Learning Representations for Counterfactual Inference

Fredrik D.Johansson, Uri Shalit, David Sontag [1]

Benjamin Dubois-Taine

Feb 12th, 2020

The University of British Columbia

Setting of Causal Inference

Define the following:

- $\mathcal{X}$ the set of contexts
- $\mathcal{T}$ the set of possible actions
- $\mathcal{Y}$ the set of possible outcomes

Setting of Causal Inference

Define the following:

- $\mathcal{X}$ the set of contexts
- $\mathcal{T}$ the set of possible actions
- $\mathcal{Y}$ the set of possible outcomes

For all $t \in \mathcal{T}$, denote $Y_t(x) \in \mathcal{Y}$ the potential outcome for $x \in \mathcal{X}$.

Setting of Causal Inference

Define the following:

- $\mathcal{X}$ the set of contexts
- $\mathcal{T}$ the set of possible actions
- $\mathcal{Y}$ the set of possible outcomes

For all $t \in \mathcal{T}$, denote $Y_t(x) \in \mathcal{Y}$ the potential outcome for $x \in \mathcal{X}$.

Fundamental problem of causal inference: we can only observe $Y_t(x)$ for one specific value of t.

More setting

We will only look at the case where $\mathcal{T} = \{0, 1\}$.
Two quantities of interest are then

- **Individual Treatment Effect**

$$\text{ITE}(x) = Y_1(x) - Y_0(x)$$

- **Average Treatment Effect**

$$\text{ATE} = \mathbb{E}_{x \sim p(x)}\Big[\text{ITE}(x)\Big]$$

More settings

Finally, we define

- the observed outcome associated with x as the **factual outcome**, denoted $y^F(x)$.
- the unobserved outcome associated with x as the **counterfactual outcome**, denoted $y^{CF}(x)$.

Goal of The Paper

Come up with a framework to train models for factual and counterfactual inference.

A First Supervised Approach

- Given n samples $\{x_i, t_i, y_i^F\}_{i=1}^n$, where $y_i^F = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$

A First Supervised Approach

- Given n samples $\{x_i, t_i, y_i^F\}_{i=1}^n$, where $y_i^F = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$
- Learn a function $h : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ such that

$$h(x_i, t_i) \approx y_i^F$$

A First Supervised Approach

- Given n samples $\{x_i, t_i, y_i^F\}_{i=1}^n$, where $y_i^F = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$
- Learn a function $h : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ such that

$$h(x_i, t_i) \approx y_i^F$$

- To compute ITE on training data we could do

$$\hat{\text{ITE}}(x_i) = \begin{cases} y_i^F - h(x_i, t_i - 1) & \text{if } t_i = 1 \\ h(x_i, 1 - t_i) - y_i^F & \text{if } t_i = 0 \end{cases}$$

A First Supervised Approach

- Given n samples $\{x_i, t_i, y_i^F\}_{i=1}^n$, where $y_i^F = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$
- Learn a function $h : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ such that

$$h(x_i, t_i) \approx y_i^F$$

- To compute ITE on training data we could do

$$\hat{\text{ITE}}(x_i) = \begin{cases} y_i^F - h(x_i, t_i - 1) & \text{if } t_i = 1 \\ h(x_i, 1 - t_i) - y_i^F & \text{if } t_i = 0 \end{cases}$$

What is the problem with this ?

- We are training on the set

$$\hat{P}^F = \{(x_i, t_i)\}_{i=1}^n$$

with $\hat{P}^F \sim P^F$, the empirical **factual distribution**.

- We are training on the set

$$\hat{P}^F = \{(x_i, t_i)\}_{i=1}^n$$

with $\hat{P}^F \sim P^F$, the empirical **factual distribution**.

- We are inferring on the set

$$\hat{P}^{CF} = \{(x_i, 1 - t_i)\}_{i=1}^n$$

with $\hat{P}^{CF} \sim P^{CF}$, the empirical **counterfactual distribution**.

- We are training on the set

$$\hat{P}^F = \{(x_i, t_i)\}_{i=1}^n$$

with $\hat{P}^F \sim P^F$, the empirical **factual distribution**.

- We are inferring on the set

$$\hat{P}^{CF} = \{(x_i, 1 - t_i)\}_{i=1}^n$$

with $\hat{P}^{CF} \sim P^{CF}$, the empirical **counterfactual distribution**.

We do not want to make assumptions on the treatment assignment.

The Approach Proposed

The authors propose a general approach for causal inference

- Learn a representation $\Phi : \mathcal{X} \to \mathbb{R}^d$.

The Approach Proposed

The authors propose a general approach for causal inference

- Learn a representation $\Phi : \mathcal{X} \to \mathbb{R}^d$.
- Learn a function h from a hypothesis class $\mathcal{H}$, such that $h : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}$ predicts the outcome.

The Approach Proposed

We want the built representation (Φ, h) to balance the trade-offs between

- being able to achieve low-error prediction on the factual outcomes

The Approach Proposed

We want the built representation (Φ, h) to balance the trade-offs between

- being able to achieve low-error prediction on the factual outcomes
- being able to achieve low-error prediction on unobserved counterfactual outcomes.

The Approach Proposed

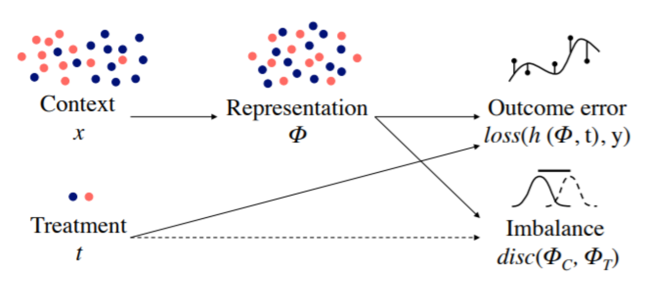
We want the built representation (Φ, h) to balance the trade-offs between

- being able to achieve low-error prediction on the factual outcomes
- being able to achieve low-error prediction on unobserved counterfactual outcomes.
- the distribution of treatment populations under Φ are similar/balanced.

The Approach Proposed

We want the built representation (Φ, h) to balance the trade-offs between

- being able to achieve low-error prediction on the factual outcomes
- being able to achieve low-error prediction on unobserved counterfactual outcomes.
- the distribution of treatment populations under Φ are similar/balanced.



How to evaluate performance of (Φ, h) on factual outcomes ?

How to evaluate performance of (Φ, h) on factual outcomes ?

- That is simple, we can simply compute

$$\frac{1}{n}\sum_{i=1}^{n}|h(\Phi(x_i), t_i) - y_i^F|$$

How to evaluate performance of (Φ, h) on factual outcomes ?

- That is simple, we can simply compute

$$\frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F|$$

How to evaluate performance of (Φ, h) on counterfactual outcomes ?

How to evaluate performance of (Φ, h) on factual outcomes ?

- That is simple, we can simply compute

$$\frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F|$$

How to evaluate performance of (Φ, h) on counterfactual outcomes ?

- For any x_i, compute

$$j(i) = \underset{j \in \{1,\dots,n\} \text{ with } t_j = 1 - t_i}{\arg\min} d(x_i, x_j)$$

How to evaluate performance of (Φ, h) on factual outcomes ?

- That is simple, we can simply compute

$$\frac{1}{n}\sum_{i=1}^{n}|h(\Phi(x_i), t_i) - y_i^F|$$

How to evaluate performance of (Φ, h) on counterfactual outcomes ?

- For any x_i, compute

$$j(i) = \underset{j\in\{1,\dots,n\} \text{ with } t_j = 1 - t_i}{\arg\min} d(x_i, x_j)$$

- Then the error term is

$$\frac{1}{n}\sum_{i=1}^{n}|h(\Phi(x_i), 1 - t_i) - y_{j(i)}^F|$$

How to encourage similarity between the empirical factual and counterfactual distributions $\hat{P}^F_\Phi$ and $\hat{P}^{CF}_\Phi$?

How to encourage similarity between the empirical factual and counterfactual distributions $\hat{P}^F_\Phi$ and $\hat{P}^{CF}_\Phi$?

- By controlling the discrepancy between them, namely given our hypothesis class $\mathcal{H}$ and a loss function L, we have

$$\text{disc}_{\mathcal{H}}(\hat{P}^F_\Phi, \hat{P}^{CF}_\Phi) = \max_{\beta, \beta' \in \mathcal{H}} \left[\mathop{\mathbb{E}}_{z \sim \hat{P}^F_\Phi} [L(\beta(z), \beta'(z))] - \mathop{\mathbb{E}}_{z \sim \hat{P}^{CF}_\Phi} [L(\beta(z), \beta'(z))] \right]$$

How to encourage similarity between the empirical factual and counterfactual distributions $\hat{P}_{\Phi}^{F}$ and $\hat{P}_{\Phi}^{CF}$?

- By controlling the discrepancy between them, namely given our hypothesis class $\mathcal{H}$ and a loss function L, we have

$$\mathrm{disc}_{\mathcal{H}}(\hat{P}_{\Phi}^{F}, \hat{P}_{\Phi}^{CF}) = \max_{\beta,\beta' \in \mathcal{H}} \left[\mathop{\mathbb{E}}_{z \sim \hat{P}_{\Phi}^{F}} [L(\beta(z), \beta'(z))] - \mathop{\mathbb{E}}_{z \sim \hat{P}_{\Phi}^{CF}} [L(\beta(z), \beta'(z))] \right]$$

- In this paper we only deal L being the square loss

How to encourage similarity between the empirical factual and counterfactual distributions $\hat{P}_\Phi^F$ and $\hat{P}_\Phi^{CF}$?

- By controlling the discrepancy between them, namely given our hypothesis class $\mathcal{H}$ and a loss function L, we have

$$\text{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}) = \max_{\beta,\beta' \in \mathcal{H}} \left[\mathop{\mathbb{E}}_{z \sim \hat{P}_\Phi^F}[L(\beta(z), \beta'(z))] - \mathop{\mathbb{E}}_{z \sim \hat{P}_\Phi^{CF}}[L(\beta(z), \beta'(z))] \right]$$

- In this paper we only deal L being the square loss
- Discrepancy in the case of linear hypotheses class, namely $\mathcal{H} \subset \mathbb{R}^{d+1}$, has a closed form formula.

How to encourage similarity between the empirical factual and counterfactual distributions $\hat{P}_\Phi^F$ and $\hat{P}_\Phi^{CF}$?

- By controlling the discrepancy between them, namely given our hypothesis class $\mathcal{H}$ and a loss function L, we have

$$\mathrm{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}) = \max_{\beta,\beta' \in \mathcal{H}} \left[\mathop{\mathbb{E}}_{z \sim \hat{P}_\Phi^F}[L(\beta(z), \beta'(z))] - \mathop{\mathbb{E}}_{z \sim \hat{P}_\Phi^{CF}}[L(\beta(z), \beta'(z))] \right]$$

- In this paper we only deal L being the square loss
- Discrepancy in the case of linear hypotheses class, namely $\mathcal{H} \subset \mathbb{R}^{d+1}$, has a closed form formula.
- From now on we restrict the study to linear hypotheses.

This gives rise to the following objective function

$$
\begin{aligned}
B_{\mathcal{H},\alpha,\gamma}(\Phi, h) = & \frac{1}{n} \sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F| \\
& + \frac{\gamma}{n} \sum_{i=1}^{n} |h(\Phi(x_i), 1 - t_i) - y_{j(i)}^F| + \\
& + \alpha \mathrm{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF})
\end{aligned}
$$

This gives rise to the following objective function

$$
\begin{aligned}
B_{\mathcal{H},\alpha,\gamma}(\Phi, h) =& \frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F| \\
&+ \frac{\gamma}{n}\sum_{i=1}^{n} |h(\Phi(x_i), 1 - t_i) - y_{j(i)}^F| + \\
&+ \alpha \mathrm{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF})
\end{aligned}
$$

Algorithm 1 Balancing counterfactual regression

1: **Input:** $X, T, Y^F; \mathcal{H}, \mathcal{N}; \alpha, \gamma, \lambda$

2: $\Phi^*, g^* = \underset{\Phi \in \mathcal{N}, g \in \mathcal{H}}{\arg\min}\, B_{\mathcal{H},\alpha,\gamma}(\Phi, g)$ (2)

3: $h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (h(\Phi, t_i) - y_i^F)^2 + \lambda \|h\|_{\mathcal{H}}$

4: **Output:** h^*, Φ^*

Theoretical Motivation behind Algorithm 1

- The former analysis gave an intuition on the form of the objective function $B_{\mathcal{H},\alpha,\gamma}(\Phi, h)$

Theoretical Motivation behind Algorithm 1

- The former analysis gave an intuition on the form of the objective function $B_{\mathcal{H},\alpha,\gamma}(\Phi, h)$
- Existence of a theoretical bound

A Theoretical Bound

Theorem 1. *For a sample* $\{(x_i, t_i, y_i^F)\}_{i=1}^n$, $x_i \in \mathcal{X}$, $t_i \in \{0,1\}$ *and* $y_i \in \mathcal{Y}$, *and a given representation function* $\Phi : \mathcal{X} \to \mathbb{R}^d$, *let* $\hat{P}_\Phi^F = (\Phi(x_1), t_1), \ldots, (\Phi(x_n), t_n)$, $\hat{P}_\Phi^{CF} = (\Phi(x_1), 1-t_1), \ldots, (\Phi(x_n), 1-t_n)$. *We assume that* $\mathcal{X}$ *is a metric space with metric* d, *and that the potential outcome functions* $Y_0(x)$ *and* $Y_1(x)$ *are Lipschitz continuous with constants* K_0 *and* K_1 *respectively, such that* $\mathrm{d}(x_a, x_b) \leq c \implies |Y_t(x_a) - Y_t(x_b)| \leq K_t \cdot c$ *for* $t = 0, 1$.

Let $\mathcal{H}_l \subset \mathbb{R}^{d+1}$ *be the space of linear functions* $\beta : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$, *and for* $\beta \in \mathcal{H}_l$, *let* $\mathcal{L}_P(\beta) = \mathbb{E}_{(x,t,y)\sim P}[L(\beta(x,t), y)]$ *be the expected loss of* β *over distribution* P. *Let* $r = max\left(\mathbb{E}_{(x,t)\sim P^F}[\|[\Phi(x), t]\|_2], \mathbb{E}_{(x,t)\sim P^{CF}}[\|[\Phi(x), t]\|_2]\right)$ *be the maximum expected radius of the distributions. For* $\lambda > 0$, *let* $\hat{\beta}^F(\Phi) = \arg\min_{\beta \in \mathcal{H}_l} \mathcal{L}_{\hat{P}_\Phi^F}(\beta) + \lambda\|\beta\|_2^2$, *and* $\hat{\beta}^{CF}(\Phi)$ *similarly for* $\hat{P}_\Phi^{CF}$, *i.e.* $\hat{\beta}^F(\Phi)$ *and* $\hat{\beta}^{CF}(\Phi)$ *are the ridge regression solutions for the factual and counterfactual empirical distributions, respectively.*

Let $\hat{y}_i^F(\Phi, h) = h^\top[\Phi(x_i), t_i]$ *and* $\hat{y}_i^{CF}(\Phi, h) = h^\top[\Phi(x_i), 1 - t_i]$ *be the outputs of the hypothesis* $h \in \mathcal{H}_l$ *over the representation* $\Phi(x_i)$ *for the factual and counterfactual settings of* t_i, *respectively. Finally, for each* $i, j \in \{1 \ldots n\}$, *let* $\mathrm{d}_{i,j} \equiv \mathrm{d}(x_i, x_j)$ *and* $j(i) \in \arg\min_{j \in \{1\ldots n\} \text{ s.t. } t_j = 1-t_i} \mathrm{d}(x_j, x_i)$ *be the nearest neighbor in* $\mathcal{X}$ *of* x_i *among the group that received the opposite treatment from unit* i. *Then for* both $Q = P^F$ *and* $Q = P^{CF}$ *we have:*

$$\frac{\lambda}{\mu r}(\mathcal{L}_Q(\hat{\beta}^F(\Phi)) - \mathcal{L}_Q(\hat{\beta}^{CF}(\Phi)))^2 \leq$$

$$disc_{\mathcal{H}_l}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}) + \tag{3}$$

$$\min_{h \in \mathcal{H}_l} \frac{1}{n} \sum_{i=1}^{n} (|\hat{y}_i^F(\Phi, h) - y_i^F| + |\hat{y}_i^{CF}(\Phi, h) - y_i^{CF}|) \leq \tag{4}$$

$$disc_{\mathcal{H}_l}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}) +$$

$$\min_{h \in \mathcal{H}_l} \frac{1}{n} \sum_{i=1}^{n} \left(|\hat{y}_i^F(\Phi, h) - y_i^F| + |\hat{y}_i^{CF}(\Phi, h) - y_{j(i)}^F|\right) + \tag{5}$$

$$\frac{K_0}{n} \sum_{i:t_i=1} \mathrm{d}_{i,j(i)} + \frac{K_1}{n} \sum_{i:t_i=0} \mathrm{d}_{i,j(i)}. \tag{6}$$

A Theoretical Bound

- Let Φ be any representation function.

A Theoretical Bound

- Let Φ be any representation function.
- Let $\mathcal{H}_\ell = \mathbb{R}^{d+1}$ be the space of linear functions.

A Theoretical Bound

- Let Φ be any representation function.
- Let $\mathcal{H}_\ell = \mathbb{R}^{d+1}$ be the space of linear functions.
- Let $\hat{\beta}^F(\Phi) = \underset{\beta \in \mathcal{H}_\ell}{\arg\min} \underset{(x,t,y) \sim \hat{P}^F_\Phi}{\mathbb{E}} \left[L(\beta(x,t), y) \right] + \lambda ||\beta||_2^2$, the ridge regression solutions for the factual empirical distributions.
- Define $\hat{\beta}^{CF}(\Phi)$ similarly

A Theoretical Bound

- Let Φ be any representation function.
- Let $\mathcal{H}_\ell = \mathbb{R}^{d+1}$ be the space of linear functions.
- Let $\hat{\beta}^F(\Phi) = \arg\min_{\beta \in \mathcal{H}_\ell} \mathbb{E}_{(x,t,y)\sim \hat{P}^F_\Phi}\Big[L(\beta(x,t),y)\Big] + \lambda||\beta||_2^2$, the ridge regression solutions for the factual empirical distributions.
- Define $\hat{\beta}^{CF}(\Phi)$ similarly
- The theorem then states that for both $Q = P^F$ and $Q = P^{FC}$, we have

$$
\begin{aligned}
& c_1\Big(\mathcal{L}_Q(\hat{\beta}^F(\Phi)) - \mathcal{L}_Q(\hat{\beta}^{CF}(\Phi))\Big) \\
\leq & \min_{h \in \mathcal{H}_\ell} \frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F| + |h(\Phi(x_i), 1-t_i) - y_{j(i)}^F| \\
& + \mathrm{disc}_{\mathcal{H}_\ell}(\hat{P}^F_\Phi, \hat{P}^{CF}_\Phi) \\
& + \frac{K_0}{n}\sum_{i:t_i=1} d(x_i, x_{j(i)}) + \frac{K_1}{n}\sum_{i:t_i=0} d(x_i, x_{j(i)})
\end{aligned}
$$

A Theoretical Bound

The theorem states that for both $Q = P^F$ and $Q = P^{FC}$, we have

$$c_1\Big(\mathcal{L}_Q(\hat{\beta}^F(\Phi)) - \mathcal{L}_Q(\hat{\beta}^{CF}(\Phi))\Big)$$

$$\leq \min_{h\in\mathcal{H}_\ell} \frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F| + |h(\Phi(x_i), 1-t_i) - y_i^{CF}|$$

$$+\text{disc}_{\mathcal{H}_\ell}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF})$$

$$+\frac{K_0}{n}\sum_{i:t_i=1} d(x_i, x_{j(i)}) + \frac{K_1}{n}\sum_{i:t_i=0} d(x_i, x_{j(i)})$$

Which is close to

$$B_{\mathcal{H},\alpha,\gamma}(\Phi, h) = \frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F|$$

$$+\frac{\gamma}{n}\sum_{i=1}^{n} |h(\Phi(x_i), 1-t_i) - y_{j(i)}^F| +$$

$$+\alpha\text{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF})$$

How to Choose the Representation function Φ ?

- Two approaches are proposed.

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- First one is by directly re-weighting the features of X, namely

$$\Phi(x) = Wx$$

where W is a diagonal matrix with $w_i \geq 0$, $\sum_i w_i = 1$.

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- First one is by directly re-weighting the features of X, namely

$$\Phi(x) = Wx$$

  where W is a diagonal matrix with $w_i \geq 0$, $\sum_i w_i = 1$.
- One can then show that

$$\mathrm{disc}_{\mathcal{H}_\ell}(\hat{P}^F_\Phi, \hat{P}^{CF}_\Phi) \approx ||W(p \sum_{i:t_i=1} x_i - (1-p) \sum_{i:t_i=0} x_i||_2$$

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- First one is by directly re-weighting the features of X, namely

$$\Phi(x) = Wx$$

  where W is a diagonal matrix with $w_i \geq 0$, $\sum_i w_i = 1$.
- One can then show that

$$\text{disc}_{\mathcal{H}_\ell}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}) \approx ||W(p \sum_{i:t_i=1} x_i - (1-p) \sum_{i:t_i=0} x_i||_2$$

- Features that differ a lot between treatment groups will receive a smaller weight

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- Second is with Neural Networks

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- Second is with Neural Networks

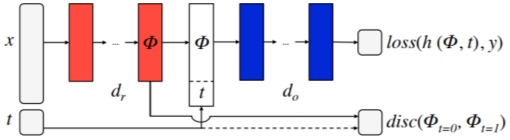


Figure 2. Neural network architecture.

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- Second is with Neural Networks

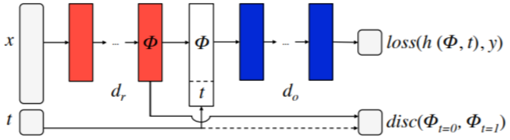


Figure 2. Neural network architecture.

- First d_r layers learn the representation Φ

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- Second is with Neural Networks

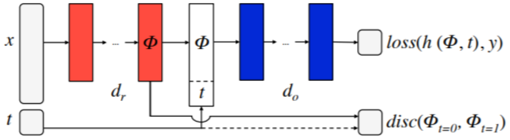


Figure 2. Neural network architecture.

- First d_r layers learn the representation Φ
- The d_o layers learn h given t

How to Choose the Representation function Φ ?

- Two approaches are proposed.
- Second is with Neural Networks

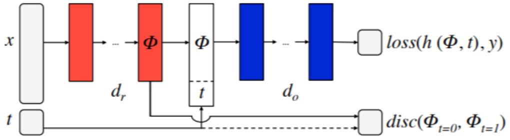


Figure 2. Neural network architecture.

- First d_r layers learn the representation Φ
- The d_o layers learn h given t
- Given Φ, the discrepancy is calculated

Experiments

- We don't have the data !
- Need to simulate

News Dataset

- The units x_i are news items in $\mathbb{N}^V$, i.e. word counts from the NY Times corpus, with $n = 5000$.

News Dataset

- The units x_i are news items in $\mathbb{N}^V$, i.e. word counts from the NY Times corpus, with $n = 5000$.
- The representation $\Phi(x_i) \in \mathbb{R}^{50}$ is the topic distribution of x_i, obtained using a LDA model with 50 topics.

News Dataset

- The units x_i are news items in $\mathbb{N}^V$, i.e. word counts from the NY Times corpus, with $n = 5000$.
- The representation $\Phi(x_i) \in \mathbb{R}^{50}$ is the topic distribution of x_i, obtained using a LDA model with 50 topics.
- The treatment t_i represents what device was used to read the news item.
  $t_i = 1$ for mobile, $t_i = 0$ for desktop.

News Dataset

- The units x_i are news items in $\mathbb{N}^V$, i.e. word counts from the NY Times corpus, with $n = 5000$.
- The representation $\Phi(x_i) \in \mathbb{R}^{50}$ is the topic distribution of x_i, obtained using a LDA model with 50 topics.
- The treatment t_i represents what device was used to read the news item.
  $t_i = 1$ for mobile, $t_i = 0$ for desktop.
- the factual outcome $y^F(x_i) \in \mathbb{R}$ is the readers experience of x_i

The News Dataset

The outcomes are generated as follows

- Pick two centroids in topic space, z_1 at random, and z_0 is the average of topic distribution

The News Dataset

The outcomes are generated as follows

- Pick two centroids in topic space, z_1 at random, and z_0 is the average of topic distribution
- The generated outcome of x_i with treatment t_i is then

$$y(x_i) = C(z(x_i)^T z_0 + t_i z(x_i)^T z_1)$$

The News Dataset

The outcomes are generated as follows

- Pick two centroids in topic space, z_1 at random, and z_0 is the average of topic distribution
- The generated outcome of x_i with treatment t_i is then

$$y(x_i) = C(z(x_i)^T z_0 + t_i z(x_i)^T z_1)$$

- Finally, we assume that the assignment of a news item x_i to a device t_i is biased towards the preferred devices, i.e.

$$p(t_i = 1 \mid x_i) = \frac{e^{\kappa z(x_i)^T z_1}}{e^{\kappa z(x_i)^T z_0} + e^{\kappa z(x_i)^T z_1}}$$

Results

The authors compare

- The balanced linear regression model (BLR), i.e. $\Phi(x) = Wx$.

Results

The authors compare

- The balanced linear regression model (BLR), i.e. $\Phi(x) = Wx$.
- A neural network with 4 layers to learn the representation, and a single linear output layer, BNN-4-0.

Results

The authors compare

- The balanced linear regression model (BLR), i.e. $\Phi(x) = Wx$.
- A neural network with 4 layers to learn the representation, and a single linear output layer, BNN-4-0.
- A neural network with 2 layers to learn the representation, followed by 2 ReLU layers and a single layer. (BNN-2-2)

Results

The authors compare

- The balanced linear regression model (BLR), i.e. $\Phi(x) = Wx$.
- A neural network with 4 layers to learn the representation, and a single linear output layer, BNN-4-0.
- A neural network with 2 layers to learn the representation, followed by 2 ReLU layers and a single layer. (BNN-2-2)
- Different classical supervised learning regression algorithms like linear regression, doubly robust linear regression, BART, etc..

Results

The quantities measured to evaluate the models are

- The RMSE of the estimated individual treatment effect

$$\epsilon_{\text{ITE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\text{ITE}}(x_i)^2}$$

Results

The quantities measured to evaluate the models are

- The RMSE of the estimated individual treatment effect

$$\epsilon_{\text{ITE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\text{ITE}}(x_i)^2}$$

- the absolute error in estimated average treatment effect

$$\epsilon_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\text{ITE}}(x_i)$$

Results

The quantities measured to evaluate the models are

- The RMSE of the estimated individual treatment effect

$$\epsilon_{\text{ITE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\text{ITE}}(x_i)^2}$$

- the absolute error in estimated average treatment effect

$$\epsilon_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\text{ITE}}(x_i)$$

- The Precision in Estimation of Heterogeneous Effect,

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \Big(\hat{y}_1(x_i) - \hat{y}_0(x_i) - (Y_1(x_i) - Y_0(x_i))\Big)^2}$$

Results

Table 2. News. Results and standard errors for 50 repeated experiments. (Lower is better.) Proposed methods: BLR, BNN-4-0 and BNN-2-2. † (Chipman et al., 2010)

	ϵ_{ITE}	ϵ_{ATE}	PEHE
LINEAR OUTCOME			
OLS	3.1 ± 0.2	0.2 ± 0.0	3.3 ± 0.2
DOUBLY ROBUST	3.1 ± 0.2	0.2 ± 0.0	3.3 ± 0.2
LASSO + RIDGE	2.2 ± 0.1	0.6 ± 0.0	3.4 ± 0.2
BLR	2.2 ± 0.1	0.6 ± 0.0	3.3 ± 0.2
BNN-4-0	2.1 ± 0.0	0.3 ± 0.0	3.4 ± 0.2
NON-LINEAR OUTCOME			
NN-4	2.8 ± 0.0	1.1 ± 0.0	3.8 ± 0.2
BART[†]	5.8 ± 0.2	0.2 ± 0.0	3.2 ± 0.2
BNN-2-2	$\mathbf{2.0 \pm 0.0}$	0.3 ± 0.0	$\mathbf{2.0 \pm 0.1}$

IDHP Dataset

- A similar experiment was conducted on clinical data from the Infant Health and Devlopment Program (IDHP).

IDHP Dataset

- A similar experiment was conducted on clinical data from the Infant Health and Devlopment Program (IDHP).
- randomized treatment assignment

IDHP Dataset

- A similar experiment was conducted on clinical data from the Infant Health and Devlopment Program (IDHP).
- randomized treatment assignment
- introduced imbalance by removing a nonrandom portion of the treatment group.

IDHP Dataset

- A similar experiment was conducted on clinical data from the Infant Health and Devlopment Program (IDHP).
- randomized treatment assignment
- introduced imbalance by removing a nonrandom portion of the treatment group.

	ϵ_{ITE}	ϵ_{ATE}	PEHE
LINEAR OUTCOME			
OLS	4.6 ± 0.2	0.7 ± 0.0	5.8 ± 0.3
DOUBLY ROBUST	3.0 ± 0.1	0.2 ± 0.0	5.7 ± 0.3
LASSO + RIDGE	2.8 ± 0.1	0.2 ± 0.0	5.7 ± 0.2
BLR	2.8 ± 0.1	0.2 ± 0.0	5.7 ± 0.3
BNN-4-0	3.0 ± 0.0	0.3 ± 0.0	5.6 ± 0.3
NON-LINEAR OUTCOME			
NN-4	2.0 ± 0.0	0.5 ± 0.0	1.9 ± 0.1
BART†	2.1 ± 0.2	0.2 ± 0.0	1.7 ± 0.2
BNN-2-2	$\mathbf{1.7 \pm 0.0}$	0.3 ± 0.0	$\mathbf{1.6 \pm 0.1}$

IDHP Dataset

- A similar experiment was conducted on clinical data from the Infant Health and Devlopment Program (IDHP).
- randomized treatment assignment
- introduced imbalance by removing a nonrandom portion of the treatment group.

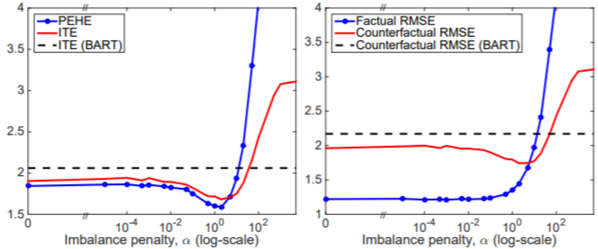


Figure 4. Error in estimated treatment effect (ITE, PEHE) and counterfactual response (RMSE) on the IHDP dataset. Sweep over α for the BNN-2-2 neural network model.

Conclusion

- This paper introduced models learning **balanced representations** for counterfactual inference, based on practical and theoretical evidence

Conclusion

- This paper introduced models learning **balanced representations** for counterfactual inference, based on practical and theoretical evidence

Some open questions

Conclusion

- This paper introduced models learning **balanced representations** for counterfactual inference, based on practical and theoretical evidence

Some open questions

- generalize this for more than 2 treatments

Conclusion

- This paper introduced models learning **balanced representations** for counterfactual inference, based on practical and theoretical evidence

Some open questions

- generalize this for more than 2 treatments
- allow for other distribution measures

Conclusion

- This paper introduced models learning **balanced representations** for counterfactual inference, based on practical and theoretical evidence

Some open questions

- generalize this for more than 2 treatments
- allow for other distribution measures
- allow for non-linear hypotheses

Any questions?

Thank you!

References

[1] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.