Review of “Learning Representations for Counterfactual Inference”

Johansson et al. (2016)

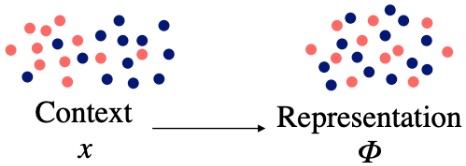
Suyeol Yun

Massachusetts Institute of Technology

June 30, 2022

What is "Representation Learning"?

- Representation learning is to find effective transformation of input data that makes it easier to perform a task like classification or prediction.
- How to perform representation learning for causal inference (CI) in observational studies to estimate the causal quantity of interest?
- How's the representation learning related to CI?
- Which kind of property is desirable for the representation for CI?
- Balancing property - what else?



Causal Inference and Domain Adaptation

- Domain adaptation is a sub-field within machine learning that aims to cope with the situation where the training and the test data fall from different distributions
- Domain adaptation aligns the disparity between domains such that the trained model can be generalized into the domain of interest
- Johansson et al. (2016) view CI as DA and suggests a objective (loss) function to learn effective representation that well generalizes the trained model in factual domain to counterfactual domain

Problem Setup

- Binary action set $\mathcal{T} = \{0, 1\}$
- Causal quantity of interest: $\text{ITE}(x) = Y_1(x) - Y_0(x)$
- Given n samples $\left\{\left(x_i, t_i, y_i^F\right)\right\}_{i=1}^n$, where $y_i^F = t_i \cdot Y_1\left(x_i\right) + \left(1 - t_i\right) Y_0\left(x_i\right)$, learn a function $h : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$ such that $h\left(x_i, t_i\right) \approx y_i^F$
- $\hat{\text{ITE}}\left(x_i\right) = \begin{cases} y_i^F - h\left(x_i, 1 - t_i\right), & t_i = 1 \\ h\left(x_i, 1 - t_i\right) - y_i^F, & t_i = 0 \end{cases}$
- $\hat{P}^F = \{(x_i, t_i)\}_{i=1}^n$, $\hat{P}^{CF} = \{(x_i, 1 - t_i)\}_{i=1}^n$
- P^F and P^{CF} need not to be equal- the problem of causal inference requires inference over a different distribution than the one from which samples are given
- In machine learning terms, this means that the feature distribution of the test set differs from that of the train set. This is a case of *covariate shift*, which is a special case of domain adaptation.

Balancing Objective

- Johansson et al. (2016) proposes a method to jointly learns a representation $\Phi : \mathcal{X} \to \mathbb{R}^d$ and $h : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}$ such that the learned representation minimizes the following objective:

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

- Let $j(i) \in \arg\min_{j \in \{1 \ldots n\} \text{ s.t. } t_j = 1 - t_i} \mathrm{d}\left(x_j, x_i\right)$ be the nearest neighbor of x_i among the group that received
- $\operatorname{disc}_{\mathcal{H}} \left(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF} \right) =$

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

Components of Objective Function

- (1) Enabling low-error prediction of the observed outcomes over the factual representation
- (2) Make the sampling distributions of factual and counterfactual to be similar
- (3) Enabling lowe-rror prediction of unobserved counterfactuals by taking into account relevant factual outcomes

$$
\begin{aligned}
B_{\mathcal{H},\alpha,\gamma}(\Phi, h) = & \frac{1}{n}\sum_{i=1}^{n}\left|h\left(\Phi\left(x_i\right), t_i\right) - y_i^F\right| && (1)\\
& + \alpha \operatorname{disc}\mathcal{H}\left(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}\right) && (2)\\
& + \frac{\gamma}{n}\sum_{i=1}^{n}\left|h\left(\Phi\left(x_i\right), 1 - t_i\right) - y_{j(i)}^F\right| && (3)
\end{aligned}
$$

Theoretical Justification of Objective Function

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

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

$$+\,\mathrm{disc}_{\mathcal{H}_l}\left(\hat{P}_\Phi^F,\hat{P}_\Phi^{CF}\right) \tag{5}$$

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

- $\mathcal{H}_l\subset\mathbb{R}^{d+1}$ be the space of linear functions
- $\mathcal{L}_{P^{CF}}(\beta)=\mathbb{E}_{(x,t,y)\sim P^{CF}}[L(\beta(x,t),y)]$ be the expected loss of β over distribution P^{CF}.
- $\hat{\beta}^F(\Phi)=\arg\min_{\beta\in\mathcal{H}_l}\mathcal{L}_{\hat{P}_\Phi^F}(\beta)+\lambda\|\beta\|_2^2$,
  $\hat{\beta}^{CF}(\Phi)=\arg\min_{\beta\in\mathcal{H}_l}\mathcal{L}_{\hat{P}_\Phi^{CF}}(\beta)+\lambda\|\beta\|_2^2$

Generalization Bound and Intepretation

- $\frac{\lambda}{\mu r}\left(\mathcal{L}_{P^{CF}}\left(\hat{\beta}^F(\Phi)\right) - \mathcal{L}_{P^{CF}}\left(\hat{\beta}^{CF}(\Phi)\right)\right)^2 \le$

$$\min_{h \in \mathcal{H}_l} \frac{1}{n} \sum_{i=1}^{n} \left(\left| \hat{y}_i^F(\Phi, h) - y_i^F \right| + \left| \hat{y}_i^{CF}(\Phi, h) - y_{j(i)}^F \right| \right) \tag{7}$$

$$+ \operatorname{disc}_{\mathcal{H}_l}\left(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}\right) \tag{8}$$

$$+ \frac{K_0}{n} \sum_{i:t_i=1} \mathrm{d}_{i,j(i)} + \frac{K_1}{n} \sum_{i:t_i=0} \mathrm{d}_{i,j(i)} \tag{9}$$

- Minimizing the bound make the estimator fit on factual distribution to generalize better over counterfactual distribution
- It's important to find Φ such that minimizes the bound
- Φ such that attains low prediction error, less discrepancy between representation space of factual and counterfactual
- The GB holds regardless of how Φ is obtained, e.g. if Φ is a neural net - it still holds.

References

- Johansson et al., Learning Representations for Counterfactual Inference (2016)
- Mansour et al., Domain adaptation: Learning bounds and algorithms (2009)