When interpretability meets causality: Causal Interpretability[1]

Zirui Yan

Department of Electrical, Computer, and Systems Engineering

Rensselaer Polytechnic Institute

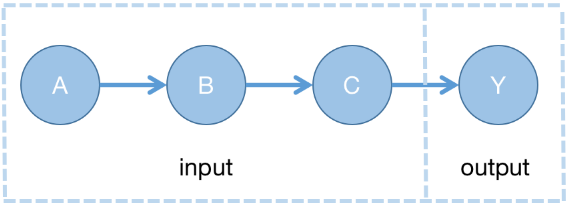
email: yanz11@rpi.edu

webpage: ziruiyan.github.io

April 26, 2022

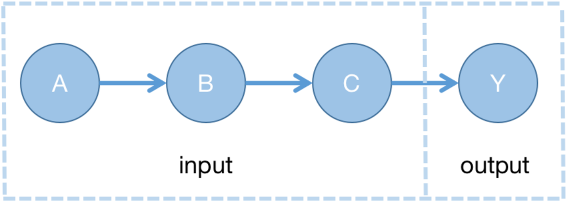
[1] https://github.com/ZiruiYan/Causal-Interpretability

Introduction



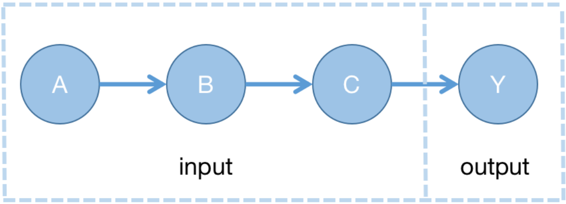
- **Problem:** Assuming the inputs and output admits some causality. If we do intervention on some variables, how output Y will change.
- **Related work:**
  - There are some interpretability methods such as Quantitative input influence (QII), Accumulated Local Effects (ALE) plot. They may consider the correlation but ignore the causality.
  - Researchers in causality community has investigate into this problem, but they focus on the binary variable and didn't link it to interpretability.

Simple Example



- A: SAT Score $\rightarrow$ B: College Addimision $\rightarrow$ C: Took TML course? $\rightarrow$ D: Got job at IBM?
- A: 1500 $\rightarrow$ B: RPI $\rightarrow$ C: Yes $\rightarrow$ D: Yes
- A: 1500 $\rightarrow$ B: Stanford $\rightarrow$ C: Yes $\rightarrow$ D: Yes

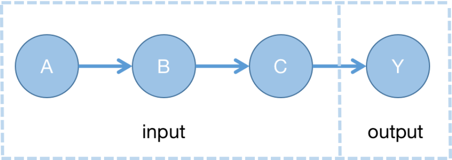
Simple Example



- A: SAT Score $\rightarrow$ B: College Addimision $\rightarrow$ C: Took TML course? $\rightarrow$ D: Got job at IBM?
- A: 1500 $\rightarrow$ B: RPI $\rightarrow$ C: Yes $\rightarrow$ D: Yes
- A: 1500 $\rightarrow$ B: Stanford $\rightarrow$ C: Yes $\rightarrow$ D: Yes
- A: 1500 $\rightarrow$ B: Stanford $\rightarrow$ C: No $\rightarrow$ D: No

Background

- Assume we know the **Directed acyclic graph (DAG)** between the variables
  Edges represent causal relations between the varialbes
  Output Y is terminal nodes



- Assume that there is **no latent variables**

- $do(X = x)$: Change X to become x, and affect the descendants of X

Interpretability Methods

- Assume we interested in the influence of $\boldsymbol{X}_s$ on function $f(\boldsymbol{X}_s, \boldsymbol{X}_c)$ and $\boldsymbol{X}_c$ are other inputs.
- Individual Conditional Expectation (ICE) for individual i with input $(\boldsymbol{x}_s^{(i)}, \boldsymbol{x}_c^{(i)})$[3]

$$f_{s,\mathrm{ICE}}^{(i)}(\boldsymbol{x}_s) = f\left(\boldsymbol{x}_s, \boldsymbol{x}_c^{(i)}\right) \quad . \tag{1}$$

- Partial Dependence Plot (PDP)[4]

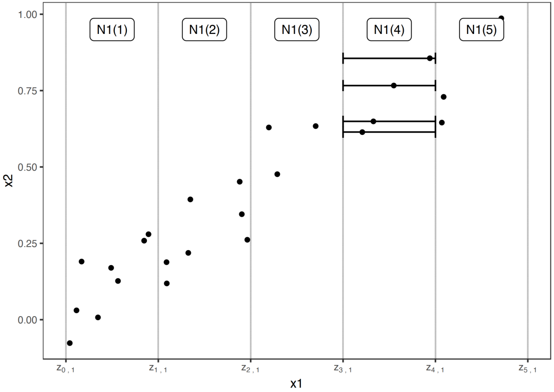
$$f_{s,\mathrm{PDP}}(\boldsymbol{x}_s) = \mathbb{E}_{\boldsymbol{X}_c}\left[f(\boldsymbol{x}_s, \boldsymbol{X}_c)\right] = \int f(\boldsymbol{x}_s, \boldsymbol{X}_c)\,\mathrm{d}\mathbb{P}(\boldsymbol{X}_c) \quad . \tag{2}$$

[3] Goldstein, Alex, et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. journal of Computational and Graphical Statistics (2015)

[4] Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics (2001)

- Accumulated Local Effects (ALE)[5]

$$f_{s,\mathrm{ALE}}\left(x_s\right)=\int_{x_0}^{x_s}\mathbb{E}_{\mathbf{X}_c|\mathbf{X}_s=z_s}\left[f\left(\mathbf{X}_s,\mathbf{X}_c\right)|\mathbf{X}_s=\mathbf{z}_s\right]\mathrm{d}\mathbf{z}_s-c\ , \tag{3}$$



[6]

[5] Apley, Daniel W., and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society (2020)

[6] Sec 8.2 in Molnar, Christoph. Interpretable machine learning. Lulu.com (2020)

Proposed Causal Interpretability Methods

- Individual Intervention Expectation (IIE) for individual i with input $(\boldsymbol{x}_s^{(i)}, \boldsymbol{x}_c^{(i)})$

$$f_{s,\mathrm{IIE}}^{(i)}(\boldsymbol{x}_s) = f\left(\boldsymbol{X}_s, \boldsymbol{X}_c \middle| (\boldsymbol{X}_s, \boldsymbol{X}_c) = (\boldsymbol{x}_s^{(i)}, \boldsymbol{x}_c^{(i)}), do(\boldsymbol{X}_s = \boldsymbol{x}_s)\right) , \tag{4}$$

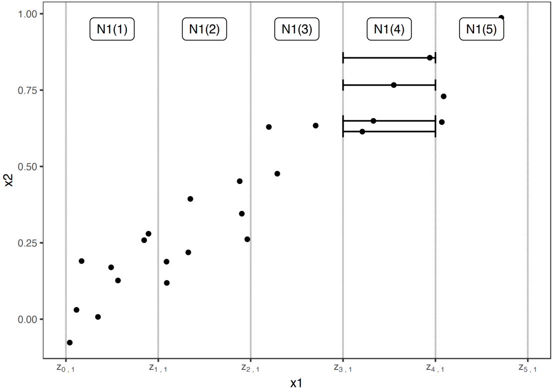
- Causal Partial Dependence Plot (CPDP)

$$f_{s,\mathrm{CPDP}}^{(i)}(\boldsymbol{x}_s) = \mathbb{E}_{(\boldsymbol{x}_s, \boldsymbol{x}_c)} f\left(\boldsymbol{X}_s, \boldsymbol{X}_c \middle| do(\boldsymbol{X}_s = \boldsymbol{x}_s)\right) , \tag{5}$$

Proposed Causal Interpretability Methods

- Accumulated Local Casual Effects (ALCE)

$$f_{S,\mathrm{ALCE}}(x_s) = \int_{x_0}^{x_s} \mathbb{E}_{\boldsymbol{X}_c|\boldsymbol{X}_s=\boldsymbol{z}_s}\left[f(\boldsymbol{X}_s, \boldsymbol{X}_c)|do(\boldsymbol{X}_s = \boldsymbol{z}_s)\right] \mathrm{d}\boldsymbol{z}_s - c. \quad (6)$$



[7]

[7] Sec 8.2 in Molnar, Christoph. Interpretable machine learning. Lulu.com (2020)

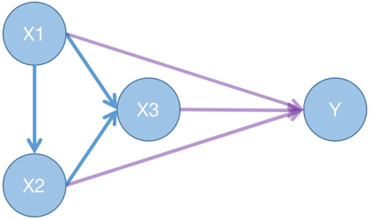
Rensselaer

Experiment setting: synthetic

- Assume we only know the causal DAG and the class of function, where $\epsilon_i \sim \mathcal{N}(0, 0.01)$

$$
\begin{aligned}
X_1 &\leftarrow U(-10, 10), \\
X_2 &\leftarrow 10\sigma(X_1) - 5 + \epsilon_2, \\
X_3 &\leftarrow 10\sigma(-X_1 - X_2) - 5 + \epsilon_3, \\
Y &\leftarrow 10\sigma(X_1 + X_2 - X_3) - 5 + \epsilon_4,
\end{aligned}
\tag{7}
$$

where σ is the sigmoid function.



- We are interested in variable X_2
- Number of samples is 50000

Experiment setting: invariant noise

- For individual i

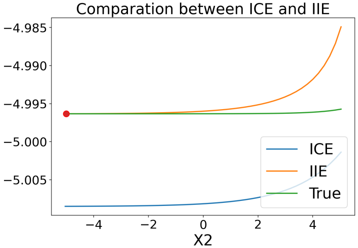
$$
\begin{aligned}
x_2^{(i)} &\leftarrow 10\sigma(x_1^{(i)}) - 5 + \epsilon_2^{(i)}, \\
x_3^{(i)} &\leftarrow 10\sigma(-x_1^{(i)} - x_2^{(i)}) - 5 + \epsilon_3^{(i)}, \\
y^{(i)} &\leftarrow 10\sigma(x_1^{(i)} + x_2^{(i)} - x_3^{(i)}) - 5 + \epsilon_4^{(i)},
\end{aligned} \tag{8}
$$

- For machine learning

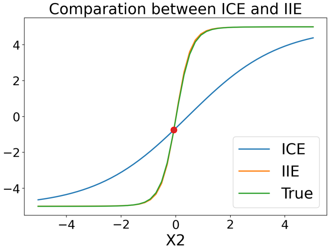
$$
\begin{aligned}
\hat{\epsilon}_2^{(i)} &= x_2^{(i)} - \hat{x}_2^{(i)}, \\
\hat{\epsilon}_3^{(i)} &= x_3^{(i)} - \hat{x}_3^{(i)},, \\
\hat{\epsilon}_4^{(i)} &= y^{(i)} - \hat{y}^{(i)},
\end{aligned} \tag{9}
$$

Rensselaer

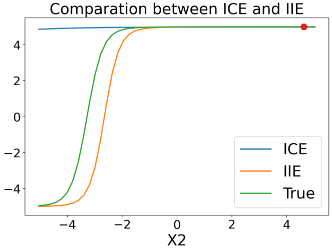
Experiment results: synthetic



Individual 1



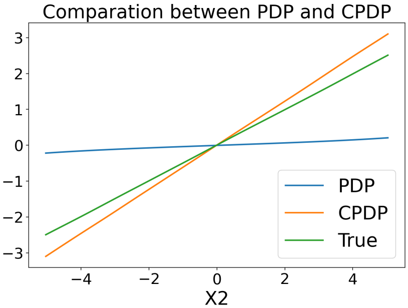
Individual 2



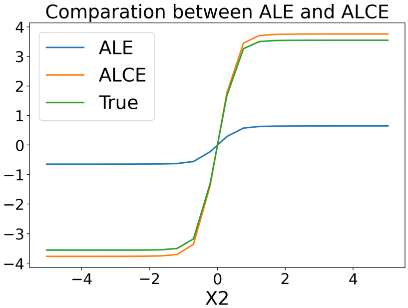
Individual 201

Figure. Comparation between ICE and IIE

Experiment results: synthetic



Comparation between PDP and CPDP

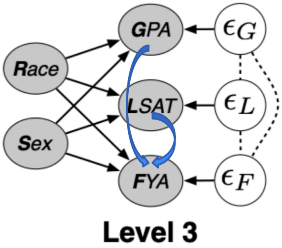


Comparation between ALE and ALCE

Figure. CPDP and ALCE plots

Experiment: real dataset

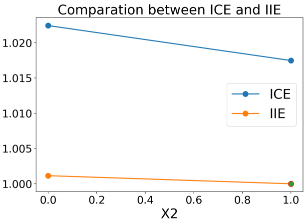
- Law School Success[8]



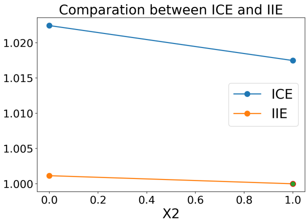
$$\begin{cases} \mathrm{GPA} = b_G + w_G^R R + w_G^S S + \epsilon_G, & \epsilon_G \sim p(\epsilon_G) \\ \mathrm{LSAT} = b_L + w_L^R R + w_L^S S + \epsilon_L, & \epsilon_L \sim p(\epsilon_L) \\ \mathrm{FYA} = b_F + w_F^R R + w_F^S S + w_F^G \mathrm{GPA} + w_F^L \mathrm{LSAT} + \epsilon_F, & \epsilon_F \sim p(\epsilon_F) \end{cases} \tag{10}$$

[8] similar setting as in Kusner, Matt J., et al. Counterfactual fairness. NeurIPS (2017).

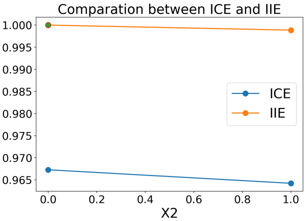
Experiment: real dataset



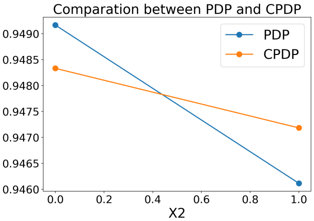
ICE and IIE for female 1



ICE and IIE for female 2



ICE and IIE for male 1



PDP and CPDP plots

Conclusion and Future work

Conclusion

- Traditional interpretability methods will meet problem when we are interested in causality
- The proposed causal interpretable methods works better in this scenario

Problem and Future work

- DAG is hard to identify based on observational dataset
  1. how to use the intervention dataset
  2. causal interpretible methods can help to identify the DAG
- Deep neural network have bad local minimum
  1. causal interpretible methods can only interpret the model itself
  2. causal interpretible methods can help to evaluation the model
- Analysis of the performance of causal interpretible methods (e.g. in high dimension)