CounterFactual Regression with Importance Sampling Weights (CFR-ISW)

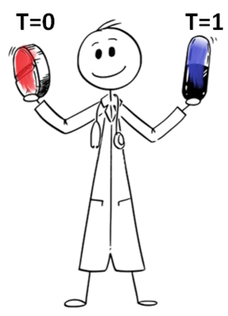
Negar Hassanpour* (hassanpo@ualberta.ca) and Russell Greiner*
*Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

1. Causal Inference from Observational Data

Ultimate 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 $\{ [x_i , t_i , y_i] \}_{i=1..n}$

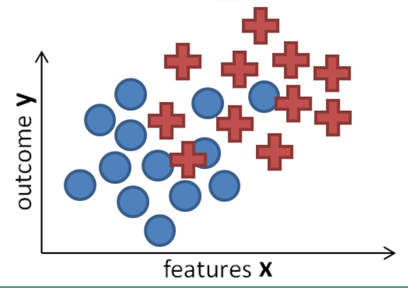
with:
- **x**: personal features
  - → *e.g.*, values of age, blood work, etc.
- **t**: received treatment chosen from a set of options
  - → *e.g.*, { 0: surgery , 1: medication }
- **y**: the observed outcome after receiving the corresponding treatment
  - → *e.g.*, survival time

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
...						



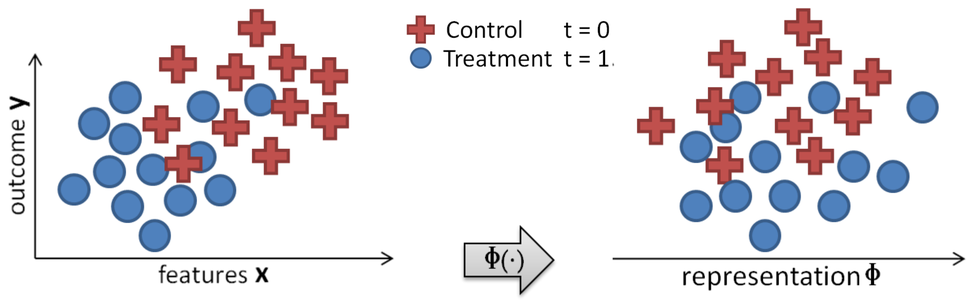
Challenges:

1. **Partial information data**. *i.e.*, 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^{-t}) is unobservable.
2. **Sample selection bias**. *i.e.*, 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**).

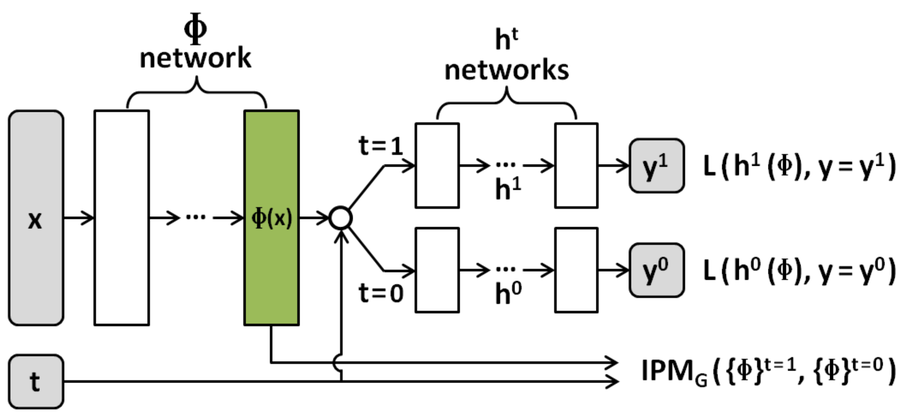


2. Shalit *et al.* (2017) Model Overview (CFR)

Their Goal: Reducing the sample 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 estimate the corresponding outcomes



Model Structure:



Objective:

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

→ regularization term

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

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

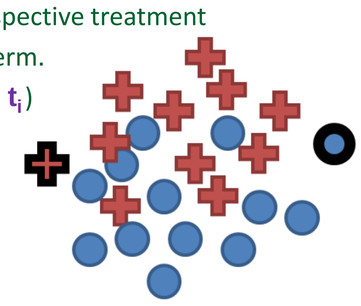
$$= \frac{1}{\Pr(t_i)} = \frac{\Pr(t_i)}{\Pr(t_i)} + \frac{1-\Pr(t_i)}{\Pr(t_i)} = 1 + \frac{\Pr(\neg t_i)}{\Pr(t_i)}$$

$\mathrm{IPM}_G(\{\Phi(x_i)\}_{i:\, t_i=0},\ \{\Phi(x_i)\}_{i:\, t_i=1})$ → Integral Probability Metric (IPM) is a measure of closeness between two probability distributions; *e.g.*, Maximum Mean Discrepancy (MMD) (Gretton *et al.*, 2012) or Wasserstein distance (Attouch *et al.*, 2014; Cuturi & Doucet, 2014).
Here, IPM measures the discrepancy between empirical $Pr(\Phi(x) \mid t=0)$ and $Pr(\Phi(x) \mid t=0)$ distributions

Once the model is trained, we can use it to predict y^1 and y^0, given as input a feature vector **x**
This will give us the individual treatment effect $ITE(x) = y^1(x) - y^0(x)$ for any (novel) **x**

3. Proposed Weighting Scheme (CFR-ISW)

Our Goal: Improving the accuracy of estimating **ITE** by incorporating the information extracted from the context of each instance $\Phi(x)$, in addition to its respective treatment **t**, to assign **sample-specific weights** in the **factual loss** term.
→ For example, if an instance x_i (with assigned treatment t_i) is **far** from other instances with the same assigned treatment (*e.g.*, samples in figure) then we force our outcome prediction network to learn this instance well.



Proposed weights:

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

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

We try to solve this multi-objective optimization problem that iteratively in two steps:
i. Minimize $J(h, \Phi)$ to update the parameters of the representation Φ and hypothesis **h** networks
ii. Minimize $C[W, b, \Phi, t]$ with fixed **h** and Φ parameters to update parameters of the propensity score function (*i.e.*, **W** and **b**).

4. Experiments

Evaluation Criteria:

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

where $\hat{y}$ 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 in the **x** space (referred to as **1-nn**) in the alternative treatment group $t_{j(i)} = \neg t_i = 1 - t_i$
- ❖ In addition to **1-nn**, we explored two alternatives:
  1. 1-nn in the Φ space; *i.e.*, $\mathbf{1\text{-nn}}_{\Phi}$
  2. outcome predicted by the Bayesian Additive Regression Trees (**BART**)

Synthetic Datasets: From the 2018 Atlantic Causal Inference Data Challenge
→ The **x** matrix is sampled from the Linked Birth and Infant Death Data (**LBIDD**)
- ❖ **100,000** instances, each with **177** features

→ **24** synthetic datasets were generated from LBIDD;
categorized into **6** groups in terms of the number of instances $\mathbf{n} \in \{1, 2.5, 5, 10, 25, 50\} \times 10^3$

5. Results

We compare performance of the following four different methods in terms of ENoRMSE:
- **1-nn**: One nearest neighbor method for finding the counterfactual outcomes
- **BART**: Bayesian Additive Regression Trees method (Chipman *et al.*, 2010) for finding the ITE
- **CFR**: CounterFactual Regression method proposed in (Shalit *et al.*, 2017) for which the best set of hyperparameters is determined based on $\mathrm{ENoRMSE}_{BART}$
- **CFR-ISW**: CounterFactual Regression with Importance Sampling Weights (*i.e.*, the proposed method)

Tables report the aggregated ENoRMSE (lower is better). The entry in **bold** is the best for each row.

Comparison of various ITE estimation methods against the proposed CFR-ISW

	Datasets	1-NN	BART	CFR	CFR-ISW
	ALL	75.32	20.03	8.92	**1.07**
# Instances	**1** k	86.44	141.58	10.51	**1.72**
	2.5 k	47.45	23.35	15.27	**0.73**
	5 k	38.04	9.51	2.81	**0.93**
	10 k	40.25	2.96	1.22	**0.81**
	25 k	25.69	1.52	**0.89**	1.03
	50 k	94.45	16.13	11.12	**1.14**

Hyperparameter selection methods: $\mathrm{ENoRMSE}_{1\text{-nn}}$ vs. $\mathrm{ENoRMSE}_{BART}$

		CFR		CFR-ISW	
	Datasets	1-NN	BART	1-NN	BART
	ALL	15.04	**8.92**	5.23	**1.07**
# Instances	**1** k	**8.08**	10.51	2.21	**1.72**
	2.5 k	36.33	**15.27**	0.82	**0.73**
	5 k	5.79	**2.81**	1.05	**0.93**
	10 k	1.45	**1.22**	15.11	**0.81**
	25 k	1.01	**0.89**	1.28	**1.03**
	50 k	17.98	**11.12**	**1.11**	1.14

Selected References:

- **(Chipman *et al.*, 2010)** Chipman, Hugh A, George, Edward I, and McCulloch, Robert E. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 2010.
- **(Shalit *et al.*, 2017)** Shalit, Uri, Johansson, Fredrik D., and Sontag, David. "Estimating individual treatment effect: generalization bounds and algorithms." *In Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3076–3085, 2017.

Get the paper @