Angrist interview
- Causal Inference for The Brave and True
- Yoshua Bengio Talk on Causal Representation Learning
- Deep Learning of Potential Outcomes
- Deep Learning for Causal Inference Tutorial - GitHub
- Causal Inference class in MIT's Machine Learning for Healthcare (Spring 2019) course