Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager and Susan Athey

解读：雷印如

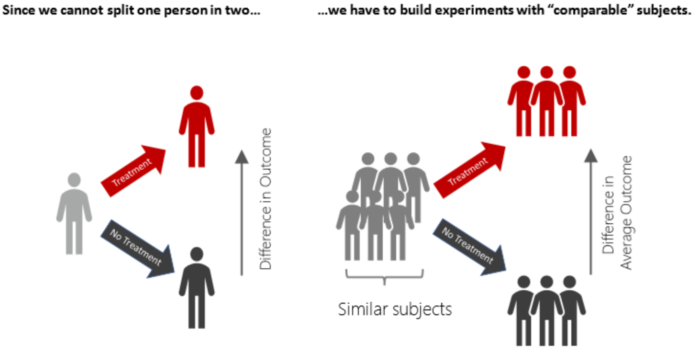
2022 年 2 月 28 日

Background-Causal Inference

- Eliminate Endogeneity
  - IV Estimation (How to find the suitable IVs?)
- Randomized Control Trials (No Counterfactual)
  - Difference-in-difference
  - Regression discontinuity design
  - Pseudo-randomized experiments
- Selection bias to Data driven

Treatment Effect

$$\tau(x) = E[Y_i^{(1)} - Y_i^{(0)} | X_i = x]$$



Key Assumption CIA

$$[Y_i^{(1)} - Y_i^{(0)}] \perp\!\!\!\perp W_i | X_i (CIA)$$

$$E[Y_i^{(1)} | W_i = 1] = E[Y_i^{(1)} | W_i = 0]$$

$$E[Y_i^{(0)} | W_i = 1] = E[Y_i^{(0)} | W_i = 0]$$

$$E[Y_i^{(1)} | W_i = 1, X_i] = E[Y_i^{(1)} | W_i = 0, X_i]$$

$$ATE = E[Y_i^{(1)} | W_i = 1, X_i] - E[Y_i^{(0)} | W_i = 0, X_i]$$

$$\tau(x) = E[Y_i^{(1)} - Y_i^{(0)} | X_i = x]$$

$$E[Y_i(\frac{W_i}{e(x)} - \frac{1 - W_i}{1 - e(x)}) | X_i = x] = \tau(x)$$

Motivation

- Most datasets have been too small to meaningfully explore heterogeneity of treatment effects beyond dividing the sample into a few subgroups.
- Classical approaches to nonparametric estimation (KNN) of heterogeneous treatment effects break down as the number of covariates increases.
- Random forests are related to KNN, but with greater widespread success at prediction and classification.

Literatures-Tree based

- Honest Tree (Athey and Imbens, 2016)
  - Risk of overfitting and lack of flexibility
- Bayesian additive regression tree (BART)
  - Markov chain Monte Carlo sampling based on a convenience prior (Green and Kern,2012; Hill, 2011; Hill and Su, 2013)
  - A limitation of this line of work is that, until now, it has lacked formal statistical inference results.

Contribution

- Complementary to showing that forest-based methods need not only be viewed as black-box heuristics, and can instead be used for rigorous asymptotic analysis
- Under regularity assumptions, causal forests can use the unconfoundedness assumption to achieve consistency without needing to explicitly estimate the propensity e(x)

Framework

- In the context of causal trees, leaves as small enough that the (Yi, Wi) pairs act as though they had come from a randomized experiment. Then, it is natural to estimate the treatment effect for any $x \in L$ as

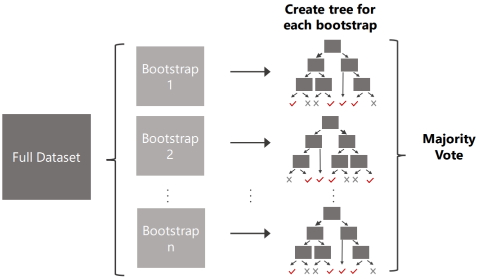
$$\hat{\tau}(x) = \frac{1}{|i : W_i = 1, X_i \in L|} \sum_{i:W_i=1,X_i \in L}^{Y_i}$$

$$-\frac{1}{|i : W_i = 0, X_i \in L|} \sum_{i:W_i=0,X_i \in L}^{Y_i}$$

Causal Forest

The forest then aggregates the predictions by average. In practice, this aggregation scheme helps reduce variance and smooths sharp decision boundaries

$$\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$$



Compare with Random Forest

A causal forest contains 1, . . . , B trees, each grown according to the following procedure based on Honest trees:

- Draw a subset of data, S_b.
- Divide S_b into two equal halves, S_{tr} and S_{est}
- Using only a subset of covariates, outcomes (Y_i) from S_{tr}, and treatment (W_i) and covariates (X_i) from both S_{tr} and S_{est}, grow tree b.
- Use the partitioning rules that define tree b to partition S_{est}

Compare with Random Forest

- Minimize intra-node variance VS. Maximize inter-node variability

$$err(C_1, C_2) = \sum_{j=1,2} P_{C_j} \frac{\sum_{i=1}^{n_{C_j}} (Y_{iC_j} - \hat{Y}_{C_j})^2}{n_{x_j}}$$

$$err_{causal}(C_1, C_2) = \sum_{j=1,2} P_{C_j} (\hat{\tau}_{C_j} - E(\hat{\tau}_{C_j}))^2$$

$$err_{causal}(C_1, C_2) = K(P) - E[\Delta(C_1, C_2)] + o(r^2)$$

- The heterogeneity of the treatment effect estimated as the treatment effects between nodes differ from each other as much as possible.

Asymptotic Inference with Causal Forests

- Our first result is that causal forests are consistent for the true treatment effect
- The predictions made by a causal forest are asymptotically Gaussian and unbiased

$$(\hat{\tau}(x) - \tau(x))/\sqrt{Var[\hat{\tau}(x)]} \Rightarrow N(0,1)$$

- Proof Omitted

Simulation: Causal Forest VS. KNN

main effect: $m(x) = 2^{-1}E[Y^{(0)} + Y^{(1)}|X = x]$

treatment effect: $\tau(x) = E[Y^{(1)} - Y^{(0)}|X = x]$

treatment propensity: $e(x) = P[W = 1|X = x]$

- Suppose $\tau(t) = 0$
- Suppose $\tau(t) = \lambda(X_1)\lambda(X_2), \lambda(x) = 1 + \frac{1}{1+e^{-20(x-1/3)}}$
- Suppose $\tau(t) = \lambda(X_1)\lambda(X_2), \lambda(x) = 1 + \frac{1}{1+e^{-12(x-1/2)}}$

Simulation: Causal Forest VS. KNN

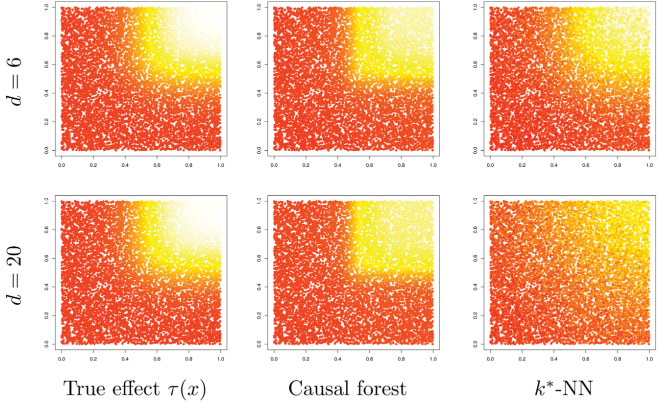
- Suppose $\tau(t) = 0$ and $\tau(t) = \lambda(X_1)\lambda(X_2)$

	Mean-squared error			Coverage		
d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	0.02 (0)	0.21 (0)	0.09 (0)	0.95 (0)	0.93 (0)	0.62 (1)
5	0.02 (0)	0.24 (0)	0.12 (0)	0.94 (1)	0.92 (0)	0.52 (1)
10	0.02 (0)	0.28 (0)	0.12 (0)	0.94 (1)	0.91 (0)	0.51 (1)
15	0.02 (0)	0.31 (0)	0.13 (0)	0.91 (1)	0.90 (0)	0.48 (1)
20	0.02 (0)	0.32 (0)	0.13 (0)	0.88 (1)	0.89 (0)	0.49 (1)
30	0.02 (0)	0.33 (0)	0.13 (0)	0.85 (1)	0.89 (0)	0.48 (1)

	Mean-squared error			Coverage		
d	CF	7-NN	50-NN	CF	7-NN	50-NN
2	0.04 (0)	0.29 (0)	0.04 (0)	0.97 (0)	0.93 (0)	0.94 (0)
3	0.03 (0)	0.29 (0)	0.05 (0)	0.96 (0)	0.93 (0)	0.92 (0)
4	0.03 (0)	0.30 (0)	0.08 (0)	0.94 (0)	0.93 (0)	0.86 (1)
5	0.03 (0)	0.31 (0)	0.11 (0)	0.93 (1)	0.92 (0)	0.77 (1)
6	0.02 (0)	0.34 (0)	0.15 (0)	0.93 (1)	0.91 (0)	0.68 (1)
8	0.03 (0)	0.38 (0)	0.21 (0)	0.90 (1)	0.90 (0)	0.57 (1)

Simulation: Causal Forest VS. KNN

- Although the causal forest faithfully captures the qualitative aspects of the true τ -surface, it does not exactly match its shape, especially in the upper-right corner where τ is largest.



Conclusion

- This article proposed a class of nonparametric methods for heterogeneous treatment effect estimation that allow for data driven feature selection
- The causal forests maintain the asympototically normal and unbiased property, which is an adaptive method for traditional KNN, provides better MSE and nominal coverage rates.
- `https://econml.azurewebsites.net/`

Application

$$\text{est} = \text{CausalForestDML}()$$

$$\text{est.fit}(\text{Y, T, X=X, W=W})$$

Feature Name	Type	Details
account_age	W	user's account age
age	W	user's age
avg_hours	W	the average hours user was online per week in the past
days_visited	W	the average number of days user visited the website per week in the past
friend_count	W	number of friends user connected in the account
has_membership	W	whether the user had membership
is_US	W	whether the user accesses the website from the US
songs_purchased	W	the average songs user purchased per week in the past
income	X	user's income
price	T	the price user was exposed during the discount season (baseline price * small discount)
demand	Y	songs user purchased during the discount season