Causal Perspectives on “Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models” by Apley and Zhu

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

September 23, 2024, RSS Journal Webinar

Acknowledgement

- Engineering and Physical Sciences Research Council (EPSRC grant EP/V049968/1).
- Dr Zijun Gao (University of Southern California).

Two problems

1. Explaining a machine
   - Notational change: Consider $\hat{Y} = \hat{f}(\boldsymbol{X})$ with two inputs $\boldsymbol{X} = (X_1, X_2)$.
   - $\hat{f}$ is treated as fixed.
2. Explaining the real world
   - Let Y be original response variable used by the regression algorithm.
   - $Y(x_1)$ is potential outcome under the intervention $X_1 = x_1$.
   - Might be helpful: Y is generated "causally" by $Y = f(X_1, X_2, E)$, so $Y(x_1) = f(x_1, X_2, E)$.

Causal perspective on the PD plot

- The **partial dependence (PD) plot** of Friedman (2001) shows

$$\hat{f}_{1,\mathrm{PD}}(x_1) = \mathbb{E}[\hat{f}(x_1, X_2)] = \int \hat{f}(x_1, x_2) p_2(x_2) \mathrm{d}x_2.$$

- Zhao and Hastie (2021) point out that this coincides with the **confounder adjustment** formula in causal inference:

$$\mathbb{E}[Y(x_1)] = \mathbb{E}\{\mathbb{E}[Y \mid X_1 = x_1, X_2]\} = \int \mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2] p_2(x_2) \mathrm{d}x_2. \tag{1}$$

- The formula (1) requries some **causal identification assumptions**:
  - Most important is **no unmeasured confounding** or **ignorability**:

$$Y(x_1) \perp\!\!\!\perp X_1 \mid X_2.$$

  - Alternative graphical condition: **back-door criterion** (Pearl 1995).
  - $+$ **consistency/SUTVA** $+$ **positivity/overlap** (closely related to the extrapolation problem).
- **Ignorability** is automatically satisfied if $\hat{Y}(x_1) = \hat{f}(x_1, \boldsymbol{X}_2)$.

Causal perspective on the ALE plot

- The **accumulated local effects** (ALE) plot shows $\hat{f}_{1,\mathsf{ALE}}(x_1)$ for which

$$\frac{\mathrm{d}\hat{f}_{1,\mathsf{ALE}}(x_1)}{\mathrm{d}x_1} = \mathbb{E}\left[\frac{\mathrm{d}\hat{f}(X_1, X_2)}{\mathrm{d}X_1} \mid X_1 = x_1\right].$$

- When $X_1 \in \{0, 1\}$, this can be replaced by

$$\hat{f}_{1,\mathsf{ALE}}(1) - \hat{f}_{1,\mathsf{ALE}}(0) = \mathbb{E}\left[\hat{f}(1, X_2) - \hat{f}(0, X_2) \mid X_1 = 0\right] = \int \{\hat{f}(1, x_2) - \hat{f}(0, x_2)\} p_{2|1}(x_2 \mid 0)\mathrm{d}x_2.$$

- This coincides with the formula for **natural direct effect** (Pearl 2001; Robins and Greenland 1992): let $Y(x_1, x_2)$ and $X_2(x_1)$ be potential outcomes and

$$\mathsf{NDE} := \mathbb{E}[Y(1, X_2(0)) - Y(0, X_2(0))] = \mathbb{E}\{\mathbb{E}[Y \mid X_1 = 1, X_2] - \mathbb{E}[Y \mid X_1 = 0, X_2] \mid X_1 = 0\}. \tag{2}$$

- For continuous X_1, ALE shows the accumulated local natural directed effect.

Identification of the natural direct effect

$$\text{NDE} := \mathbb{E}[Y(1, X_2(0)) - Y(0, X_2(0))] = \mathbb{E}\left\{\mathbb{E}[Y \mid X_1 = 1, X_2] - \mathbb{E}[Y \mid X_1 = 0, X_2] \mid X_1 = 0\right\}.$$

Main assumptions for this formula

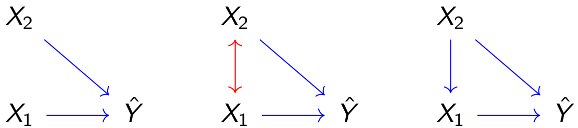
1. **No treatment-mediator confounding**: $X_2(x_1) \perp\!\!\!\perp X_1$ for all x_1.
2. **No treatment-outcome confounding**: $Y(x_1, x_2) \perp\!\!\!\perp X_1$ for all x_1, x_2.
3. **No mediator-outcome confounding**: $Y(x_1, x_2) \perp\!\!\!\perp X_2(x_1') \mid X_1$ for all x_1, x_2, x_1'.

- The last two assumptions are automatically satisfied if $\hat{Y}(x_1, x_2) = \hat{f}(x_1, x_2)$.
- But the first assumption may or may not be true.

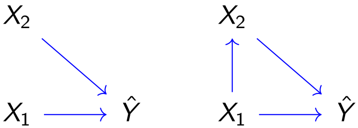
Graphs

- Directed edge $\rightarrow$ means direct causal influence.
- Bidirected edge $\leftrightarrow$ means exogenous correlation.

Causal interpretation of PD plot is valid in



Causal interpretation of ALE plot is valid in



Why explanations need to be causal?

Example 1: Multiplication of random signs

- X_1, X_2 are i.i.d. Radamacher random variables: $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$.
- Consider $\hat{Y} = \hat{f}(X_1, X_2) = X_2$.
  - Surely the "explanability" of X_1 to $\hat{Y}$ should be **zero**?
- Suppose X_2 is generated by $X_2 = X_1 X_2'$, where X_2' is another Radamacher variable independent of X_1. So $\hat{Y} = \hat{g}(X_1, X_2') = X_1 X_2'$.
  - Surely the "explanability" of X_1 to $\hat{Y}$ should be **non-zero**?
- This paradox arises because **same variable does not imply same potential outcome**: $\hat{Y}(x_1) = X_2$ in the first setting and $\hat{Y}(x_1) = x_1 X_2'$ in the second setting.

Why explanations need to be causal?

Example 2: Weatherman and Sun Wukong (the Monkey King)

- My little boy watches BBC every day and notices that the rain forecast for London has been correct in the last 5 days.[1]
- One day, he asked me: daddy, is the weatherman Sun Wukong from *Journey to the West*?

[1] If you ever lived in the UK, you will then know that this story is fictional.

Why explanations need to be causal?

Example 2: Weatherman and Sun Wukong (the Monkey King)

- My little boy watches BBC every day and notices that the rain forecast for London has been correct in the last 5 days.[1]
- One day, he asked me: daddy, is the weatherman Sun Wukong from *Journey to the West*?

[1] If you ever lived in the UK, you will then know that this story is fictional.

Why explanations need to be causal?

Example 2: Weatherman and Sun Wukong (the Monkey King)

- My little boy watches BBC every day and notices that the rain forecast for London has been correct in the last 5 days.[1]
- One day, he asked me: daddy, is the weatherman Sun Wukong from *Journey to the West*?

$$\begin{tikzcd}
\text{Weather model (unknown to my son)} \arrow[d] & \\
\text{Weatherman} \arrow[r] & \text{Weather forecast}
\end{tikzcd}
\qquad
\begin{tikzcd}
\text{Sun Wukong} \arrow[d] & \\
\text{East Dragon King} \arrow[r] & \text{Rain}
\end{tikzcd}$$

- If weatherman faithfully reports the forecast of the weather model, a PD or ALE plot will not be able to distinguish their contributions.

[1] If you ever lived in the UK, you will then know that this story is fictional.

Take-home messages

- PD and ALE plots have causal interpretations.
- But these causal interpretations requrie additional assumptions about the causal relationship between the predictors.
- Explanations of black-box machine are really meaningful only when they are causal.

References

Friedman, Jerome H. (Oct. 2001). "Greedy Function Approximation: A Gradient Boosting Machine.". In: *Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451.

Pearl, Judea (Dec. 1995). "Causal Diagrams for Empirical Research". In: *Biometrika* 82.4, pp. 669–688. DOI: 10.1093/biomet/82.4.669. (Visited on 06/22/2023).

— (Aug. 2001). "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 411–420. (Visited on 09/21/2024).

Robins, James M. and Sander Greenland (Mar. 1992). "Identifiability and Exchangeability for Direct and Indirect Effects". In: *Epidemiology* 3.2, p. 143. (Visited on 09/21/2024).

Zhao, Qingyuan and Trevor Hastie (2021). "Causal Interpretations of Black-Box Models". In: *Journal of Business & Economic Statistics* 39.272-281, pp. 1–10. DOI: 10.1080/07350015.2019.1624293.