Causal Inference and Machine Learning - Heterogeneous Trearment Effects

Alexander Quispe-The World Bank, ID4D
alexander.quispe@pucp.edu.pe

June 29, 2023

Motivation - CI & ML

In the last six years has appeared an explosion in a research collaboration between Economist (Inference) and Computer Scientists/Stats (Prediction).
What are the methods that combine CI & ML?

Susan Athey School (Stanford) - CATE	Victor Chernozhukov School (MIT) - HDM
Causal Trees Causal Forest Policy Learning Simulating data using WGANs Surrogate Index Contextual Bandits	High Dimensional Metrics (Lasso, Ridge, ElasticNet) Double Debiased Machine Learning (Lasso, Ridge, Neural Nets, Trees, Forest) DML with Sensitivity Analysis (Omitted Variable Bias)

Motivation - CI & ML

Given that most of these methodologies were proposed in academic environments, most of the packages/libraries are written in R.
However, if you transition to Industry, most Tech Firms use Python (Also Julia!!!).
Therefore, companies such as Microsoft and Uber have created Python packages that include these algorithms.

1. DoubleML - https://docs.doubleml.org/stable/index.html
2. EconML (Microsoft) - https://econml.azurewebsites.net/
3. CausalML (UBER) - https://github.com/uber/causalml
4. DoWhy - https://github.com/py-why/dowhy

PS. Sorry about Stata Lovers

Motivation - CI & ML

What if we want to start learning CI & ML from Zero to Hero? You can visit my open source project Dive into Causal Machine Learning

MGTECON 634: Machine Learning and Causal Inference - Standford

1. R Version
2. Python Version

14.388: Inference on Causal and Structural Parameters Using ML and AI - MIT

1. R Version
2. Python Version
3. Julia Version

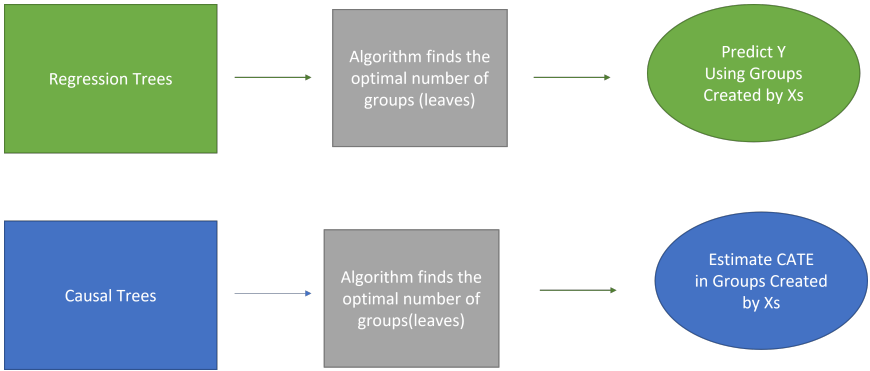
PS. Again, Sorry about Stata Lovers

Motivation - Causal Trees

- Most of our work in the WB is related to finding ATE/CATE estimators, but we **Pre-Specify analysis plan**
- Given that most of the time we use Surveys, and field experiments, we do not face the problem of HDM.
- **Goal: Data-driven search for heterogeneity in causal effects with valid standard errors**
- Applications : A/B Testing, Observational studies

Main Idea Behind Causal Trees

Figure: Trees vs Causal Trees



Conditional Average Treatment Effect

- We will learn how to estimate the conditional average treatment effect (CATE)

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

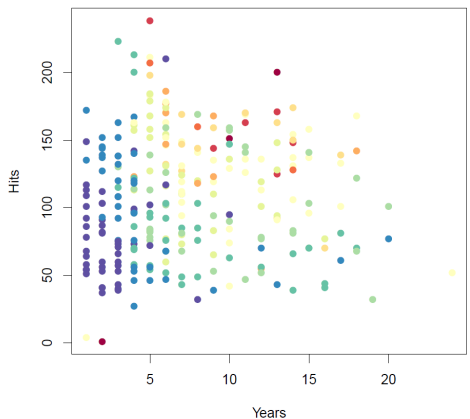
  This is the average treatment effect conditional on a vector of observable characteristics.
- However, When the observable covariates are high-dimensional, It can be hard to estimate without making strong modeling. Therefore, we will try to calculate treatment effect averages for simpler groups.

$$\tau(x) = E[Y_i(1) - Y_i(0)|G_i = g] \tag{2}$$

Decision Trees: Baseball salary data: how would you stratify it?

Salary is color-coded from low (blue, green) to high (yellow,red)

Figure: Salary-Hits-Years in Baseball (ISL page 328)



Decision tree for these data

Figure: Decision tree for baseball data (ISL page 329)

Details of previous figure

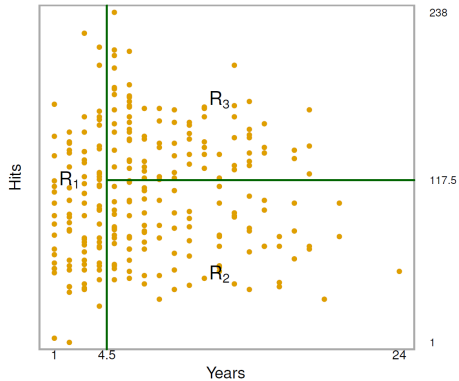
At a given node:

- $X_j < t_k$ indicates the left-hand branch emanating from that split
- $X_j >= t_k$ is the right-hand branch
- The left-hand branch corresponds to $Years < 4.5$
- The right-hand branch corresponds to $Years >= 4.5$
- The tree has **two internal nodes** and **three terminal nodes**, or **leaves**. The number in each leaf is **the mean of the response** for the observations that fall there.

Results

- $R_1 = \{X | Years < 4.5\}$
- $R_2 = \{X | Years >= 4.5, Hits < 117.5\}$
- $R_3 = \{X | Years >= 4.5, Hits > 117.5\}$

Figure: Segments (ISL page 329)



Terminology for Trees

- In keeping with the **tree** analogy, the regions R_1, R_2, and R_3 are known as **terminal nodes**
- Decision trees are typically drawn **upside down**, in the sense that the leaves are at the bottom of the tree.
- The points along the tree where the predictor space is split are referred to as **internal nodes**.
- In the hitters tree, the two internal nodes are indicated by the text $\mathit{Years} < 4.5$ and $\mathit{Hits} < 117.5$.

Interpretation of Results

- Years is the most important factor in determining **Salary**, and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of Hits that he made in the previous year seems to play little role in his **Hits Salary**.
- But among players who have been in the major leagues for five or more years, the number of Hits made in the previous year does affect Salary, and players who made more **Hits Salary Hits** last year tend to have higher salaries.

Tree-building process

- We divide the predictor space — that is, the set of possible values for $X_1, X_2, \ldots, X_p$ into J distinct and non-overlapping regions $R_1, R_2, \ldots R_J$
- For every observation that falls into the region R_j , we make the same prediction, which is simply the mean of the response values for the training observations in R_j.
- The goal is to find boxes $R1, \ldots, RJ$ that minimize the RSS, given by

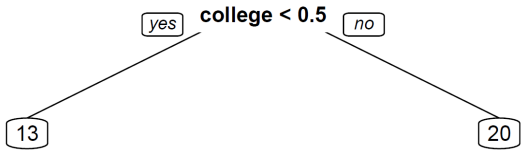
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{3}$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the *jth* box.

Top-down, greedy approach

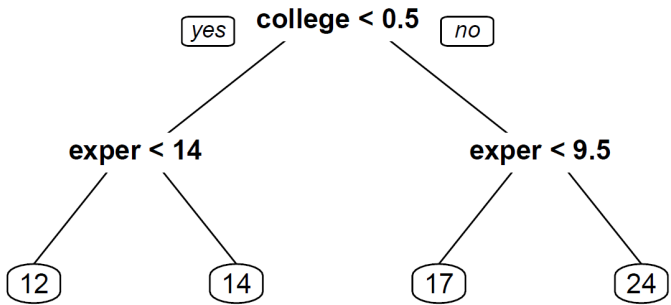
- **Top-down** because it begins at the top of the tree and then successively splits the predictor space
- **Greedy** because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.
- Example using wage data:

Figure: Wage example - Tree 1.(CMLB Ch3 Page 7)



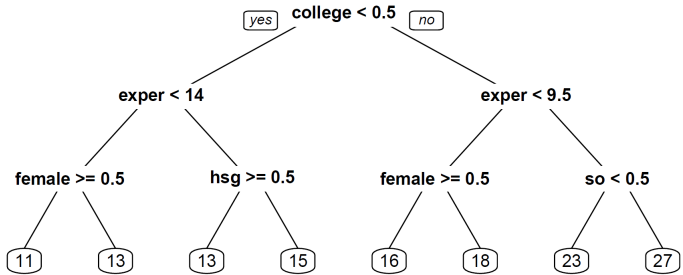
Top-down, greedy approach

Figure: Wage example - Tree 2 (CMLB Ch3 Page 7)



Top-down, greedy approach

Figure: Wage example - Tree 3 (CMLB Ch3 Page8)



Pruning Regression Trees

- First, the deeper we grow the tree, the better is our approximation to the regression function $g(Z)$
- The deeper the tree, the noisies our estimate $\hat{g}(Z)$ becomes, since there are fewer observations per terminal node to estimate the predicted value for this node.
- From a prediction point of view, we can try to find the right depth or the structure of the tree by cross-validation. For example, in the wage example the tree of depth 2 performs better in terms of cross-validated *MSE* than the trees of depth 3 or 1. The process of cutting down the branches of the tree to improve predictive performance is called **"Pruning the Tree"**.

Cost complexity pruning

we consider a sequence of trees indexed by a nonnegative tuning parameter α. For each value of α there corresponds a subtree $T \subset T0$ such that

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \quad T \subset T_o \tag{4}$$

(5)

$\theta \equiv P[\alpha_1 + 3 * \beta + U_i 1 \geq \max_{j=0,1...J}(\alpha_j + \beta * p_j + U_{ij})]$

1. T indicates the number of terminal nodes of the tree T
2. R_m is the rectangle, corresponding to the mth terminal node
3. $\hat{y}_{R_m}$ is the mean of the training observations in R_m

Cost complexity pruning

Figure: Tree algorithm

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. For each $k = 1, \ldots, K$:
   - 3.1 Repeat Steps 1 and 2 on the $\frac{K-1}{K}$th fraction of the training data, excluding the kth fold.
   - 3.2 Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013)

Causal trees and Tree Regression

Similarities

- Divide data into subgroups in order to maximize the discriminatory power of the split
- Both uses recursive binary splitting (greedy aproach) into covariate space
- Solve over-fitting problem through Croos- Validation (CV) to determine the depth of the tree

Causal trees and Tree Regression

Differences

- **Goal: estimate treatment effect heterogeneity**
- Optimize sub-groups to estimate treatment effect heterogeneity
- Divide population into subgroups to minimize MSE in treatment effects **(objective function)** instead in outcomes
- Use honesty tree, additionally, to address over-fitting.

How to address missing counterfactual

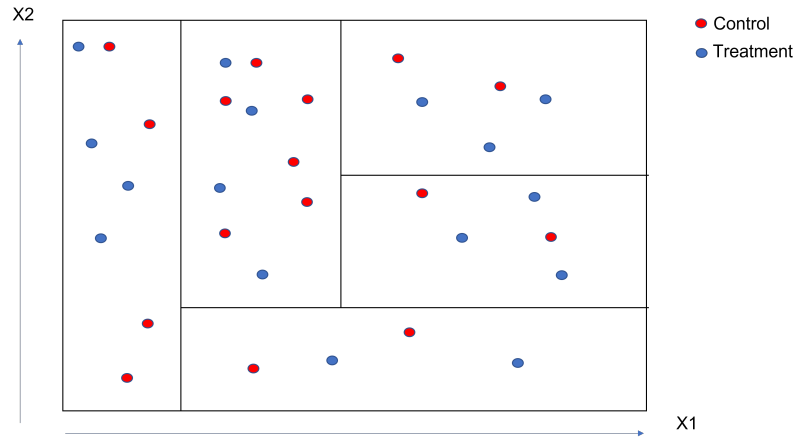
- Objective function is unfeasible $\sum_i[(\tau_i - \hat{\tau}(X_i))^2]$ (true treatment effect unobserved)

- Transform Outcome (Athey and Imbens, 2016)
- Applying off-the-shelf prediction to estimate CATE
- p is the assignment probability (if perfect RCT then p = 0.5).

$$Y_i^* = Y_i/p, \ \ if \ \ W_i = 1, \ \ and \ \ let \ \ Y_i^* = -Y_i/(1-p) \ \ if \ \ W_i = 0 \tag{6}$$

- $\mathbb{E}[Y_i^*|X_i = x] = \tau(x)$ is noisy for individual treatment effect, but unbiased estimate of CATE
- Use estimated propensity score $\hat{p}$ or AIPW score as outcome in observational studies.

Causal Tree Intuition



Honest Causal Tree

- Half sample to estimates tree and another half to estimate treatment effect.
- **Tradeoff**
- COST: sample splitting means build shallower tree, less personalized predictions and lower MSE of treatment effects
- BENEFIT: valid confidence intervals with coverage rates that do not deteriorate as data generating process get more complex or more covariates are added.
- **Inference**
- Can separate tree construction from treatment effect estimation
- tree constructed on training is independent of test sample
- Holding tree form training sample fixed, can use standard methods to conduct inference within leaf of the tree on test sample

Partition and Leaf Effects Estimates

- Three samples: model tree construction S^{tr} , estimation sample for leaf effects S^{est} and a test sample S^{te}.
- Sample average treatment effect in sample S^{est} for the leaf $l(X_i, \Pi)$ associated with covariates X_i

$$\hat{\tau}(X_i, S^{est}, \Pi) = \frac{1}{\sum_{j \in S^{est} \cap l(X_i, \Pi)} W_i} \sum_{i \in S^{est} \cap l(X_i, \Pi)} W_i Y_i \; -$$

$$\frac{1}{\sum_{j \in S^{est} \cap l(X_i, \Pi)} (1 - W_i)} \sum_{i \in S^{est} \cap l(X_i, \Pi)} (1 - W_i) Y_i$$

Estimating the MSE of treatment effects (EMSE)

- criterion for evaluating a partition Π anticipating re-estimating leaf effects using sample splitting:

$$MSE(S^{est}, S^{te}) = \frac{1}{n^{te}} \sum_{i \in S^{te}} (\tau_i - \hat{\tau}(X_i, S^{est}, \Pi))^2 \tag{7}$$

$$MSE(S^{est}, S^{te}) = \frac{1}{n^{te}} \sum_{i \in S^{te}} (\tau_i^2 - 2\tau_i * \hat{\tau}(X_i, S^{est}, \Pi) + \hat{\tau}^2(X_i, S^{est}, \Pi)) \tag{8}$$

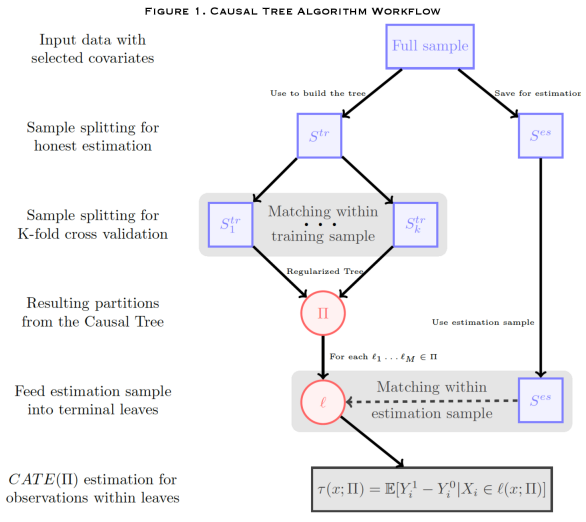
$$EMSE = E_{S^{est}, S^{te}}[MSE(S^{est}, S^{te})] \tag{9}$$

$$EMSE = \mathbf{V}_{S^{est}, S^{te}}[\hat{\tau}(X_i; \Pi, S^{est}] - E_{X_i}[\tau^2(X_i; \Pi)] + E[\tau_i^2] \tag{10}$$

Causal Tree Algorithm

- Divide data into tree-building S^{tr} and estimation S^{est} samples
- Use recursively partition in covariate space X to deep partition Π
  - Each split is selected as the one that minimizes mean - square error estimation of treatment effect over all possible binary splits
  - **Preserve minimun number of treated and control units in each child leaf**
- Use cross - validation to select the depth d^* of the partition
- Select partition Π^* by pruning Π to depth d^*
- Estimate the treatment effects in each leaf of Π^* using estimation sample

Causal Tree Algorithm



Causal Tree - Example 1

Figure: General Social Survey (GSS) (MGTECON 634 Tutorial)

The General Social Survey is an extensive survey, collected since 1972, that seeks to measure demographics, political views, social attitudes, etc. of the U.S. population.

Of particular interest to us is a **randomized experiment**, for which we have data between 1986 and 2010.

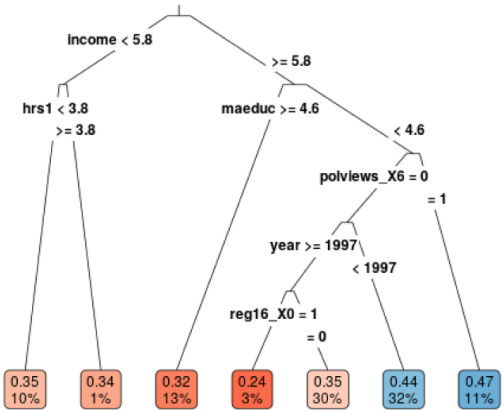
- **Question A**: Are we spending too much, too little, or about the right amount on **welfare**?
- **Question B**: Are we spending too much, too little, or about the right amount on **assistance to the poor**?

Treatment effect: how much less likely are people to answer **too much** to question B than to question A.

- We want to understand how the treatment effect depends on **covariates**: political views, income, age, hours worked, ...

Causal Tree - Example 1

Figure: Causal Tree (MGTECON 634 Tutorial)

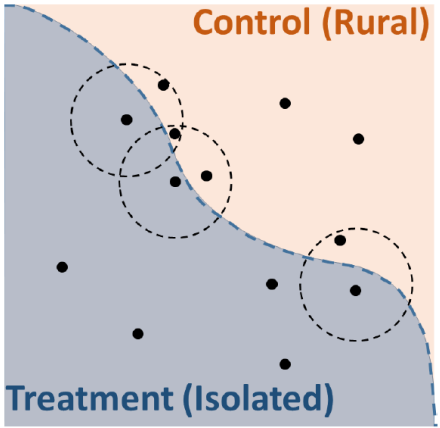



```
Linear hypothesis test

Hypothesis:
leaf1:W - leaf2:W = 0
leaf1:W - leaf3:W = 0
leaf1:W - leaf4:W = 0
leaf1:W - leaf5:W = 0
leaf1:W - leaf6:W = 0
leaf1:W - leaf7:W = 0

Model 1: restricted model
Model 2: Y ~ leaf + W:leaf

  Res.Df Df      F    Pr(>F)
1   5272
2   5266  6 4.4771 0.0001575 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Causal Tree - Example 2

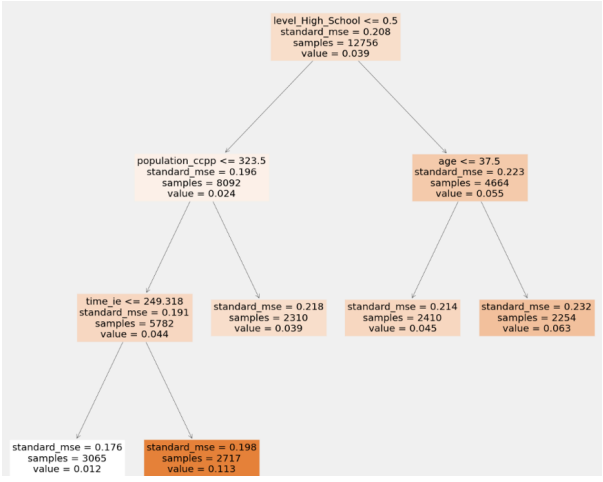
Estimating Heterogeneous Effects : How cash bonuses retain teachers in remote locations in Peru.

Figure: Unobserved Boundaries (Perez-Leon 2018)



Causal Tree - Example 2

Estimating Heterogeneous Effects : How cash bonuses retain teachers in remote locations in Peru.



Citation

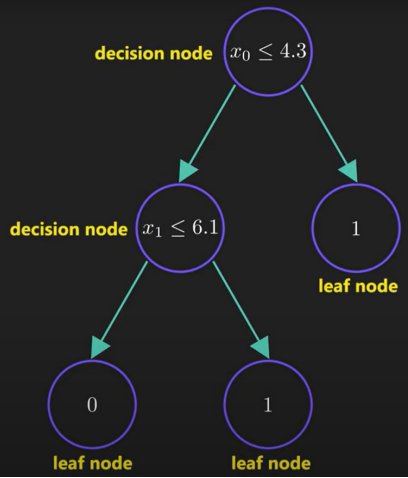
Images are taken from Sujan Dutta notes on Random Forest. Notes on Causal Trees and Causal Forest are taken from Susan Athey Lecture Notes on Machine Learning and Causal Inference, 2021.

Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

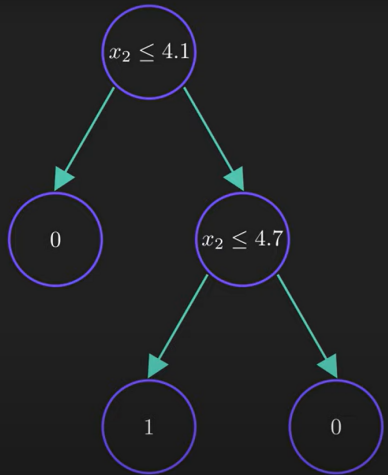
Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

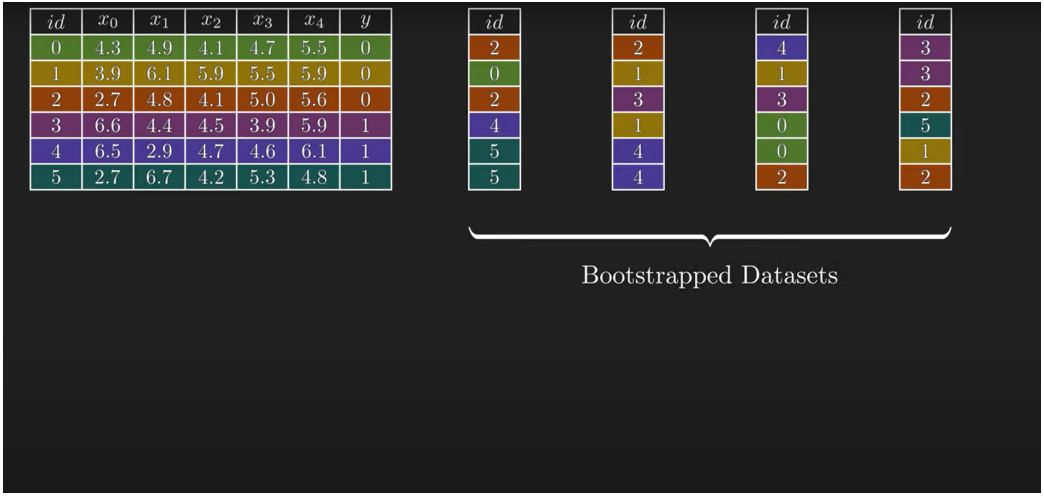
id
2
0
2
4
5
5

id
2
1
3
1
4
4

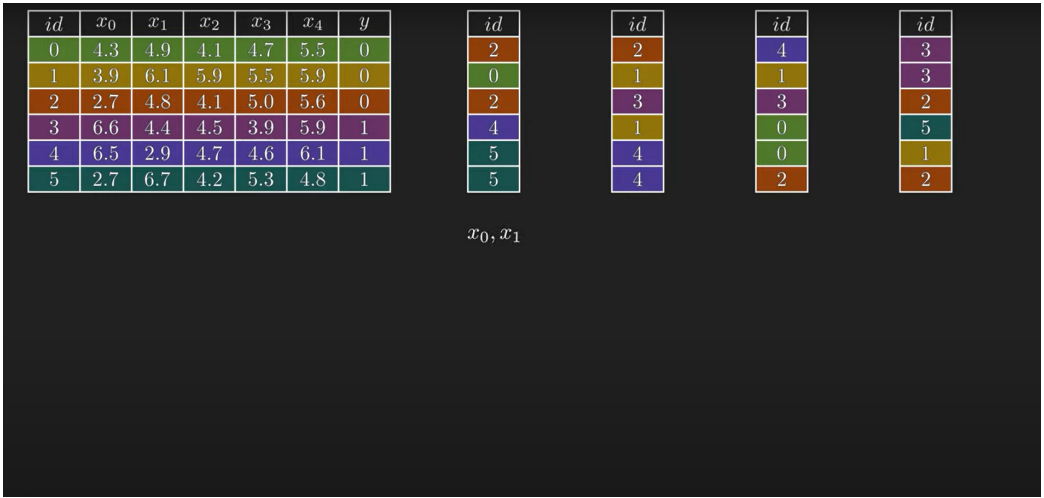
id
4
1
3
0
0
2

id
3
3
2
5
1
2

Random Forest



Random Forest



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

x_0, x_1

id
2
1
3
1
4
4

x_2, x_3

id
4
1
3
0
0
2

x_2, x_4

id
3
3
2
5
1
2

x_1, x_3

Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

x_0, x_1

id
2
1
3
1
4
4

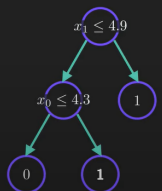
x_2, x_3

id
4
1
3
0
0
2

x_2, x_4

id
3
3
2
5
1
2

x_1, x_3



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

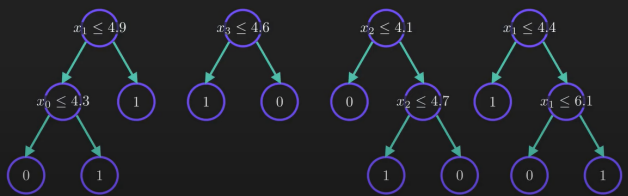
id
2
0
2
4
5
5

id
2
1
3
1
4
4

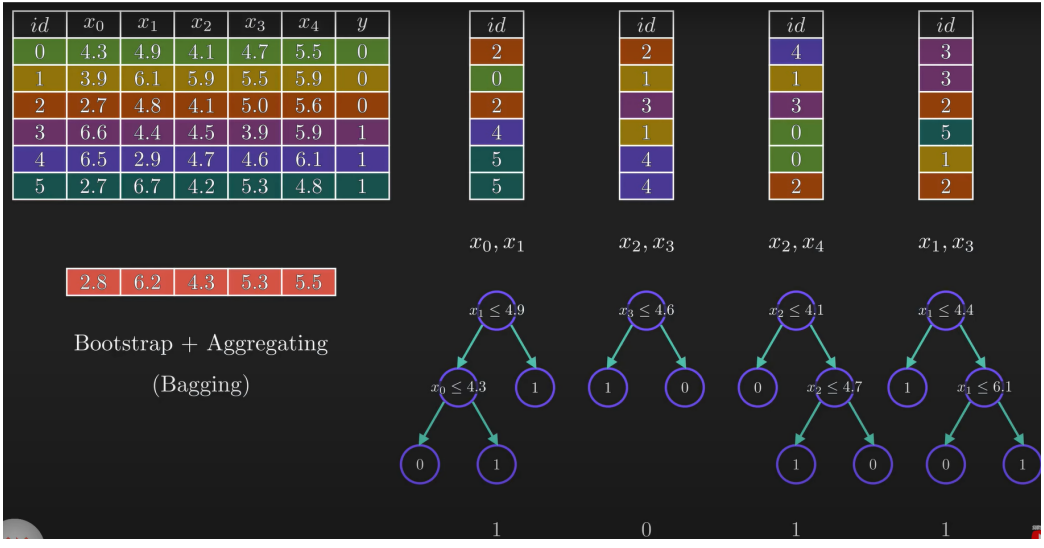
id
4
1
3
0
0
2

id
3
3
2
5
1
2

x_0, x_1 x_2, x_3 x_2, x_4 x_1, x_3



Random Forest



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

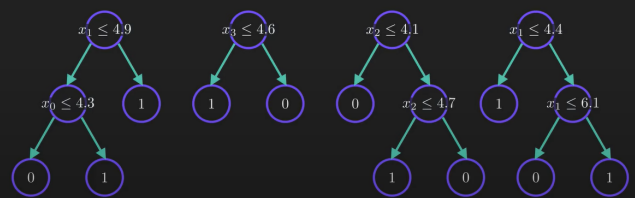
id
2
1
3
1
4
4

id
4
1
3
0
0
2

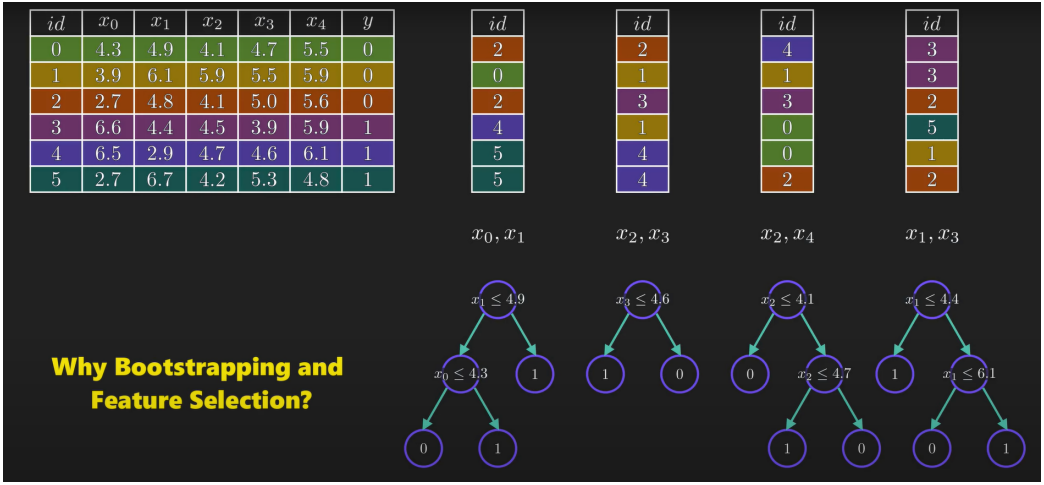
id
3
3
2
5
1
2

x_0, x_1 x_2, x_3 x_2, x_4 x_1, x_3

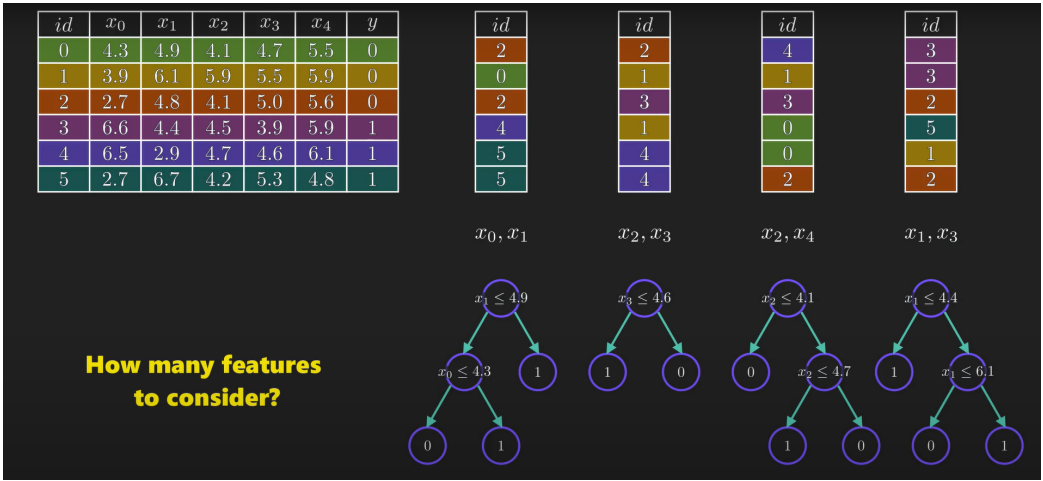
Why is it called random?



Random Forest



Random Forest



Random Forest - Concluding Remarks

From ISL2 we have :

- Random forests provide an improvement over bagged trees by way of a small tweak that decorrelates the trees. This reduces the variance when we average the trees.
- But when building these decision trees, each time a split in a tree is considered, **a random selection of m predictors is chosen** as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$, that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors

Baseline method: k-NN matching

Consider the k-**NN matching** estimator for $\tau(x)$:

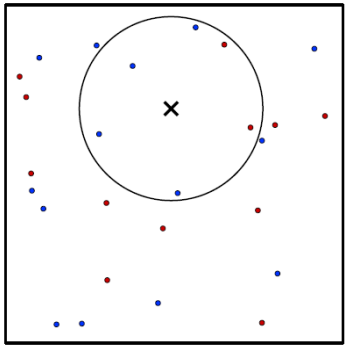
$$\hat{\tau}(x) = \frac{1}{k} \sum_{\mathcal{S}_1(x)} Y_i - \frac{1}{k} \sum_{\mathcal{S}_0(x)} Y_i,$$

where $\mathcal{S}_{0/1}(x)$ is the set of k-nearest cases/controls to x. This is consistent given **unconfoundedness** and regularity conditions.

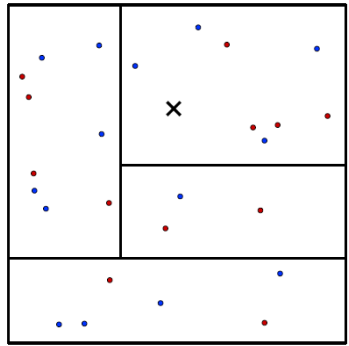
- **Pro:** Transparent asymptotics and good, robust performance when p is small.
- **Con:** Acute curse of dimensionality, even when $p = 20$ and $n = 20k$.

NB: Kernels have similar qualitative issues as k-NN.

Making k-NN matching adaptive



Euclidean neighborhood, for k-NN matching.



Tree-based neighborhood.

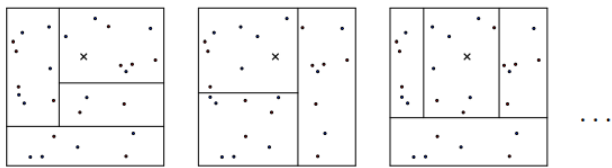
From Trees to Random Forest

- Training set $(X_i, Y_i, W_i)_{i=1}^n$, while tree predictor for a test point (x)

$$\hat{\tau} = T(x; X_i, Y_i, W_{ii=1}^n) \tag{11}$$

- Random Forest: build and average many different trees T^*
- Create alternative tress (T_b^*) by bagging (sampling with replacement) or sub-sampling the training set

$$\hat{\tau} = \frac{1}{B} \sum_{b=1}^{B} T_b^*(x; X_i, Y_i, W_{i i=1}^n) \tag{12}$$



Statistical inference with regression forest

Regresssion forest are asymptotically Gaussian and centered:

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \rightarrow \mathbb{N}(0,1), \sigma_n^2(x) \rightarrow_p 0 \tag{13}$$

technical conditions

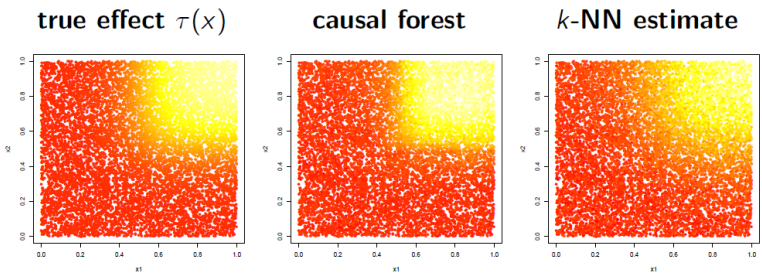
- Individual trees are honest (**Honesty**)
- Individual trees built random sub-samples of size s n^{β}, where $\beta_{min} < \beta < 1$ (**Subsampling**)
- X_i density from 0 and ∞ (**Continuos features**)
- Conditional mean function $\mu(x) = \mathbb{E}[Y|X = x]$ is Lipschitz continues (**Lipschitz response**)

Causal Forest Example

Figure: True effect, Causal Forest, KNN estimate

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 6$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

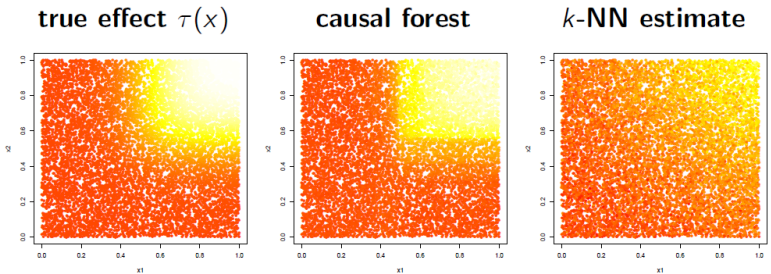


Causal Forest Example

Figure: True effect, Causal Forest, KNN estimate

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 20$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.



Sources

- Lecture Notes of Susan Athey's Machine Learning and Causal Inference Course at Stanford.
- Lecture Notes of Victor Chernozhukov's nference on Causal and Structural Parameters Using ML and AI at MIT .
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Perez-Leon (2018). Inducing Teacher Retention in Remote Locations: Evidence from Peru.