CounterFactual Regression with Importance Sampling Weights (CFR-ISW)

Negar Hassanpour (hassanpo@ualberta.ca) and Russell Greiner
Department of Computing Science, University of Alberta
Alberta Machine Intelligence Institute (Amii)
Edmonton, Alberta, Canada

1. Causal Inference from Observational Data

Goal: Finding a model that estimates the Individual Treatment Effect: $ITE(x) = y^1(x) - y^0(x)$

from an observational dataset in the form of $\{[x_i, t_i, y_i]\}_{i=1:n}$

with: x: personal features

→ *e.g.*, values of age, blood work, etc.

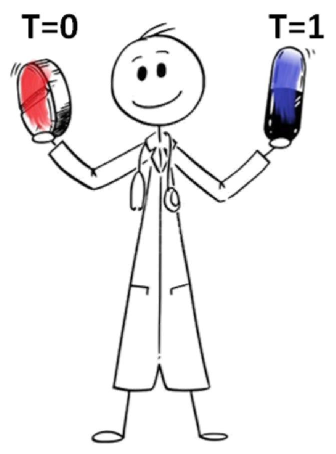
t: received treatment chosen from a set of options

→ *e.g.*, { 0: medication, 1: surgery}

y: the observed outcome after receiving the corresponding treatment

→ *e.g.*, survival time

ID	X			T	Y^0	Y^1
	Gender	Age	BMI			
Mr. Smith	Male	35	20	0	15	
Mr. Green	Male	22	32	0	22	
Ms. Jones	Female	20	23	1		31
…						

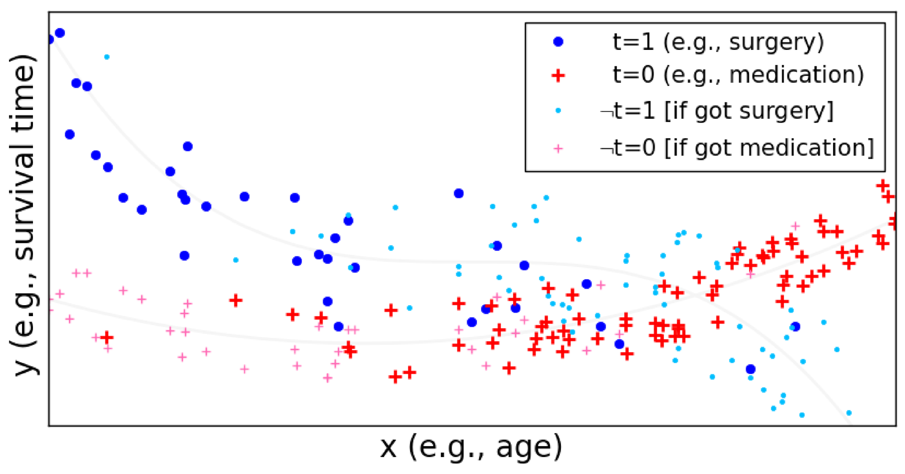


Challenges:

1. Partial information data. depending on the received treatment t, we observe (**factual outcome** y^t) either y^0 or y^1, but never both. The other outcome (**counterfactual outcome** $y^{\neg t}$) is unobservable.

2. Selection bias. both outcome y and the treatment t assignment are dependent on (some) context information x.

→ *e.g.*, **younger** {**older**} patients (part of x) are more likely to receive treatment t : **surgery** {**medication**} because they tend to have a **faster** {**complicated**} recovery (outcome y).



2. Related Work

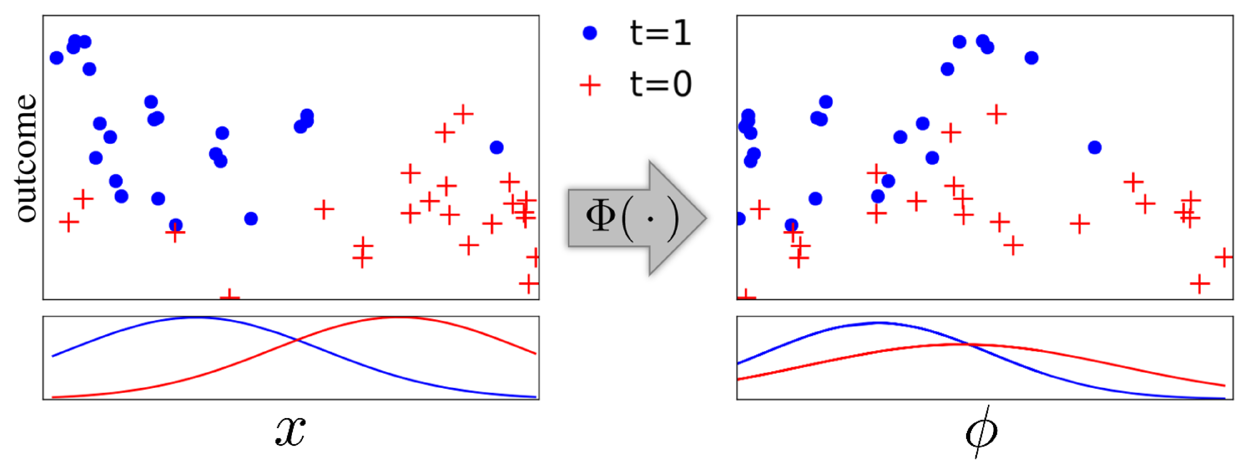
Representation Learning:

Reducing the selection bias by learning a common representation space $\phi(x)$ such that:

→ $\Pr(\phi(x) \mid t = 0)$ and $\Pr(\phi(x) \mid t = 1)$ are as close as possible to each other

→ provided that $\phi(x)$ retains enough information to accurately predict factual outcomes

→ by a learned hypothesis network for each treatment arm (*i.e.*, $h^t(x)$) that estimates the corresponding outcomes



[Shalit *et al.*, 2017]'s approach:

$$\arg\min_{h,\Phi} J(h,\Phi) = \arg\min_{h,\Phi} \Big[\frac{1}{n}\sum_{i=1}^{n} \omega_i \cdot L[\, h^{t_i}(\Phi(x_i)),\, y_i \,] + \alpha \cdot \mathrm{IPM}_G(\, \{\Phi(x_i)\}_{i:\ t_i=0},\ \{\Phi(x_i)\}_{i:\ t_i=1} \,) + \lambda \cdot \mathcal{R}(h) \Big]$$

where $L[\, h^{t_i}(\Phi(x_i)),\, y_i \,] = \left[h^{t_i}(\Phi(x_i)) - y_i \right]^2$ → **factual loss**

$$\omega_i = \frac{t_i}{u} + \frac{1-t_i}{(1-u)}, \quad \text{with} \quad u = \frac{1}{n}\sum_{i=1}^{n} t_i = \Pr(t=1)$$

$$= \frac{1}{\Pr(t_i)} = \frac{\Pr(t_i)}{\Pr(t_i)} + \frac{1-\Pr(t_i)}{\Pr(t_i)} = 1 + \boxed{\frac{\Pr(\neg t_i)}{\Pr(t_i)}}$$

$\mathrm{IPM}_G(\, \{\Phi(x_i)\}_{i:\ t_i=0},\ \{\Phi(x_i)\}_{i:\ t_i=1} \,)$ → Integral Probability Metric (IPM) is a measure of distance between two probability distributions (*e.g.*, Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012]), here between empirical $\Pr(\phi(x) \mid t = 0)$ and $\Pr(\phi(x) \mid t = 1)$ distributions

Once the model is trained, use it to predict y^0 and y^1, given as input a feature vector x

→ Gives the individual treatment effect $ITE(x) = y^1(x) - y^0(x)$ for any (novel) x

Get the paper @

Selected References:

- [Chipman *et al.*, 2010] Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 2010.
- [Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander Smola. A Kernel Two-sample Test. *JMLR*, 13(Mar):723–773, 2012.
- [Johansson *et al.*, 2018] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning Weighted Representations for Generalization Across Designs. *arXiv preprint arXiv:1802.08598*, 2018.
- [Shalit *et al.*, 2017] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. *ICML*, 2017.

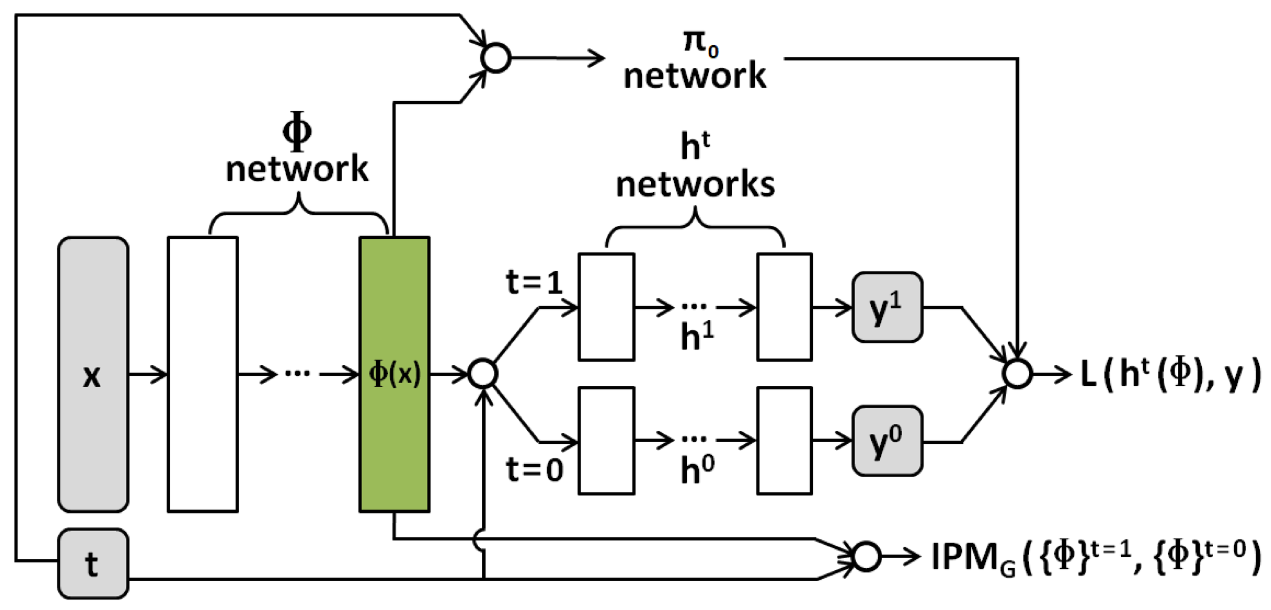
3. Proposed Method

Idea: Representation learning does not (and should not) remove selection bias completely

Importance Sampling Weighting on top of Representation Learning

→ Incorporate **context-aware weights** in the **factual loss** term.

Proposed Model Architecture:



$$\boxed{\frac{\Pr(\Phi(x_i) \mid \neg t_i)}{\Pr(\Phi(x_i) \mid t_i)}} = \frac{\dfrac{\pi(\neg t_i|\Phi(x_i)) \cdot \Pr(\Phi(x_i))}{\Pr(\neg t_i)}}{\dfrac{\pi(t_i|\Phi(x_i)) \cdot \Pr(\Phi(x_i))}{\Pr(t_i)}} = \frac{\Pr(t_i)}{1-\Pr(t_i)} \cdot \frac{1-\pi(t_i|\Phi(x_i))}{\pi(t_i|\Phi(x_i))}$$

where $\pi(t \mid \phi(x))$ is the probability of assigning treatment t given the context in ϕ representation space (a.k.a., **propensity score**).

→ We use Logistic Regression (LR) with parameters **[W, b]** to fit the propensity score function:

$$\pi(t \mid \Phi(x)) = \frac{1}{1 + e^{-(2t-1)(\Phi(x)\cdot \mathbf{W} + \mathbf{b})}}$$

and learn the parameters by minimizing: $\min_{\mathbf{W},\mathbf{b}} \frac{1}{n}\sum_{i=1}^{n} C[\, \mathbf{W}, \mathbf{b}, \Phi(x), t \,]$

where $C[\, \mathbf{W}, \mathbf{b}, \Phi(x), t \,] = -\log[\, \pi(t_i|\Phi(x_i)) \,]$

We try to solve this multi-objective optimization problem alternatively, repeating the two steps:

i. Minimize $J(h, \phi)$ to update the parameters of the representation ϕ and hypothesis h networks

ii. Minimize $C[\mathbf{W}, \mathbf{b}, \phi, t]$ with fixed h and ϕ parameters to update parameters of the propensity score function (*i.e.*, $\mathbf{W}$ and $\mathbf{b}$).

4. Experiments

Evaluation Criteria:

$$\epsilon_{ATE} = \left| \mathrm{ATE} - \widehat{\mathrm{ATE}} \right|$$

$$\mathrm{PEHE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{e}_i - e_i)^2}$$

$$\mathrm{ENoRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(1 - \frac{\hat{e}_i}{e_i}\right)^2} \quad \text{with} \quad \begin{cases} \hat{e}_i = \hat{y}_i^1 - \hat{y}_i^0 \\ e_i = y_i^1 - y_i^0 \end{cases}$$

where $\hat{y} = [\hat{y}^0, \hat{y}^1]$ indicates an outcome predicted by the trained model

Hyperparameter Selection: As counterfactual outcomes are **inherently unobservable**, it is not possible to use standard internal cross-validation to select hyperparameters (*e.g.*, α, λ etc.).

→ An estimation of the true effect is needed as a **surrogate** for the e term.

- Shalit *et al.* [2017] used the observed outcome $y_{j(i)}$ of the nearest neighbor (**1-NN**) in the x space (referred to as **1-NN**) in the alternative treatment group $t_{j(i)} = \neg t_i = 1 - t_i$
- We also considered outcome predicted by the Bayesian Additive Regression Trees (**BART**)

Benchmarks:

Infant Health and Development Program (**IHDP**)

- The observational study is sub-sampled from an RCT by removing a non-random subset of the treated population
- Includes 747 instances with 25 covariates

Atlantic Causal Inference Conference 2018 (**ACIC'18**)

- The x matrix is sub-sampled from the Linked Birth and Infant Death Data (LBIDD)
- The ys are synthesized by the challenge organizers
- Includes 100,000 instances with 177 features

5. Results

We compare performance of four methods:

- **1-NN**: One nearest neighbor method for finding the counterfactual outcomes
- **BART**: Bayesian Additive Regression Trees method [Chipman *et al.*, 2010]
- **CFR**: CounterFactual Regression method proposed in [Shalit *et al.*, 2017]
- **RCFR**: Re-weighted CFR [Johansson *et al.*, 2018]
- **CFR-ISW**: CounterFactual Regression with Importance Sampling Weights (our method)

Comparison of *ENoRMSE*, *PEHE*, and *bias of ATE* (lower is better) on the **IHDP** benchmark according to various hyperparameter selection criteria: **P1**: $PEHE_{1\text{-}NN}$, **PB**: $PEHE_{BART}$, and **EB**: $ENoRMSE_{BART}$

	Methods	ENoRMSE	PEHE	ϵ_{ATE}
	1-NN	24.6 (189)	4.85 (6.29)	0.67 (1.27)
	BART	2.13 (11.3)	1.57 (2.41)	0.22 (0.30)
	CFR†		0.78 (0.0)	0.31 (0.01)
	RCFR‡		0.65 (0.04)	
P1	CFR	2.65 (1.67)	0.88 (0.10)	0.20 (0.03)
P1	CFR-ISW	3.82 (3.17)	**0.77** (0.10)	**0.19** (0.03)
PB	CFR	1.87 (1.29)	0.65 (0.05)	0.21 (0.03)
PB	CFR-ISW	2.50 (2.05)	**0.55** (0.05)	**0.20** (0.03)
EB	CFR	1.18 (0.29)	0.84 (0.07)	0.23 (0.03)
EB	CFR-ISW	**0.88** (0.29)	**0.66** (0.05)	**0.16** (0.02)

Aggregated ENoRMSE (lower is better) on the **ACIC'18** benchmark. Hyperparameters for both CFR and CFR-ISW methods are selected according to $ENoRMSE_{BART}$

	Datasets	1-NN	BART	CFR	CFR-ISW
	All	54.56	9.35	5.43 (5.78)	**1.03** (0.27)
# Instances	1 k	66.70	73.66	7.08 (8.97)	1.54 (0.87)
# Instances	2.5 k	33.31	15.12	8.33 (14.78)	0.68 (0.31)
# Instances	5 k	31.89	8.15	2.00 (2.28)	0.88 (0.35)
# Instances	10 k	31.46	2.60	0.86 (1.00)	0.74 (0.39)
# Instances	25 k	19.47	1.27	0.85 (0.30)	1.00 (0.28)
# Instances	50 k	75.43	12.27	8.23 (8.63)	1.13 (0.23)

Entries in **bold** indicate significantly better performance (Welch's unpaired t-test with α=0.05)