On Measuring Causal Contribution via *do*-interventions

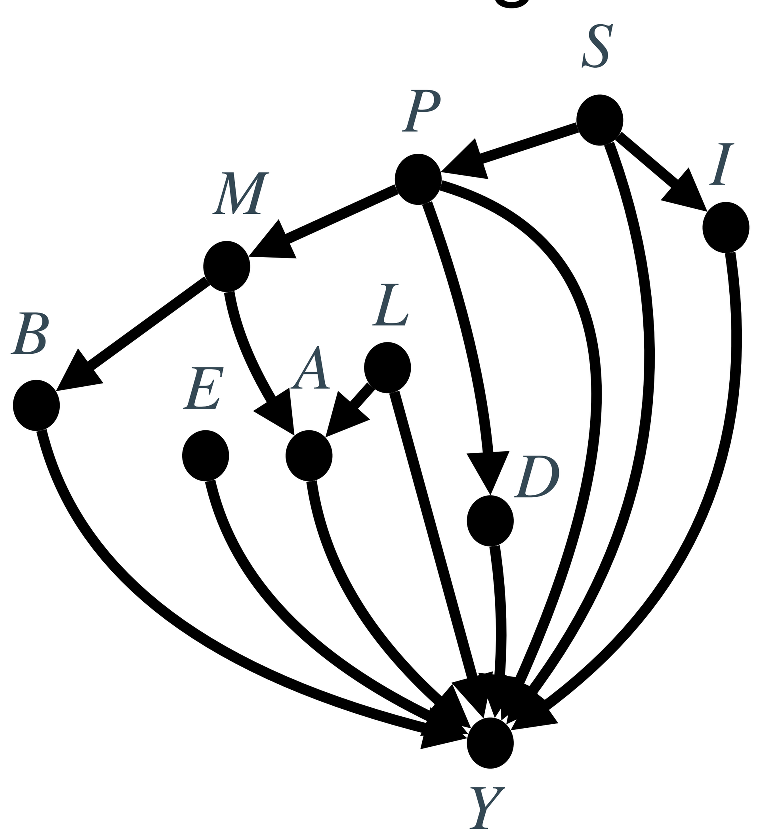
Yonghan Jung — Purdue University
Shiva Prasad Kasiviswanathan — amazon
Jin Tian — Iowa State University
Dominik Janzing — amazon
Patrick Blöbaum — amazon
Elias Bareinboim — Columbia University in the City of New York

ICML — International Conference On Machine Learning

Summary

In this work, we propose *do*-Shapley values, a method for measuring the contribution of each input feature based on its causal impact on the outcome.

Measuring Causal Contribution



S: Sales calls
P: Production needs
I: Interaction with customers
M: Monthly usages
D: Discounts
L: Last upgrade
E: Economic factors
A: Ad spend
B: Bugs reported
Y: Customer Retention
$\mathbf{V} := \{S, P, I, M, D, L, E, A, B\}$

Task: Given the causal effect of a hypothetical intervention $\mathbb{E}[Y \mid do(\mathbf{v})]$, our goal is to measure the contribution of each realization $v_i \in \mathbf{v}$ based on the causal impact of v_i on Y.

Application to Interpreting ML models: If the outcome Y is an ML model output; i.e., $Y = f(\mathbf{V})$ where f is the ML model, then the task reduces to measuring the contribution of $v_i \in \mathbf{v}$ of the ML outcome $f(\mathbf{v})$.

Comparing with other methods

- Existing methods for attribution contribution of input features measure the contribution based on the correlation, instead of causation, of inputs to the outcome; e.g., [1].
- Other methods considering causality assume that a model for target Y is *accessible* and can be generated for arbitrary input features [2,3].
- The proposed method measures the contribution based on identifiable causal effects without relying on accessibility.
  - Note Intrinsic Causal Contribution (ICC) [4] captures a fundamentally different kind of contribution because it describes the causal contribution of a node that has not been inherited from its ancestors.

Axiomatic Characterization

***do*-Shapley:** We propose the *do*-Shapley as a method for measuring the contribution of each input feature based on its causal impact on the outcome.

$$\phi_{v_i} := \frac{1}{n} \sum_{S \subseteq [n]} \binom{n-1}{|S|}^{-1} \{\mathbb{E}[Y \mid do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \mid do(\mathbf{v}_S)]\}$$

where n is the number of variables and $[n] := \{1, 2, \cdots, n\}$.

Theorem: Axiomatic Characterization of do-Shapley

The do-Shapley is a *unique* contribution method satisfying the following properties:

1. **Assignment**: Its sum equals to $\mathbb{E}[Y \mid do(\mathbf{v})] = \sum_{v_i \in \mathbf{v}} \phi_{v_i}$.
2. **Causal Irrelevance**: $\phi_{v_i} = 0$, if $\mathbb{E}[Y \mid do(v_i, \mathbf{v}_S)] = \mathbb{E}[Y \mid do(v_i', \mathbf{v}_S)]$ for all $\mathbf{V}_S \subseteq \mathbf{V}$ (i.e., v_i is causally irrelevant to Y).
3. **Causal Symmetry**: $\phi_{v_i} = \phi_{v_j}$ if $\mathbb{E}[Y \mid do(v_i), do(\mathbf{v}_S)] = \mathbb{E}[Y \mid do(v_j), do(\mathbf{v}_S)]$ for all $\mathbf{V}_S \subseteq \mathbf{V}$ (i.e., v_i, v_j have the same causal explanatory power).
4. **Linearity**: ϕ_{v_i} is a linear function of $\mathbb{E}[Y \mid do(\mathbf{v}_S)]$ $\forall \mathbf{V}_S \subseteq \mathbf{V}$.

do-Shapley Identification

- To determine the identifiability of the do-Shapley, the identifiability of $\mathbb{E}[Y \mid do(\mathbf{v}_S)]$ $\forall S \subseteq [n]$ should be determined, which takes exponential time.
- We introduce graphical criteria where the identifiability of the do-Shapley can be determined efficiently.

Theorem: do-Shapley Identifiability

The do-Shapley is identifiable if $V_i \in \mathbf{V}$ and its children are not connected by the bidirected paths (the path where unmeasured variables confound variables on the paths).

[1] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[2] Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." International Conference on artificial intelligence and statistics. PMLR, 2020.

[3] Heskes, Tom, et al. "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models." Advances in neural information processing systems 33 (2020): 4778-4789.

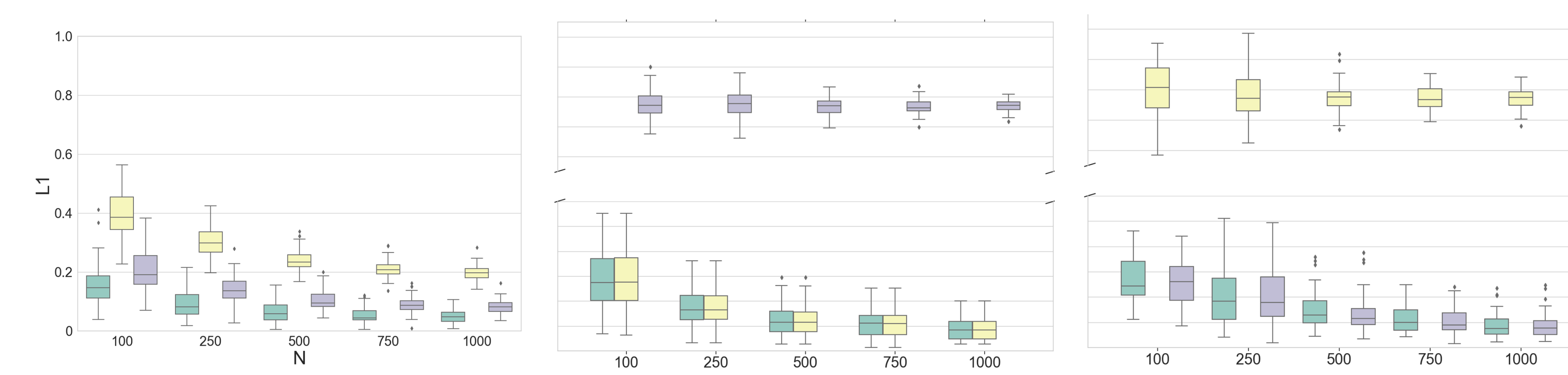
do-Shapley Estimation

- Computing do-Shapley takes exponential time because it iterates all $S \subseteq [n]$. Also, $\mathbb{E}[Y \mid do(\mathbf{v}_S)]$ $\forall S \subseteq [n]$ must be well-approximated from finite samples for accurately computing the do-Shapley.
- We developed do-DML-Shapley, based on the double/debiased machine learning (DML) [5], which can be evaluated efficiently, and exhibits robustness properties (debiasedness, doubly robustness) against errors.

Theorem: Robustness of do-DML-Shapley

The do-DML-Shapley $\hat{\phi}_{v_i}$ converges to the do-Shapley ϕ_{v_i} fast even when the nuisance parameters of $\hat{\phi}_{v_i}$ converges slowly or misspecified.

Simulation



(a) Debiasedness (b) Doubly Robustness - 1 (c) Doubly Robustness - 2

- We compared the proposed method (do-DML-Shapley) with competing methods (regression, inverse-probability-weighting based).
- The result shows that the do-DML-Shapley converges faster than other methods, exhibiting robustness properties against model errors.

Conclusion

We developed the do-Shapley and corresponding estimators for measuring the contribution of each input feature based on its causal impact on the outcome.

[4] Janzing, Dominik, et al. "Quantifying causal contributions via structure preserving interventions." *arXiv preprint arXiv:2007.00714* (2020).

[5] Chernozhukov, Victor, et al. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21.1 (2018): C1-C68.