DOTE 6635: Artificial Intelligence for Business Research

Heterogeneous Treatment Effect

Renyu (Philip) Zhang

1

Heterogeneous Treatment Effect

- Why do we care about heterogeneous treatment effect (HTE), or conditional average treatment effect (CATE)?
- In PS-type of estimations (IPW, AIPW, PSTRT, PSM, etc.), we are essentially estimating CATE on units that have sufficient overlapping.
- More generally, we care about how HTE affects what we are estimating, and we also care about how to best estimate HTE.
- HTE for insights: Provide supportive evidence on proposed mechanisms (also called moderating effects).
  - What HTEs you are testing should be informed by theory.
- HTE for prescriptions: Decide precisely how to assign costly treatments or there are responses in both directions.
  - ML methods are most powerful in estimating HTE for prescriptions.
    - Personalized medicine: Which cancer therapy should be administered to this specific patient?
    - Personalized targeting in ads and promotions.
    - ML Fairness: If we screen job/loan candidates with ML, how do we ensure we are not discriminating against certain people?

2

2

HTE Literature

- Causal Tree/Forest: This literature directly solves the HTE problem by modelling HTE as a tree/forest.
- DML Literature: This literature does not solve the HTE problem directly but provides a better way to estimate ATE under unconfoundedness. But the byproduct, nuisance parameter training, can help with the HTE estimation at large.
- Uplift modeling: Use meta-learning approaches in CS to estimate HTE, without properly quantifying standard errors.

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

Quasi-oracle estimation of heterogeneous treatment effects
X Nie, S Wager - Biometrika, 2021 - academic.oup.com
… to **estimating heterogeneous treatment effects** that addresses both of the above concerns. Our framework allows for fully automatic specification of **heterogeneous treatment effect** …
☆ Save 99 Cite Cited by 967 Related articles All 11 versions Web of Science: 232

Metalearners for estimating heterogeneous treatment effects using machine learning
SR Künzel, JS Sekhon, PJ Bickel, B Yu - Proceedings of the national …, 2019 - pnas.org
… To **estimate** the CATE, we build on regression or supervised **learning** methods in statistics and **machine learning**, … (or **metalearners**) for **estimating** the CATE in a binary **treatment** setting. …
☆ Save 99 Cite Cited by 1415 Related articles All 13 versions Web of Science: 477

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

3

3

Causal Tree/Forest Literature

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

- The literature provides methods for estimating heterogeneity in causal effects in RCT or observational studies and conducting hypothesis tests about the magnitude of differences in treatment effects across different subgroups of the population.
- Causal tree/forest offers a data-driven automated approach to partition the data into subpopulations that differ in the magnitude of their treatment effects.
  - The method also constructs valid confidence intervals for heterogeneous treatment effects, even with many covariates relative to sample size, and without the "sparsity" assumptions.
- Athey and Imbens (2016) is one of the first papers that study the automated way of analyzing HTEs.
- This literature makes the standard unconfoundedness, overlapping, and SUTVA assumptions.

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

4

4

Causal Tree Setup

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748 »

- Parameter of interest:

Conditional Average Treatment Effects and Partitioning. Define the conditional average treatment effect

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

- Assumptions:
  - Unfoundedness
  - Overlapping
  - SUTVA
  - Binary Treatment

- To understand the causal tree algorithm, let's revisit the decision tree model for prediction first.

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEghRgHNezS30lI49G-

Decision Tree

- 5000 data points

- <u>Goal</u>: Predict whether the customer will accept the personal loan offer.

- Need to find: $f : X \rightarrow Y \in \{0,1\}$

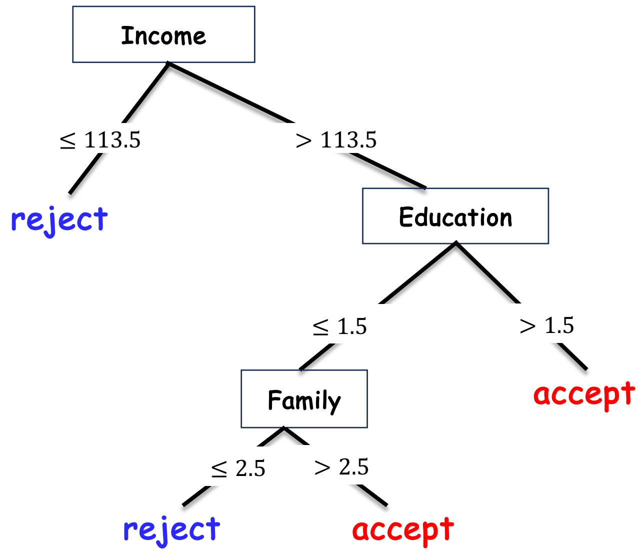
- Binary outcome for now.

Personal.Loan	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities.Account	CD.Account	Online	CreditCard
0	25	1	49	4	1.6	1	0	1	0	0	0
0	45	19	34	3	1.5	1	0	1	0	0	0
0	39	15	11	1	1.0	1	0	0	0	0	0
0	35	9	100	1	2.7	2	0	0	0	0	0
0	35	8	45	4	1.0	2	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
0	29	3	40	1	1.9	3	0	0	0	1	0
0	30	4	15	4	0.4	1	85	0	0	1	0
0	63	39	24	2	0.3	3	0	0	0	0	0
0	65	40	49	3	0.5	2	0	0	0	1	0
0	28	4	83	3	0.8	1	0	0	0	1	1

Y X

Decision Tree

- Each internal node is split based on one feature.
- Each leaf node is assigned with one Y.
- Benefits:
  - A smart data structure to scale kNN.
  - Flexible and large space of functions.
  - Easy for human interpretation.

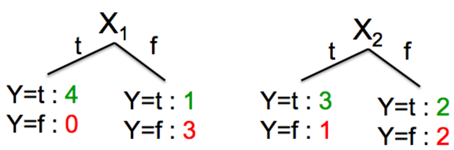


7

7

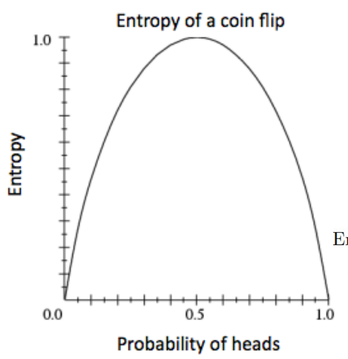
Fitting a Decision Tree

- Given a training set, identifying the best decision tree in NP-complete.



On which feature should we split?

- Heuristics to build a tree:
  1. Start with an empty tree.
  2. Split on the feature that gives the largest reduction in impurity.
  3. Repeat step 2 until a stopping criterion is met.
  4. Assign the major class (classification) or the average outcome to each leaf (regression).



Entropy: $H(Y) = -\sum_{i=1}^{k} \mathbb{P}(Y = y_i) log_2 \mathbb{P}(Y = y_i)$

- Measurement of impurity:
  - **Classification:** Gini index; Entropy
  - **Regression:** Squared Error $\sum_i [(Y_i - \hat{\mu}(X_i))^2]$

8

8

From Decision Tree to Causal Tree

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

- The potential problems of using decision trees for causal inference:
  - We use the same sample to construct the tree and use the tree for inference (i.e., overfitting bias).
  - We are predicting the conditional average outcomes, but not the CATEs.
- Recall the fundamental challenge in causal inference: We do not observe the counterfactuals and the treatment effects.
- Ideally, we would like our objective function defined directly on the treatment effects: $\sum_i [(\tau_i - \hat{\tau}(X_i))^2]$
- The HTE literature basically approximates the loss function in a reasonable way.

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

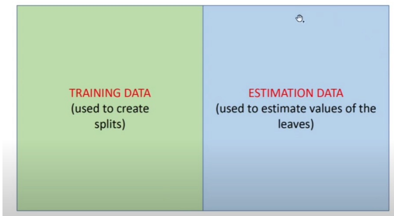
9

9

From Decision Tree to Causal Tree

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

- Split the data into training and estimation samples.



$$\hat{\mu}(x; \mathcal{S}, \Pi) \equiv \frac{1}{\#(i \in \mathcal{S} : X_i \in \ell(x; \Pi))} \sum_{i \in \mathcal{S} : X_i \in \ell(x; \Pi)} Y_i$$

- The target is no longer the in-sample fit of Y, but to fit on estimation sample:

Honest Target

$$\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi) \equiv \frac{1}{\#(\mathcal{S}^{\text{te}})} \sum_{i \in \mathcal{S}^{\text{te}}} \left\{ (Y_i - \hat{\mu}(X_i; \mathcal{S}^{\text{est}}, \Pi))^2 - Y_i^2 \right\}.$$

Adjustment independent of estimator

Tree-split based on training data

The (adjusted) expected MSE is the expectation of $\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)$ over test and estimation samples:

$$\text{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}} [\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)], \qquad Q^H(\pi) \equiv -\mathbb{E}_{\mathcal{S}^{\text{est}}, \mathcal{S}^{\text{est}}, \mathcal{S}^{\text{tr}}} [\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi(\mathcal{S}^{\text{tr}}))]$$

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

10

10

Decision Tree vs. Causal Tree

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …

- Let's take a closer took at the difference between decision tree and causal tree.
- Original decision tree:

equivalent to maximizing the sum of the final nodes' squared predictions

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \boxed{\hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})}$$

test set

training set

tree partitions based on training set

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

11

11

Decision Tree vs. Causal Tree

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

- Let's take a closer took at the difference between decision tree and causal tree.

Causal Trees:

treatment effects

only using training set

$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \boxed{\mathcal{S}^{tr}, \Pi^{tr}})}_{\text{rewards high heterogeneity}} + \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

variance ...

test set

estimation set

tree partitions based on training set

in the sample of control observations in leaf (l)

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

12

12

Decision Tree vs. Causal Tree

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

	Regression Tree	Causal Tree
Predictions based on	training sample	estimation sample
Splitting rule minimizes in-sample	RSS	Honest Target
Segments X for heterogeneity	in outcomes	in treatment effects

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

13

13

Summary

Recursive partitioning for heterogeneous causal effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… In this paper we propose methods for estimating heterogeneity in causal effects in … treatment effects across subsets of the population. We provide a data-driven approach to partition the …
☆ 保存 99 引用 被引用次数：2224 相关文章 所有 17 个版本 Web of Science: 748

- Athey and Imbens (2016) first propose a new way of automatically estimating HTEs using decision tree.
- 2 major changes compared with traditional decision trees:
  - Inference done using a different estimation sample.
  - Splitting done based on treatment effects rather than outcomes.
- Causal tree has well-performed simulation results without theoretical guarantee.
- To find theoretical guarantee (i.e., root-n consistency and asymptotic normality), we leverage random forests.

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

14

14

Estimation and inference of heterogeneous treatment effects using random forests
S Wager, S Athey - Journal of the American Statistical Association, 2018 - Taylor & Francis
… of **treatment effect heterogeneity**. In this article, we develop a nonparametric … **forest** for **estimating heterogeneous treatment effects** that extends Breiman's widely used **random forest** …

Causal Forest

- Wager and Athey (2018) is a follow-up of Athey and Imbens (2016), building a causal forests that averages across different causal trees.
- It builds on a general statistical framework to build valid confidence intervals (though sometimes too large…) for the estimated CATE.
- Asymptotic normality:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1,\ X_i \in L\}|} \sum_{\{i:W_i=1,\, X_i\in L\}} Y_i - \frac{1}{|\{i : W_i = 0,\ X_i \in L\}|} \sum_{\{i:W_i=0,\, X_i\in L\}} Y_i. \quad (5)$$

In the following sections, we will establish that such trees can be used to grow causal forests that are consistent for $\tau(x)$.[3]

Finally, given a procedure for generating a single causal tree, a causal forest generates an ensemble of B such trees, each of which outputs an estimate $\hat{\tau}_b(x)$. The forest then aggregates their predictions by averaging them: $\hat{\tau}(x) = B^{-1}\sum_{b=1}^{B}\hat{\tau}_b(x)$. We always assume that the individual causal trees in the forest are built using random subsamples of s training examples, where $s/n \ll 1$; for our theoretical results, we will assume that $s \asymp n^{\beta}$ for some $\beta < 1$. The advantage of a forest over a single tree is that it is not always clear what the "best" causal tree is.

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

To define the variance estimates, let $\hat{\tau}_b^*(x)$ be the treatment effect estimate given by the bth tree, and let $N_{ib}^* \in \{0,\ 1\}$ indicate whether or not the ith training example was used for the bth tree.[4] Then, we set

$$\widehat{V}_{IJ}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*[\hat{\tau}_b^*(x),\ N_{ib}^*]^2, \quad (8)$$

where the covariance is taken with respect to the set of all the trees $b = 1, \ldots, B$ used in the forest. The term $n(n-1)/(n-s)^2$ is a finite-sample correction for forests grown by subsampling without replacement; see Proposition 5. We show that this variance estimate is consistent, in the sense that $\widehat{V}_{IJ}(x)/\,\mathrm{Var}[\hat{\tau}(x)] \to_p 1$.

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

Algorithms

Procedure 1. Double-Sample Trees

Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits.

Input: n training examples of the form $(X_i,\ Y_i)$ for regression trees or $(X_i,\ Y_i,\ W_i)$ for causal trees, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample of size s from $\{1, \ldots, n\}$ without replacement, and then divide it into two disjoint sets of size $|\mathcal{I}| = \lfloor s/2 \rfloor$ and $|\mathcal{J}| = \lceil s/2 \rceil$.
2. Grow a tree via recursive partitioning. The splits are chosen using any data from the $\mathcal{J}$ sample and X- or W-observations from the $\mathcal{I}$ sample, but without using Y-observations from the $\mathcal{I}$-sample.
3. Estimate leafwise responses using only the $\mathcal{I}$-sample observations.

Double-sample *regression* trees make predictions $\hat{\mu}(x)$ using (4) on the leaf containing x, only using the $\mathcal{I}$-sample observations. The splitting criteria is the standard for CART regression trees (minimizing mean-squared error of predictions). Splits are restricted so that each leaf of the tree must contain k or more $\mathcal{I}$-sample observations.

Double-sample *causal* trees are defined similarly, except that for prediction we estimate $\hat{\tau}(x)$ using (5) on the $\mathcal{I}$ sample. Following Athey and Imbens (2016), the splits of the tree are chosen by maximizing the variance of $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$; see Remark 1 for details. In addition, each leaf of the tree must contain k or more $\mathcal{I}$-sample observations of *each* treatment class.

Estimation and inference of heterogeneous treatment effects using random forests
S Wager, S Athey - Journal of the American Statistical Association, 2018 - Taylor & Francis
… of **treatment effect heterogeneity**. In this article, we develop a nonparametric … **forest** for **estimating heterogeneous treatment effects** that extends Breiman's widely used **random forest** …
☆ Save 99 Cite Cited by 3705 Related articles All 14 versions »

Procedure 2. Propensity Trees

Propensity trees use only the treatment assignment indicator W_i to place splits, and save the responses Y_i for estimating τ.

Input: n training examples $(X_i,\ Y_i,\ W_i)$, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample $\mathcal{I} \in \{1, \ldots, n\}$ of size $|\mathcal{I}| = s$ (no replacement).
2. Train a classification tree using sample $\mathcal{I}$ where the outcome is the treatment assignment, that is, on the $(X_i,\ W_i)$ pairs with $i \in \mathcal{I}$. Each leaf of the tree must have k or more observations of *each* treatment class.
3. Estimate $\tau(x)$ using (5) on the leaf containing x.

In step 2, the splits are chosen by optimizing, for example, the Gini criterion used by CART for classification (Breiman et al. 1984).

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

Summary

Estimation and inference of heterogeneous treatment effects using random forests
S Wager, S Athey - Journal of the American Statistical Association, 2018 - Taylor & Francis
… of **treatment effect heterogeneity**. In this article, we develop a nonparametric … **forest** for **estimating heterogeneous treatment effects** that extends Breiman's widely used **random forest** …
☆ Save 99 Cite Cited by 3705 Related articles All 14 versions

- Causal forest is built upon Athey and Imbens (2016).
- You can build a forest and average them, which obtains the asymptotically optimal CATE.
- Valid confidence intervals can be constructed.
- It still only applies to binary treatments.

Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

17

17

Generalized Random Forest

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …
☆ 保存 99 引用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784

- Generalized random forest works well with categorical treatments or even continuous treatments.
  - Be careful with the confidence interval.
- A new perspective of RF-based HTE estimators: Adaptive clustering algorithm.

grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

18

18

kNN Matching

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …
☆ 保存 ⁹⁹ 引用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784

- The most intuitive way of estimating HTEs on a general set of treatment variables is kNN matching.

- kNN matching:
  - For any covariate X where you want to estimate treatment effects.
  - Find a nearby neighborhood of X.
  - Run a simple ATE regression on all the neighbors of X where units are weighted through some distance metric.

grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

19

19

Generalized Random Forest

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …
☆ 保存 ⁹⁹ 引用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784

- Random forest does kNN with adaptively putting weights on observations near each covariate X.

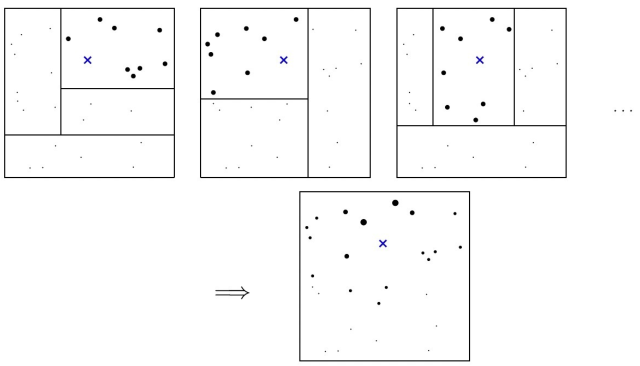


FIG. 1. *Illustration of the random forest weighting function. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x.*

grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

20

20

GRF Model

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …
☆ 保存 ⑰ 引用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784

Definition 1. (GRF model)

Suppose that a sequence of i.i.d. random vector $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1,\ldots,n}$ satisfies

$$\mathbb{E}\left[\psi_{\theta(x)}(Y_i) | X_i = x\right] = 0 \quad \text{for all } x \in \mathcal{X} \qquad (1)$$

where

- $\theta \in \Theta = \{\theta : \mathcal{X} \to \mathbb{R}\}$: parameter of interest
- $\psi : \Theta \times \mathcal{Y} \to \mathbb{R}$: some scoring function

(Note) ψ depends on the parameter for example

- mean : $\psi_{\theta(x)}(y) = y - \theta(x)$
- quantile : $\psi_{\theta(x)}(y) = \tau - \mathbf{1}_{\{y \leq \theta(x)\}}$ for some $\tau \in (0, 1)$
- likelihood : $\psi_{\theta(x)}(y) = \nabla \log(f_{\theta(x)}(y))$ for some (localized) p.d.f f

For HTE inference (RCT): $\psi = y - \theta(x)D$

Definition 2. (GRF estimator)

Given a data $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1,\ldots,n}$ satisfying (1), an estimator of $\theta = (\theta(x))_{x \in \mathcal{X}} \in \Theta$ (defined in Def 1) is defined by

$$\hat{\theta}(x) \in \arg\min_{e \in \mathbb{R}} \left\{ \left| \sum_{i=1}^{n} \alpha_i(x) \psi_e(Y_i) \right| \right\} \quad \text{for all } x \in \mathcal{X}$$

where

- $\alpha_i(x) \in [0, 1]$: weight function based on Random Forests

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{bi}(x), \quad \alpha_{bi}(x) = \frac{\mathbf{1}_{\{X_i \in L_b(x)\}}}{|L_b(x)|}$$

- B : number of trees
- $L_b(x)$: "leaf" of b-th tree containning the test point $x \in \mathcal{X}$
- $|L_b(x)|$: subsample size falling in the leaf $L_b(x)$

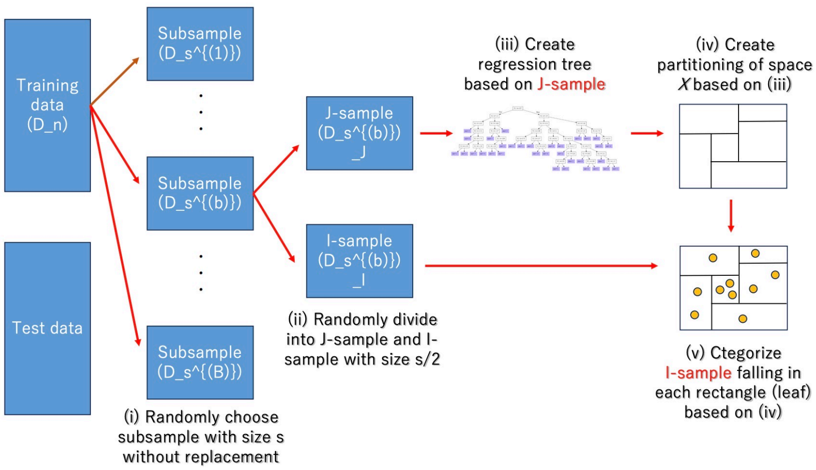
GRF Slides: https://math.bu.edu/BKT2023/slides/Shiraishi-slides.pdf
grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

21

21

Double Sample Procedure

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …
☆ 保存 ⑰ 引用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784



GRF Slides: https://math.bu.edu/BKT2023/slides/Shiraishi-slides.pdf
grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

22

22

GRF Algorithms

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …

Algorithm 1 Generalized random forest with honesty and subsampling

All tuning parameters are prespecified, including the number of trees B and the sub-sampling s rate used in SUBSAMPLE. This function is implemented in the package grf for R and C++.

procedure GENERALIZEDRANDOMFOREST(set of examples $\mathcal{S}$, test point x)
  weight vector $\alpha \leftarrow$ ZEROS($|\mathcal{S}|$)
  for $b = 1$ to total number of trees B **do**
    set of examples $\mathcal{I} \leftarrow$ SUBSAMPLE($\mathcal{S}, s$)
    sets of examples $\mathcal{J}_1, \mathcal{J}_2 \leftarrow$ SPLITSAMPLE($\mathcal{I}$)
    tree $\mathcal{T} \leftarrow$ GRADIENTTREE($\mathcal{J}_1, \mathcal{X}$) ▷ See Algorithm 2.
    $\mathcal{N} \leftarrow$NEIGHBORS($x, \mathcal{T}, \mathcal{J}_2$) ▷ Returns those elements of $\mathcal{J}_2$ that fall into the same leaf as x in the tree $\mathcal{T}$.
    for all example $e \in \mathcal{N}$ **do**
      $\alpha[e] += 1/|\mathcal{N}|$
  output $\hat{\theta}(x)$, the solution to (2) with weights α/B

$$(\hat{\theta}(x), \hat{\nu}(x)) \in \operatorname{argmin}_{\theta,\nu}\left\{\left\|\sum_{i=1}^{n}\alpha_i(x)\psi_{\theta,\nu}(O_i)\right\|_2\right\}$$

The function ZEROS creates a vector of zeros of length $|\mathcal{S}|$; SUBSAMPLE draws a subsample of size s from $\mathcal{S}$ without replacement; and SPLITSAMPLE randomly divides a set into two evenly-sized, nonoverlapping halves. The step (2) can be solved using any numerical estimator. Our implementation grf provides an explicit plug-in point where a user can write a solver for (2) appropriate for their ψ-function. $\mathcal{X}$ is the domain of the X_i. In our analysis, we consider a restricted class of generalized random forests satisfying Specification 1.

Algorithm 2 Gradient tree

Gradient trees are grown as subroutines of a generalized random forest.

procedure GRADIENTTREE(set of examples $\mathcal{J}$, domain $\mathcal{X}$)
  node $P_0 \leftarrow$ CREATENODE($\mathcal{J}, \mathcal{X}$)
  queue $\mathcal{Q} \leftarrow$ INITIALIZEQUEUE(P_0)
  while NOTNULL(node $P \leftarrow$ POP($\mathcal{Q}$)) **do**
    $(\hat{\theta}_P, \hat{\nu}_P, A_P) \leftarrow$ SOLVEESTIMATINGEQUATION(P) ▷ Computes (4) and (7).
    vector $R_P \leftarrow$ GETPSEUDOOUTCOMES($\hat{\theta}_P, \hat{\nu}_P, A_P$) ▷ Applies (8) over P.
    split $\Sigma \leftarrow$ MAKECARTSPLIT(P, R_P) ▷ Optimizes (9).
    if SPLITSUCCEEDED(Σ) **then**
      SETCHILDREN(P, GETLEFTCHILD(Σ), GETRIGHTCHILD(Σ))
      ADDTOQUEUE($\mathcal{Q}$, GETLEFTCHILD(Σ))
      ADDTOQUEUE($\mathcal{Q}$, GETRIGHTCHILD(Σ))
  output tree with root node P_0

The function call INITIALIZEQUEUE initializes a queue with a single element; POP returns and removes the oldest element of a queue $\mathcal{Q}$, unless $\mathcal{Q}$ is empty in which case it returns null. MAKECARTSPLIT runs a CART split on the pseudo-outcomes, and either returns two child nodes or a failure message that no legal split is possible.

grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

23

23

Summary

Generalized random forests
S Athey, J Tibshirani, S Wager - 2019 - projecteuclid.org
… In line with this approach, our generalized random forest software package builds on the carefully optimized ranger implementation of regression forest splitting rules [Wright and Ziegler …
☆ 保存 59 引用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784

- The causal forest literature provides an elegant way of looking at HTEs using adaptive weighting.
- Causal forests helps estimate the treatment effects as a function of a large number of covariates.
- It is very challenging to evaluate the HTE estimation with real data.
- Some additional interpretation techniques, such as policy evaluations and best linear predictions, are introduced.
- Orthogonal Random Forest (ORF) = GRF + DML

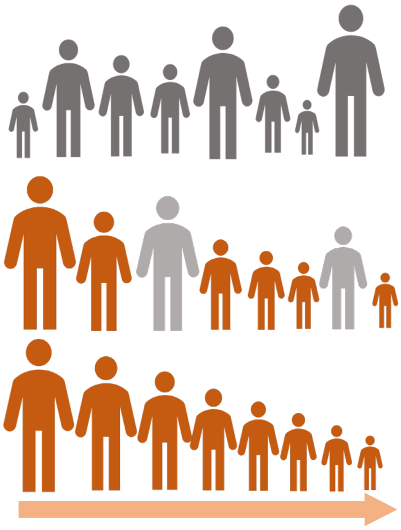
Orthogonal random forest for causal inference
M Oprescu, V Syrgkanis, ZS Wu - … Conference on Machine …, 2019 - proceedings.mlr.press
… We propose the orthogonal random forest, an algorithm that combines Neyman-orthogonality to … to estimation error of nuisance parameters with generalized random forests (Athey et al.…

grf package: https://grf-labs.github.io/grf/
Stanford Causal Inference Tutorial (Chapter 4): https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_lXOUlvQAoWZEqhRqHNezS30lI49G-

24

24

HTE Estimation Evaluation



Random Ranking of Treatment Effects

Far Away

Ranking of Treatment Effect by HTE Model

Close

Ranking of Ground-Truth HTEs

25

25

Cumulative Gain Chart

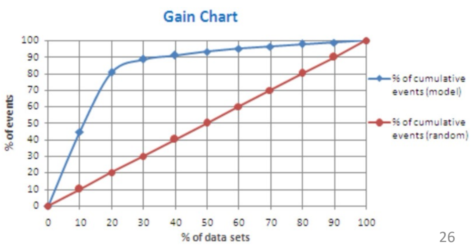
HTE Estimation Evaluation

Predicted Quantile	Treatment Sample #	Control Sample #	Cumulative Treatment Outcomes	Cumulative Control Outcomes
(90%, 100%]	N^T_1	N^C_1	Y^T_1	Y^C_1
(80%, 90%]	N^T_2	N^C_2	Y^T_2	Y^C_2
(70%, 80%]	N^T_3	N^C_3	Y^T_3	Y^C_3
…	…	…	…	…

Cumulative Gain of the first t Buckets (RCT)

$$CumulativeGain(t) = \left(\frac{\sum_{i=1}^{t} Y_i^T}{\sum_{i=1}^{t} N_i^T} - \frac{\sum_{i=1}^{t} Y_i^C}{\sum_{i=1}^{t} N_i^C}\right)\left(\sum_{i=1}^{t} N_i^T + \sum_{i=1}^{t} N_i^C\right)$$

AUC of the Cumulative Gain Chart would be a good metric to evaluate HTE estimation.



26

26