Comparison of meta-learners for estimating multi-valued treatment heterogeneous effects

ICML 2023 – Honolulu, Hawai

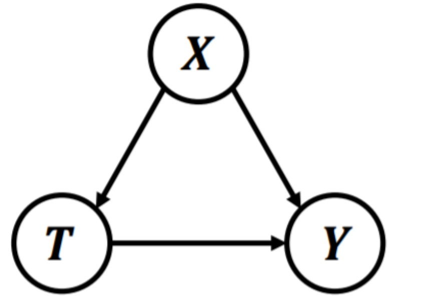
Naoufal Acharki[♠,◇], Ramiro Lugo[◇], Antoine Bertoncello[◇] and Josselin Garnier [♠]

June 30, 2023

[♠]CMAP, Ecole polytechnique, Institut Polytechnique de Paris.
[◇]TotalEnergies One Tech.

Context: Heterogeneous Treatments Effects estimation

- $i = 1, \dots, n$: an individual subject to a treatment.
- T: the discrete treatment assignment variable.
- $\mathcal{T} = \{t_0, t_1, \dots, t_K\}$: the support of the treatment assignement.
- $\boldsymbol{X} \in \mathbb{R}^d$: vector of d covariates (confounders).
- $Y_{\text{obs}} \in \mathbb{R}$: the observed outcome corresponding to the treatment T.
- $Y(t)$: the counterfactual outcome that would have been observed under treatment level $t \in \mathcal{T}$.



Rubin Causal Model [Rubin, 1974]

Goal: Estimate the Conditional Average Treatment Effect (CATE) of T on Y

$$\tau_k(\boldsymbol{x}) = \mathbb{E}[Y(t_k) - Y(t_0) | \boldsymbol{X} = \boldsymbol{x}] \quad \text{for } k = 1, \dots, K$$

Context: Heterogeneous Treatments Effects estimation

Binary treatments Various ML-based models are built to estimate the CATE (e.g. Causal Forests [Wager and Athey, 2018], Bayesian Causal Forests [Hahn et al., 2020], SIN [Kaddour et al., 2021] etc.)

Discrete and continuous treatments Most existing work tend to extend naivly existing approaches for binary treatments.

Problem 1: We need to simplify the selection task and the model's interpretation

Problem 2: The heterogeineity of the treatment and the heterogeineity of effects cannot be distinguished [Heiler and Knaus, 2022].

Problem 3: We cannot identify the key parameters on the performances of estimators (e.g. the number of treatments K).

Tools: Meta-Learners for estimating the CATE

A Meta-learner [Künzel et al., 2019] is a statistical framework (developed initially for binary treatments) that models and estimates the CATE

$$\tau_k(\boldsymbol{x}) = \mathbb{E}[Y(t_k) - Y(t_0)|\boldsymbol{X} = \boldsymbol{x}].$$

Purpose: Understand the strengths and weaknesses of algorithms from a theoretical viewpoint.

Remark: Most previous ML algorithms fall are seen theoretically as a meta-learner.

Direct plug-in meta-learners

Naive estimators that estimate the CATE directly by a plug-in difference.

The **T-learner** (T stands for *two*): Compute the CATE as plug-in difference $\widehat{\tau}_k^{(\mathrm{T})}(\boldsymbol{x}) = \widehat{\mu}_{t_k}(\boldsymbol{x}) - \widehat{\mu}_{t_0}(\boldsymbol{x})$ using two models models μ_{t_k} and μ_{t_0}.

The **S-learner** (S stands for *single*): Compute the CATE as plug-in difference $\widehat{\tau}_k^{(\mathrm{S})}(\boldsymbol{x}) = \widehat{\mu}(\boldsymbol{x}, t_k) - \widehat{\mu}(\boldsymbol{x}, t_0)$ using a $\mu(w, \boldsymbol{x}) = \mathbb{E}(Y_{\mathrm{obs}} \mid T = w, \boldsymbol{X} = \boldsymbol{x})$.

The **naive X-learner** (X- stands for *cross*): Compute the CATE as plug-in difference $\widehat{\tau}_k^{(\mathrm{nv,X})}(\boldsymbol{x}) = g(\boldsymbol{x})\, \widehat{\tau}^{(k)}(\boldsymbol{x}) + (1 - g(\boldsymbol{x}))\, \widehat{\tau}^{(0)}(\boldsymbol{x})$. with g some given weighting function.

Pseudo-outcome meta-learners

Debiased learners that estimate CATE by regressing a pseudo-outcome Z_k: $\mathbb{E}(Z_k \mid \boldsymbol{X}) = \tau_k(\boldsymbol{X})$.

M-learner: M stands for the *modified* weighted outcome:

$$Z_k^M = \frac{\mathbf{1}\{T = t\}}{\widehat{r}(t, \boldsymbol{X})} Y_{\text{obs}} - \frac{\mathbf{1}\{T = t_0\}}{\widehat{r}(t_0, \boldsymbol{X})} Y_{\text{obs}} \text{ where } r(T, \boldsymbol{X}) = \mathbb{P}(T \mid \boldsymbol{X}).$$

DR-learner: DR stands for the *Doubly-Robustness* with respect to mispecification of $\widehat{r}$ and $\widehat{\mu}_t$:

$$Z_k^{DR} = \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t_k, \boldsymbol{X})} \mathbf{1}\{T = t_k\} - \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t_0, \boldsymbol{X})} \mathbf{1}\{T = t_0\} + \widehat{\mu}_{t_k}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X}).$$

X-learner: X stands for the *Cross* estimation procedure over all treatments:

$$\begin{aligned} Z_k^X = \mathbf{1}\{T = t_k\}(Y_{\text{obs}} - \widehat{\mu}_{t_0}(\boldsymbol{X})) + \sum_{k' \neq k} \mathbf{1}\{T = t_{k'}\}(\widehat{\mu}_{t_k}(\boldsymbol{X}) - Y_{\text{obs}}) \\ + \sum_{k' \neq k} \mathbf{1}\{T = t_{k'}\}(\widehat{\mu}_{t_{k'}}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})). \end{aligned}$$

$\widehat{r}$ and $\widehat{\mu}_t$ are two estimators of $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mu_t(\boldsymbol{x}) = \mathbb{E}[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}]$.

Neyman orthogonality based learners

Learners that use the generalized Robinson [1988] decomposition and estimate CATEs by minimizing (jointly or separatly) a loss function (ℓ_R for **R-learner** or $\ell_{R,Bin}$ for **Bin R-learner**).

The **R-learner** estimates all K CATE models $\{\tau_t\}_k$ by addressing the problem:

$$\{\widehat{\tau}_k^{(\mathrm{R})}\}_{k=1}^K = \arg\min_k \frac{1}{n}\sum_{i=1}^n \ell_R(\boldsymbol{X}_i).$$

The **Bin R-learner** estimates separately CATE models τ_k by addressing the problem:

$$\widehat{\tau}^{(\mathrm{R,Bin})} = \arg\min_k \frac{1}{n}\sum_{i=1}^n \ell_{R,Bin}(\boldsymbol{X}_i).$$

Upper bounds on error

Assuming that $Y(t) = f(t, \boldsymbol{X}) + \varepsilon(t)$ where $f(t, \boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\top \beta_t$, the Ordinary Least Square estimators $\widehat{\beta}^*_{t_k}$ have covariance matrix $\mathbb{V}(\widehat{\beta}^*_{t_k}) = \mathbf{C}/n$ whose terms are bounded by:

Theorem

$$\mathcal{E}^T = \mathcal{E}^{X,nv} = \mathcal{O}\left(\frac{1}{\rho(t_k)} + \frac{1}{\rho(t_0)}\right) \text{ for the T- and naive X-learners,}$$

$$\mathcal{E}^M = \mathcal{O}\left(\frac{1}{r_{\min}^{1+\epsilon}}\right) \text{ for the M-learner,}$$

$$\mathcal{E}^{DR} = \mathcal{O}\left(\frac{\mathrm{err}(\widehat{\mu}_{t_k}) + \mathrm{err}(\widehat{\mu}_{t_0})}{r_{\min}^{1+\epsilon}}\right) \text{ for the DR-learner,}$$

$$\mathcal{E}^X = \mathcal{O}\Big(K^2 \sum_{k' \neq k} \mathrm{err}(\widehat{\mu}_{t_{k'}})\Big) \text{ for the X-learner.}$$

where $\mathbb{P}(T = t) = \rho(t) > 0$, $r_{\min}$ *the lower bound of propensity score and for all* $\epsilon > 0$

Summary table of multi-treatments meta-learners

Meta-learner	Advantages	Disadvantages
T-learner (naive X-learner)	Simple approach	Selection bias Low samples
S-learner	Simple approach	Confounding effects Regularization bias
M-learner	Consistency	High variance
DR-learner	Consistency Doubly Robust	Possibly high variance
X-learner	Consistency Low variance	Non-intuitive
R-learner	Interaction effects	Non-identifiability
Bin R-learner	Identifiability	Computational cost

Conclusion

- Highlighting the difference between the naive and generalization versions of both X- and R-learners.
- Demonstrating theoretically the X-learner performances with multi-treatments.
- Identifying the impact of the number of treatment levels and the lower bound of the propensity score on the M-, DR and X-learners.
- To-do: Extend this analysis to Causal Inference with continuous treatments.

References

P. Heiler and M. C. Knaus. Effect or treatment heterogeneity? policy evaluation with aggregated and disaggregated treatments, 2022.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116:4156–4165, Feb 2019. ISSN 1091-6490.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

D. Rubin. Estimating causal effects if treatment in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 01 1974.