

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 88

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 88

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 88

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

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 88

- 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/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

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 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/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30I49G-

5

5

Decision Tree

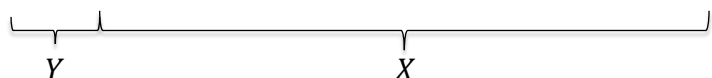
- 5000 data points

- Goal: 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

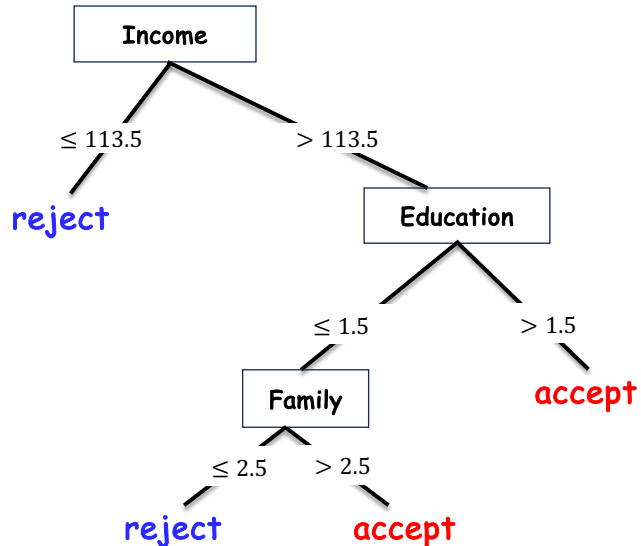


6

6

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.

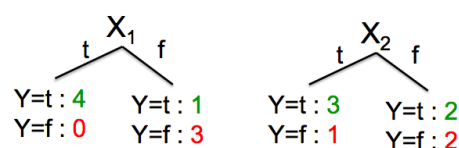


7

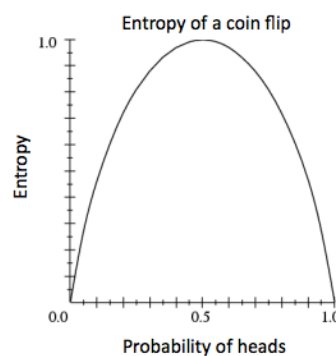
7

Fitting a Decision Tree

- Given a training set, identifying the best decision tree is **NP-complete**.
- Heuristics to build a tree:
 - Start with an **empty tree**.
 - Split on the feature that gives the **largest reduction in impurity**.
 - Repeat step 2 until a stopping criterion is met.
 - Assign the major class (classification) or the average outcome to each leaf (regression).
- Measurement of impurity:
 - Classification:** Gini index; Entropy
 - Regression:** Squared Error $\sum_i [(Y_i - \hat{\mu}(X_i))^2]$



On which feature should we split?



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

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 88

- 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/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
 Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOUlvQAoWZEghRqHNezS30II496-

9

9

From Decision Tree to Causal Tree

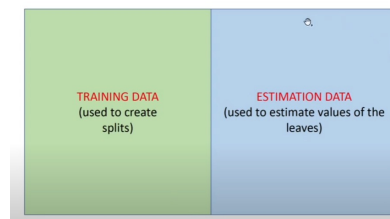
Recursive partitioning for heterogeneous causal effects

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 88

- 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

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:

Tree-split based on training data

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

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
 Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOUlvQAoWZEghRqHNezS30II496-

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 ...

☆ 保存 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 look at the difference between decision tree and causal tree.
- Original decision tree:

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

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

test set

training set

tree partitions based on training set

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

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 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 look at the difference between decision tree and causal tree.

Causal Trees:

$$\widehat{EMSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

test set

estimation set

tree partitions based on training set

treatment effects

only using training set

variance ...

rewards high heterogeneity

$$+ \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{l \in \Pi^{tr}} \left(\frac{S_{\mathcal{S}^{tr}}^2(l)}{p} + \frac{S_{\mathcal{S}^{est}}^2(l)}{1-p} \right)$$

Penalizes splits leading to small leafs

in the sample of control observations in leaf (l)

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

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 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/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

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 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/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

14

14

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 $s \propto n^\beta$ 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.

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

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

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

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

where the covariance is taken with respect to the set of all the trees $b = 1, \dots, B$ 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 $\hat{V}_{ij}(x) / \text{Var}[\hat{\tau}(x)] \rightarrow_p 1$.

15

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 .

- Draw a random subsample of size s from $\{1, \dots, n\}$ without replacement, and then divide it into two disjoint sets of size $|\mathcal{I}| = \lfloor s/2 \rfloor$ and $|\mathcal{J}| = \lfloor s/2 \rfloor$.
- 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.
- 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 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 .

- Draw a random subsample $\mathcal{I} \in \{1, \dots, n\}$ of size $|\mathcal{I}| = s$ (no replacement).
- 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.
- 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/stanfordqdsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II496-

16

Summary

- 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**.

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 引用 Cited by 3705 Related articles All 14 versions 》》

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanforddgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOUlvQAoWZEghRqHNezS30II49G-

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 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/stanforddgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOUlvQAoWZEghRqHNezS30II49G-

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 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/stanforddgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II49G-

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 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 .

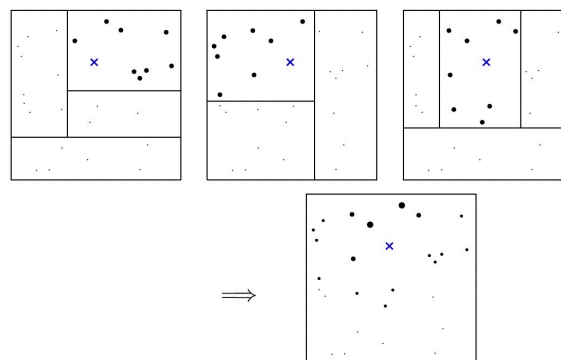


FIG. 1. Illustration of the random forest weighting function. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then the forest averages all these tree-based weightings, and 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/stanforddgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II49G-

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 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 $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1, \dots, n}$ satisfies

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

where

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

Definition 2. (GRF estimator)

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

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

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 containing 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/stanfordqsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II49G-

21

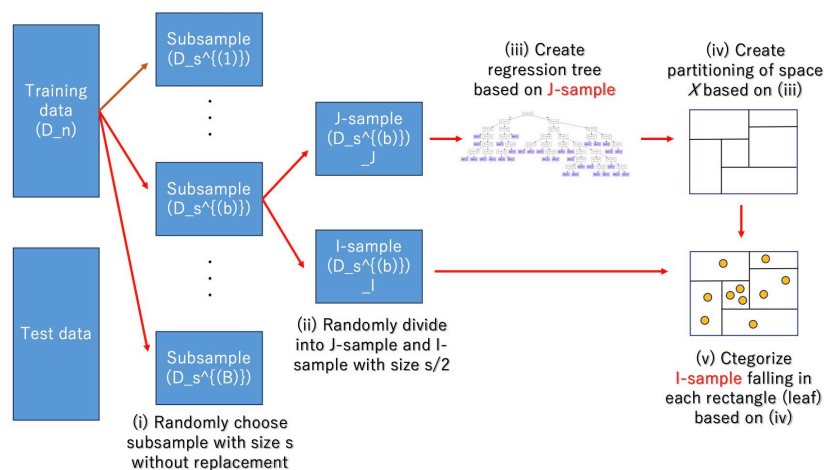
21

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 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/stanfordqsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II49G-

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 ...

☆ 保存 引 用 被引用次数：2435 相关文章 所有 18 个版本 Web of Science: 784

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++.

```

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

```

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.

grf package: <https://grf-labs.github.io/grf/>

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbslab/ml-ci-tutorial/hte-i-binary-treatment.html>

Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II49G-

23

23

Algorithm 2 Gradient tree

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

```

1: procedure GRADIENTTREE(set of examples  $\mathcal{J}$ , domain  $\mathcal{X}$ )
2:   node  $P_0 \leftarrow \text{CREATENODE}(\mathcal{J}, \mathcal{X})$ 
3:   queue  $Q \leftarrow \text{INITIALIZEQUEUE}(P_0)$ 
4:   while NOTNULL(node  $P \leftarrow \text{POP}(Q)$ ) do
5:      $(\hat{\theta}_P, \hat{v}_P, A_P) \leftarrow \text{SOLVEESTIMATINGEQUATION}(P)$   $\triangleright$  Computes (4) and (7).
6:     vector  $R_P \leftarrow \text{GETPSEUDOOUTCOMES}(\hat{\theta}_P, \hat{v}_P, A_P)$   $\triangleright$  Applies (8) over  $P$ .
7:     split  $\Sigma \leftarrow \text{MAKECARTSPLIT}(P, R_P)$   $\triangleright$  Optimizes (9).
8:     if SPLITSUCEEDED( $\Sigma$ ) then
9:       SETCHILDREN( $P$ , GETLEFTCHILD( $\Sigma$ ), GETRIGHTCHILD( $\Sigma$ ))
10:      ADDTOQUEUE( $Q$ , GETLEFTCHILD( $\Sigma$ ))
11:      ADDTOQUEUE( $Q$ , GETRIGHTCHILD( $\Sigma$ ))
12:  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 Q , unless 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.

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 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 Oprea, 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....

☆ 保存 引 用 被引用次数：140 相关文章 所有 8 个版本

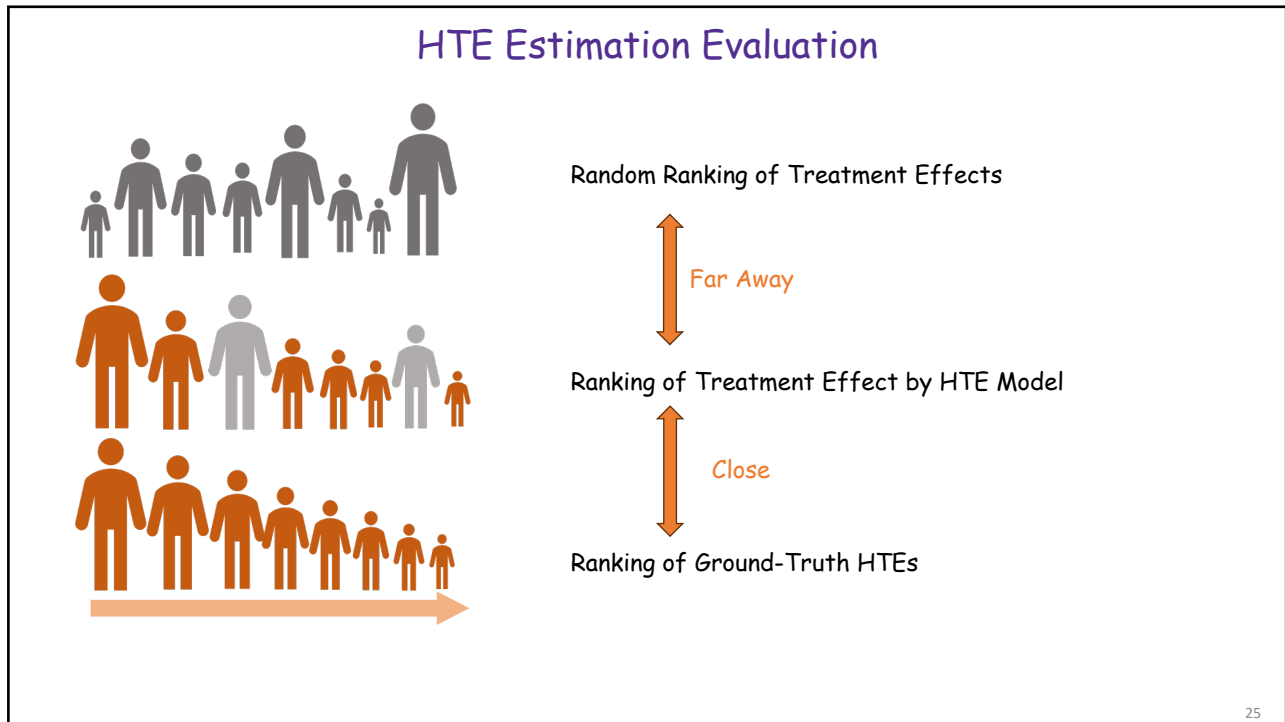
grf package: <https://grf-labs.github.io/grf/>

Stanford Causal Inference Tutorial (Chapter 4): <https://bookdown.org/stanfordqdsbslab/ml-ci-tutorial/hte-i-binary-treatment.html>

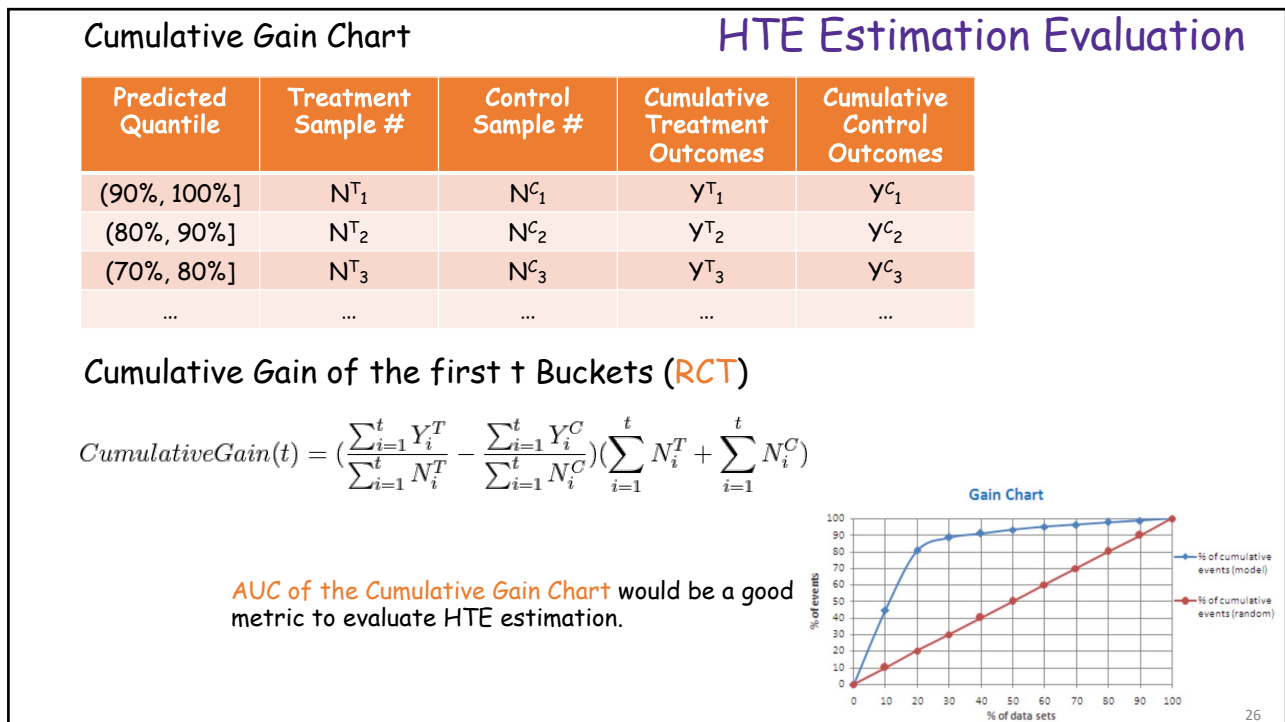
Stanford Causal Inference Course (Lecture 9-15): https://www.youtube.com/playlist?list=PLxq_IXOULvQAoWZEghRqHNezS30II49G-

24

24



25



26