Lecture 21: More on Trees (w. causality)

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

October 26, 2021

Agenda

1. Recap w. Example: Beauty in the classroom
2. Causal Trees
   - Causality Review: ATE, CATE, HTE
   - Empirical Example
3. Causal Trees: Theory Details
   - Honest Inference for Treatment Effects
   - Observational Studies with Unconfoundedness
4. Review & Next Steps
5. Further Readings

CART: Recap w. Example: Beauty in the classroom

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Economics of Education Review 24 (2005) 369–376

Economics of Education Review

www.elsevier.com/locate/econedurev

Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity

Daniel S. Hamermesh*, Amy Parker

Department of Economics, University of Texas, Austin, TX 78712-1173, USA

Received 14 June 2004; accepted 21 July 2004

CART: Recap w. Example: Beauty in the classroom

Motivation

- An immense literature in social psychology (summarized by Hatfield and Sprecher, 1986) has examined the impact of human beauty on a variety of noneconomic outcomes.
- Economists have considered how beauty affects labor market outcomes, particularly earnings (Hamermesh and Biddle, 1994; Biddle and Hamermesh, 1998).
- The impacts on these monetary outcomes are implicitly the end results of the effects of beauty on productivity;
- But there seems to be no direct evidence of the impacts of beauty on productivity in a context in which we can be fairly sure that productivity generates economic rewards.

CART: Recap w. Example: Beauty in the classroom

Motivation

- A substantial amount of research has indicated that academic administrators pay attention to teaching quality in setting salaries (Becker and Watts, 1999).
- The question is what generates the measured productivity for which the economic rewards are being offered.
- One possibility is simply that descriptive characteristics, such as beauty, trigger positive responses by students and lead them to evaluate some teachers more favorably, so that their beauty earns them higher economic returns.

CART: Recap w. Example: Beauty in the classroom

Motivation

- They take a sample of student instructional ratings for a group of university teachers and acquire six independent measures of their beauty, and a number of other descriptors of them and their classes.

```
data("TeachingRatings", package = "AER")
tr <- subset(TeachingRatings, credits == "more")
str(tr)
```

```
## 'data.frame':    436 obs. of  12 variables:
##  $ minority   : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ age        : int  36 59 51 40 31 62 33 51 33 47 ...
##  $ gender     : Factor w/ 2 levels "male","female": 2 1 1 2 2 1 2 2 2 1 ...
##  $ credits    : Factor w/ 2 levels "more","single": 1 1 1 1 1 1 1 1 1 1 ...
##  $ beauty     : num  0.29 -0.738 -0.572 -0.678 1.51 ...
##  $ eval       : num  4.3 4.5 3.7 4.3 4.4 4.2 4 3.4 4.5 3.9 ...
##  $ division   : Factor w/ 2 levels "upper","lower": 1 1 1 1 1 1 1 1 1 1 ...
##  $ native     : Factor w/ 2 levels "yes","no": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tenure     : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 1 ...
##  $ students   : num  24 17 55 40 42 182 33 25 48 16 ...
##  $ allstudents: num  43 20 55 46 48 282 41 41 60 19 ...
##  $ prof       : Factor w/ 94 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

CART: Recap w. Example: Beauty in the classroom

OLS

```
tr_lm <- lm(eval ~ beauty + gender + minority + native + tenure + division,
  data = tr, weights = students)

tr_lm_gender <- lm(eval ~ beauty:gender + minority + native + tenure + division,
  data = tr, weights = students)

tr_lm_male <- lm(eval ~ beauty + minority + native + tenure + division,
  data = tr[tr$gender=="male",], weights = students)

tr_lm_female <- lm(eval ~ beauty + minority + native + tenure + division,
  data = tr[tr$gender=="female",], weights = students)

stargazer::stargazer(tr_lm,tr_lm_gender,tr_lm_male,tr_lm_female,type="text")
```

CART: Recap w. Example: Beauty in the classroom

OLS

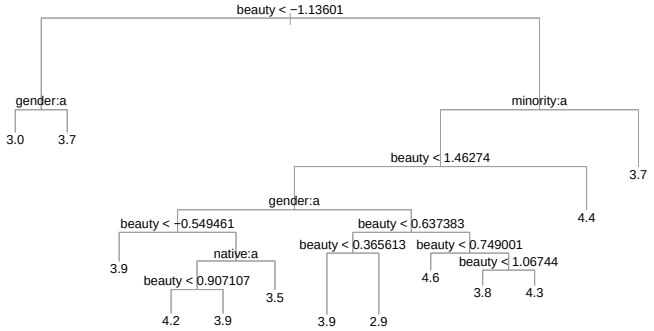
	Dependent variable:			
	eval			
	(1)	(2)	(3)	(4)
beauty	0.283***		0.383***	0.133***
	(0.028)		(0.037)	(0.045)
genderfemale	-0.213***			
	(0.048)			
minorityyes	-0.327***	-0.363***	-0.014	-0.279***
	(0.086)	(0.084)	(0.194)	(0.098)
nativeno	-0.217*	-0.229*	-0.388**	-0.288
	(0.125)	(0.124)	(0.188)	(0.175)
tenureyes	-0.132**	-0.032	-0.053	-0.064
	(0.065)	(0.065)	(0.088)	(0.098)
divisionlower	-0.050	-0.071	0.004	-0.244***
	(0.044)	(0.045)	(0.054)	(0.078)
beauty:gendermale		0.407***		
		(0.038)		
beauty:genderfemale		0.132***		
		(0.043)		
Constant	4.216***	4.058***	4.101***	4.027***
	(0.068)	(0.062)	(0.089)	(0.084)
Observations	436	436	250	186
R2	0.271	0.275	0.335	0.170
Note:				*p<0.1; **p<0.05; ***p<0.01

CART: Recap w. Example: Beauty in the classroom

Trees

```
library("tree")

pstree <- tree(eval ~beauty + gender + minority + native + tenure + division, data=tr, mincut=1)
par(mfrow=c(1,1))
plot(pstree, col=8)
text(pstree, digits=2)
```



CART: Recap w. Example: Beauty in the classroom

Trees: Constructing the partition

- How to choose the partition?
- Start with the trivial partition with one element
- Greedy algorithm (CART): Iteratively split an element of the partition, such that the in-sample prediction improves as much as possible.
- That is: Given $(R_1 \dots, R_M)$
  - For each R_m $m = 1, \dots, M$ and
  - For each X_j $j = 1, \dots, p$
  - find the $s_{j,m}$ that minimizes the mean squared error, if we split R_m along variable X_j at $s_{j,m}$
  - then pick the (m, j) that minimizes the MSE and construct a new partition with $M + 1$ elements
  - Iterate

CART: Recap w. Example: Beauty in the classroom

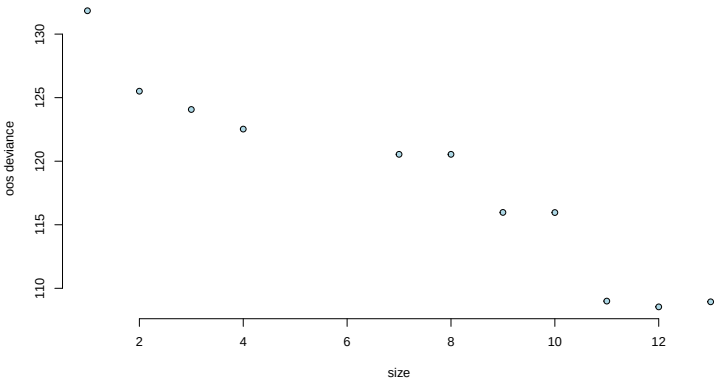
Trees: Tuning and pruning

- Key tuning parameter: Total number of splits M.
- We can optimize this via cross-validation.
- CART can furthermore be improved using "pruning."
- Idea:
  - Fit a flexible tree (with large M) using CART.
  - Then iteratively remove (collapse) nodes.
  - To minimize the sum of squared errors,
  - plus a penalty for the number of elements in the partition.
- This improves upon greedy search. It yields smaller trees for the same mean squared error.

CART: Recap w. Example: Beauty in the classroom

Trees

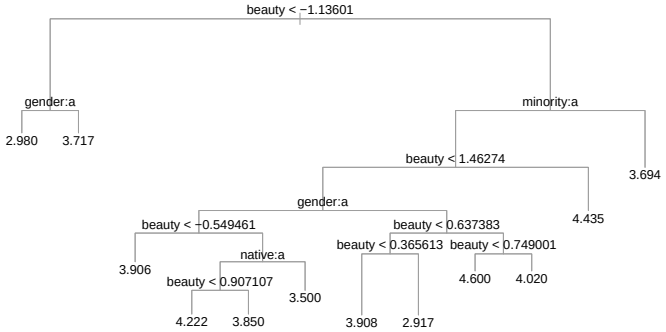
```
## Use cross-validation to prune the tree
cvpst <- cv.tree(pstree, K=10)
plot(cvpst$size, cvpst$dev, xlab="size", ylab="oos deviance", pch=21, bg="lightblue", bty="n")
```



CART: Recap w. Example: Beauty in the classroom

Trees

```
pstcut <- prune.tree(pstree, best=12)
plot(pstcut, col=8)
text(pstcut)
```



CARTs

- Smart way to represent nonlinearities. Most relevant variables on top.
- Very easy to communicate.
- Reproduces human decision-making process.
- Trees are intuitive and do OK, but
  - They are not very good at prediction
  - If the structure is linear, CART does not work well.
  - Not very robust

Causal Trees

Treatment Effects Review

- We observe a sequence of triples $\{(W_i, Y_i, X_i)\}_i^N$, where
  - $W_i \in \{0, 1\}$: is a binary variable indicating whether the individual was treated (1) or not (0)
  - $Y_i^{obs} \in \mathbb{R}$: a real variable indicating the observed outcome for that individual
  - X_i: is a p-dimensional vector of observable pre-treatment characteristics
- Moreover, in the Neyman-Rubin potential-outcomes framework, we will denote by
  - $Y_i(1)$: the outcome unit i would attain if they received the treatment
  - $Y_i(0)$: the outcome unit i would attain if they were part of the control group

Treatment Effects Review

The individual treatment effect for subject i can then be written as

$$Y_i(1) - Y_i(0)$$

Unfortunately, in our data we can only observe one of these two potential outcomes.

Education (X_i)	Treated W_i	No Subsidy $Y_i(0)$	Subsidy $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
High	1	?	$Y_1(1)$	?
High	0	$Y_2(0)$	?	?
Low	0	$Y_3(0)$	?	?
Low	1	?	$Y_4(1)$	?

Using the potential outcome notation above, the observed outcome can also be written as

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

Average Treatment Effects Review

- Computing the difference for each individual is impossible.
- But we can get the Average Treatment Effect (ATE):

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

- Heterogeneous Treatment Effects: Same treatment may affect different individuals differently
- Conditional Average Treatment Effect (CATE)

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

Heterogeneous Treatment Effects Review

Concerns

- Issues:
  - Ad hoc searches for particularly responsive subgroups may mistake noise for a true treatment effect.
  - Concerns about ex-post "data-mining" or p-hacking
    - preregistered analysis plan can protect against claims of data mining
    - But may also prevent researchers from discovering unanticipated results and developing new hypotheses
- But how is researcher to predict all forms of heterogeneity in an environment with many covariates?
- Athey and Imbens to the rescue
  - Allow researcher to specify set of potential covariates
  - Data-driven search for heterogeneity in causal effects with valid standard errors

Causal Trees: Empirical Example (Green and Kern)

- To illustrate how it works let me use this experiment from the General Social Survey (GSS)
- GSS conducts surveys regular surveys on Americans think feel about different issues
- For decades, scholars studying Americans' support for social welfare spending have noted the special disdain that americans harbor for programs labeled "welfare"
- This phenomenon became the subject of sustained experimental inquiry in the mid-1980s, when the GSS included a question-wording experiment in its national survey of adults.

Causal Trees: Empirical Example

- Respondents in each survey were randomly assigned to one of two questions about public spending.
- *"too much" money is spent on assistance to the Poor (control) or Welfare (treatment)*
- Various explanations put forward: stereotypes associated with welfare recipients and poor people, particularly racial stereotypes, and to political orientations such as individualism and conservatism .
- Some authors consider the interaction between the treatment and attributions, e.g.
  - Federico (2004) examines a complicated three-way interaction between the treatment, education, and racial perceptions.
  - Jacoby (2000) suggests that party and ideology may make some respondents especially receptive to the more specific program (should strong and weak Democrats be treated as separate subgroups or should they be combined?)

Causal Trees

```
#load packages
library(fBasics)     # Summary statistics
library(rpart)       # Classification and regression trees
library(rpart.plot)  # Plotting trees
library(treeClust)   # Predicting leaf position for causal trees
library(car)         # linear hypothesis testing for causal tree
library(kableExtra)  # Tables
library(causalTree) # For causal trees (Athey and Imbens, 2016)
library(dplyr) # For data wrangling
# Set seed for reproducibility
set.seed(201911)
# Load Data
df<-readRDS("welfare.rds")
str(df)
```

```
'data.frame':   13198 obs. of  34 variables:
 $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Y       : num  0 0 1 1 1 0 0 0 1 0 ...
 $ W       : num  1 1 1 0 0 1 1 0 0 1 ...
 $ hrs1    : num  40 35 30 40 35 38 27 40 32 50 ...
 $ partyid : num  4 1 2 2 1 2 0 1 3 3 ...
 $ income  : num  12 12 12 12 11 12 12 11 12 12 ...
 ...
```

Causal Trees

ATE

```
difference_in_means <- function(dataset) {
  treated_idx <- which(dataset$W == 1)
  control_idx <- which(dataset$W == 0)

  # Filter treatment / control observations, pulls outcome variable as a vector
  y1 <- dataset[treated_idx, "Y"] # Outcome in treatment grp
  y0 <- dataset[control_idx, "Y"] # Outcome in control group

  n1 <- sum(df[,"W"])     # Number of obs in treatment
  n0 <- sum(1 - df[,"W"]) # Number of obs in control

  # Difference in means is ATE
  tauhat <- mean(y1) - mean(y0)

  # 95% Confidence intervals
  se_hat <- sqrt( var(y0)/(n0-1) + var(y1)/(n1-1) )
  lower_ci <- tauhat - 1.96 * se_hat
  upper_ci <- tauhat + 1.96 * se_hat

  return(c(ATE = tauhat, lower_ci = lower_ci, upper_ci = upper_ci))
}

tauhat_rct <- difference_in_means(df)
tauhat_rct
```

```
       ATE   lower_ci   upper_ci
-0.3697802 -0.3841123 -0.3554481
```

Causal Trees

ATE

```
outcome_variable_name <- "Y"
treatment_variable_name <<- "W"
covariate_names <<- c("hrs1", "partyid", "income", "rincome",
                      "wrkstat", "wrkslf","age", "polviews",
                      "educ", "earnrs", "race","wrkslf",
                      "marital","sibs","childs", "occ80",
                      "prestg80", "indus80","res16","reg16",
                      "mobile16", "family16", "parborn",
                      "maeduc","degree","sex","race",
                      "born","hompop","babies",
                      "preteen","teens","adults")

fmla <- paste("Y ~ W +",paste(covariate_names, collapse = " + "))
print( fmla)
```

```
[1] "Y ~ W + hrs1 + partyid + income + rincome + wrkstat + wrkslf + age
+ polviews + educ + earnrs + race + wrkslf + marital + sibs + childs
+ occ80 + prestg80 + indus80 + res16 + reg16 + mobile16 + family16
+ parborn + maeduc + degree + sex + race + born + hompop + babies + preteen + teens + adults"
```

Causal Trees

ATE

```
reg_simple<-lm(Y~W,data=df)
reg_controls<-lm(fmla,data=df)
stargazer::stargazer(reg_simple,reg_controls,type="latex")
```

Table 1

	Dependent variable:	
	Y	
	(1)	(2)
W	-0.370^{***}	-0.368^{***}
	(0.007)	(0.007)
Constant	0.481^{***}	0.223^{***}
	(0.005)	(0.069)
Controls	No	Yes
Observations	13,198	13,198
R^2	0.166	0.215

Causal Trees

HTE

- We need to proceed in steps
- Step 1: Split the dataset. Why? $\rightarrow$ Athey and Imbens innovation
  - In order to ensure valid estimates of the treatment effect within each subgroup, Athey and Imbens propose a sample-splitting approach that they refer to as honesty:
  - a method is honest if it uses one subset of the data to estimate the model parameters, and a different subset to produce estimates given these estimated parameters.
  - In the context of causal trees, honesty implies that the asymptotic properties of treatment effect estimates within leaves are the same as if the tree partition had been exogenously given, and it is one of the assumptions required to produce unbiased and asymptotically normal estimates of the treatment effect.

Causal Trees

HTE

- Divide the data 40%-40%-20% for honest estimation and validation.

```
train_fraction <- 0.80  # Use train_fraction % of the dataset to train our models

df_train <- sample_frac(df, replace=F, size=train_fraction)
df_test <- anti_join(df,df_train, by = "ID")#need to check on larger datasets


split_size <- floor(nrow(df_train) * 0.5)
df_split <- sample_n(df_train, replace=FALSE, size=split_size)

# Make the splits
df_est <- anti_join(df_train,df_split, by ="ID")
```

Causal Trees

- Step 2: Fit the tree

```
fmla_ct <- paste("factor(Y) ~", paste(covariate_names, collapse = " + "))

ct_unpruned <- honest.causalTree(
  formula = fmla_ct,            # Define the model
  data = df_split,              # Subset used to create tree structure
  est_data = df_est,            # Which data set to use to estimate effects

  treatment = df_split$W,       # Splitting sample treatment variable
  est_treatment = df_est$W,     # Estimation sample treatment variable

  split.Rule = "CT",            # Define the splitting option
  cv.option = "TOT",            # Cross validation options
  cp = 0,                       # Complexity parameter

  split.Honest = TRUE,          # Use honesty when splitting
  cv.Honest = TRUE,             # Use honesty when performing cross-validation

  minsize = 25,                 # Min. number of treatment and control cases in each leaf
  HonestSampleSize = nrow(df_est)) # Num obs used in estimation after building the tree
```

Causal Trees

- Step 3: Crossvalidate

```
# Table of cross-validated values by tuning parameter.
ct_cptable <- as.data.frame(ct_unpruned$cptable)
# Obtain optimal complexity parameter to prune tree.
selected_cp <- which.min(ct_cptable$xerror)
optim_cp_ct <- ct_cptable[selected_cp, "CP"]
# Prune the tree at optimal complexity parameter.
ct_pruned <- prune(tree = ct_unpruned, cp = optim_cp_ct)
ct_pruned
```

```
n= 5279

node), split, n, deviance, yval
      * denotes terminal node

 1) root 5279 912.78610 -0.3753160
   2) partyid>=1.5 3530 654.60930 -0.3822570
     4) polviews>=3.5 2826 532.11460 -0.3997024
       8) reg16>=0.5 2658 500.98290 -0.4043439
        16) hrs1>=44.5 1063 203.85490 -0.4271320
          32) wrkslf< 1.5 182  35.70208 -0.4264330
            64) indus80< 526 81  15.81056 -0.3757764 *
            65) indus80>=526 101  19.59552 -0.4610849 *
          33) wrkslf>=1.5 881 167.91090 -0.4283712 *
        17) hrs1< 44.5 1595 295.01770 -0.3896395 *
       9) reg16< 0.5 168  30.09444 -0.3315372 *
     5) polviews< 3.5 704 115.34030 -0.3167446 *
```

Causal Trees

- Step 4: Predict point estimates (on estimation sample)

```
tauhat_ct_est <- predict(ct_pruned, newdata = df_est)
head(tauhat_ct_est)
```

```
         1          2          3          4          5          6
-0.3843850 -0.4283712 -0.3843850 -0.3896395 -0.3896395 -0.3843850
```

Causal Trees

- Step 5: Compute standard errors
- The `causalTree` package does not compute standard errors by default, but we can compute them using the following trick.
  - First, define L_l to indicate assignment to leaf l
  - Second, consider the following linear model.

$$Y = \sum_l L_l \alpha_l + W L_l \beta_l \tag{3}$$

  - The interaction coefficients in this regression recover the average treatment effects in each leaf, since

$$E[Y|W=1, L=1] - E[Y|W=0, L=1] = (\alpha_1 + \beta_1) - \alpha_1 = \beta_1 \tag{4}$$

  - Therefore, the standard error around the coefficients is also the standard error around the treatment effects.
  - We will also use these statistics to test hypothesis about leaf estimates.

Causal Trees

```
# Create a factor column 'leaf' indicating leaf assignment
num_leaves <- length(unique(tauhat_ct_est))  #There are as many leaves as there are predictions
df_est$leaf <- factor(tauhat_ct_est, labels = seq(num_leaves))
# Run the regression
ols_ct <- lm(as.formula("Y ~ 0 + leaf + W:leaf"), data= df_est)
ols_ct_summary <- summary(ols_ct)
```

Table 2: Average treatment effects per leaf

	Estimate	Std. Error
leaf1:W	-0.4611	0.0817
leaf2:W	-0.4284	0.0276
leaf3:W	-0.3896	0.0205
leaf4:W	-0.3844	0.0214
leaf5:W	-0.3758	0.0920
leaf6:W	-0.3315	0.0633
leaf7:W	-0.3167	0.0309
leaf8:W	-0.2124	0.0497

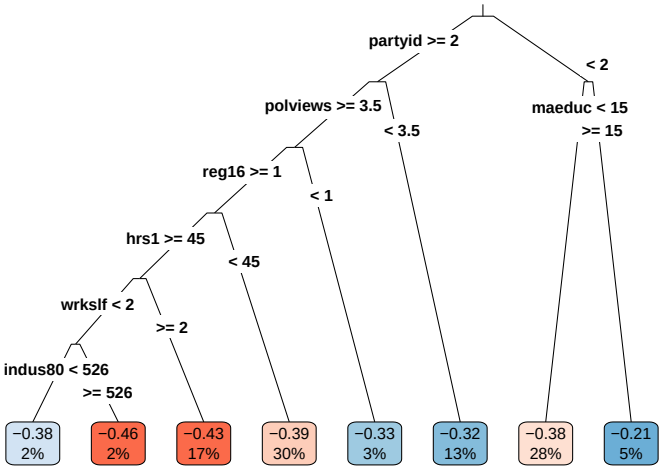
Causal Trees

- Step 6: Predict point estimates (on test set)

```
tauhat_ct_test <- predict(ct_pruned, newdata = df_test)


rpart.plot(
  x = ct_pruned,         # Pruned tree
  type = 3,              # Draw separate split labels for the left and right directions
  fallen = TRUE,         # Position the leaf nodes at the bottom of the graph
  leaf.round = 1,        # Rounding of the corners of the leaf node boxes
  extra = 100,           # Display the percentage of observations in the node
  branch = 0.1,           # Shape of the branch lines
  box.palette = "RdBu") # Palette for coloring the node
```

Causal Trees



Causal Trees

```
# Null hypothesis: all leaf values are the same
hypothesis <- paste0("leaf1:W = leaf", seq(2, num_leaves), ":W")
ftest <- linearHypothesis(ols_ct, hypothesis, test="F")

kable_styling(kable(data.frame(ftest, check.names = FALSE, row.names = NULL)[2,],
                    "latex", digits = 4,
                    caption="Testing null hypothesis: Average treatment effect is same across leaves"),
              bootstrap_options=c("striped", "hover", "condensed", "responsive"),
              full_width=FALSE)
```

Table 3: Testing null hypothesis: Average treatment effect is same across leaves

	Res.Df	RSS	Df	Sum of Sq	F	Pr($> F$)
2	5263	884.921	7	3.4114	2.8984	0.005

Causal Trees: Theory Details

Causal Trees: Theory Details

Causal Tree: Theory Details

- Work well in RCTs
- Issue: we do not observe the ground truth
- Honest estimation (Innovation):
  - One sample to choose partition
  - One sample to estimate leaf effects
- Why is the split critical?
- Fitting both on the training sample risks overfitting: Estimating many "heterogeneous effects" that are really just noise idiosyncratic to the sample.
- We want to search for true heterogeneity, not noise

Heterogeneous Treatment Effects Assumptions

- Before proceeding we need to make a couple of assumptions
- Assumption 1: Unconfoundedness

$$Y_i(1), Y_i(0) \perp W_i \mid X_i \tag{5}$$

  - The *unconfoundedness* assumption states that, once we condition on observable characteristics, the treatment assignment is independent to how each person would respond to the treatment.
  - i.e., the rule that determines whether or not a person is treated is determined completely by their observable characteristics.
  - This allows, for example, for experiments where people from different genders get treated with different probabilities,
  - **rules out** experiments where people self-select into treatment due to some characteristic that is not observed in our data.

Heterogeneous Treatment Effects

- Assumption 2: Overlap

$$\forall\, x \in \text{supp}\,(X), \qquad 0 < P\,(W = 1 \mid X = x) < 1 \tag{6}$$

  - The *overlap* assumption states that at every point of the covariate space we can always find treated and control individuals.
  - i.e., in order to estimate the treatment effect for a person with particular characteristics $X_i = x$, we need to ensure that we are able to observe treated and untreated people with those same characteristics so that we can compare their outcomes.

Trees

- A simple tree

$$MSE_0 = \frac{1}{N}\sum(Y_i - \bar{Y})^2 \qquad \text{All observations}$$

$$MSE_1 = \frac{1}{N}\sum(Y_i - \bar{Y}_{j:x_j \in l(x_i|\Pi)})^2 \qquad X_i < c_1 \quad X_i \geq c_2$$

- Partition $\Pi \in P$

$$\{l_1 = \{x_i : x_i < c_1\}, l_2 = \{x_i : x_i \geq c_2\}\} \tag{7}$$

- Prediction is

$$\hat{\mu}(x) = \bar{Y}_{j:x_j \in l(x_i|\Pi)} \tag{8}$$

The Honest Target: Athey and Imbens Innovation

- Given a partition Π define

$$MSE_{\mu}(S^{te}, S^{est}, \Pi) = \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ \left(Y_i - \hat{\mu}(X_i, S^{est}, \Pi) \right)^2 - Y_i^2 \right\} \tag{9}$$

- The expected MSE is the expectation of $MSE_{\mu}(S^{te}, S^{est}, \Pi)$ over estimation and test samples (independent)

$$EMSE_{\mu}(\Pi) = E_{S^{te}, S^{est}} \left[MSE_{\mu}(S^{te}, S^{est}, \Pi) \right] \tag{10}$$

The Honest Target: Athey and Imbens Innovation

- The ultimate goal is to construct and assess an algorithm $\pi(.)$ that maximizes the honest criterion

$$max\, Q^H(\pi) = -E_{S^{te},S^{est},S^{tr}}\left[MSE_\mu(S^{te}, S^{est}, S^{tr}, \pi(S^{tr})\right] \tag{11}$$

- In CART the target is different (adaptive target)

$$max\, Q^C(\pi) = -E_{S^{te},S^{tr}}\left[MSE_\mu(S^{te}, S^{tr}, \pi(S^{tr})\right] \tag{12}$$

The Honest Criterion

$$max\, Q^H(\pi) = -E_{S^{te},S^{est},S^{tr}}\left[MSE_\mu(S^{te}, S^{est}, S^{tr}, \pi(S^{tr})\right] \tag{13}$$

The Honest Criterion

- Understanding $EMSE_\mu(\Pi)$:

$$-EMSE_\mu(\Pi) = -E_{S^{te},S^{est}}\left[\left(Y_i - \hat{\mu}(X_i, S^{est}, \Pi)\right)^2 - Y_i^2\right] \tag{14}$$

$$= -E_{S^{te},S^{est}}\left[\left(Y_i - \mu(X_i,\Pi) + \mu(X_i,\Pi) - \hat{\mu}(X_i, S^{est}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -E_{S^{te},S^{est}}\left[\left(Y_i - \mu(X_i,\Pi)\right)^2 - Y_i^2\right]$$

$$-E_{S^{te},S^{est}}\left[\left(\mu(X_i,\Pi) - \hat{\mu}(X_i, S^{est}, \Pi)\right)^2\right]$$

$$-E_{S^{te},S^{est}}\left[2\left(Y_i - \mu(X_i,\Pi)\right)\left(\mu(X_i,\Pi) - \hat{\mu}(X_i, S^{est}, \Pi)\right)^2\right]$$

Heterogeneous Treatment Effects

$$= -E_{(Y_i,X_i),S^{est}}\left[(Y_i - \mu(X_i,\Pi))^2 - Y_i^2\right] - E_{X_i,S^{est}}\left[\left(\mu(X_i,\Pi) - \hat{\mu}(X_i,S^{est},\Pi)\right)^2\right]$$

$$= -E_{(Y_i,X_i),S^{est}}\left[Y_i^2 - 2Y_i\mu(X_i,\Pi) + \mu^2(X_i,\Pi) - Y_i^2\right] - E_{X_i,S^{est}}\left[\left(\mu(X_i,\Pi) - \hat{\mu}(X_i,S^{est},\Pi)\right)^2\right]$$

$$= -E_{(Y_i,X_i),S^{est}}\left[-2Y_i\mu(X_i,\Pi) + \mu^2(X_i,\Pi)\right] - E_{X_i,S^{est}}\left[\left(\mu(X_i,\Pi) - \hat{\mu}(X_i,S^{est},\Pi)\right)^2\right]$$

Note $E_{(Y_i,X_i),S^{est}}(Y_i) = E_{X_i,S^{est}}\mu(X_i,\Pi)$

$$= -E_{(Y_i,X_i),S^{est}}\left[\mu^2(X_i,\Pi)\right] - E_{X_i,S^{est}}\left[V(\hat{\mu}(X_i,S^{est},\Pi))\right]$$

The Honest Criterion

- How to estimate this quantities?
- First $E_{X_i,S^{est}}\left[V(\hat{\mu}(X_i,S^{est},\Pi))\right]$

$$V(\hat{\mu}(X_i,S^{est},\Pi)) = \frac{S^2_{S^{tr}}\left(l\left(x|\Pi\right)\right)}{N^{est}\left(l\left(x|\Pi\right)\right)}$$

$$\hat{E}_{X_i,S^{est}}\left[V(\hat{\mu}(X_i,S^{est},\Pi))|i\in S^{te}\right] = \sum_l p_l \frac{S^2_{S^{tr}}\left(l\right)}{N^{est}\left(l\right)}$$

$$= \sum_l \frac{1}{\#(l)} \frac{S^2_{S^{tr}}\left(l\right)}{N^{est}\left(l\right)}$$

$$= \frac{1}{N^{est}} \sum_{l\in\Pi} S^2_{S^{tr}}\left(l\right)$$

The Honest Criterion

- Next $E_{(Y_i,X_i),S^{est}}\left[\mu^2(X_i,\Pi)\right]$
- Note $V(\hat{\mu}|x,\Pi) = E(\hat{\mu}^2|x,\Pi) - [E(\hat{\mu}|x,\Pi)]^2$

$$\frac{S^2_{S^{tr}}\left(l\left(x|\Pi\right)\right)}{N^{tr}\left(l\left(x|\Pi\right)\right)} \approx \hat{\mu}^2(x|S^{tr},\Pi) - \mu^2(x|\Pi)$$

$$\mu^2(x|\Pi) \approx \hat{\mu}^2(x|S^{tr},\Pi) - \frac{S^2_{S^{tr}}\left(l\left(x|\Pi\right)\right)}{N^{tr}\left(l\left(x|\Pi\right)\right)}$$

$$\hat{E}_{X_i}\left(\mu^2(X_i|\Pi) \approx \frac{1}{N^{tr}}\sum_{i\in S^{tr}}\hat{\mu}^2(x|S^{tr},\Pi) - \sum_l \frac{1}{\#l}\frac{S^2_{S^{tr}}(l)}{N^{tr}/\#l}\right.$$

The Honest Criterion

- Finally

$$-EMSE_{\mu}(\Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(x|S^{tr}, \Pi) - \sum_{l} \frac{1}{N^{tr}} S^2_{S^{tr}}(l) - \frac{1}{N^{est}} \sum_{l \in \Pi} S^2_{S^{tr}}(l)$$

$$= \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(x|S^{tr}, \Pi) - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{l \in \Pi} S^2_{S^{tr}}(l)$$

Honest Inference for Treatment Effects

- Given a tree Π, define for all x and both treatment levels w the population average outcome

$$\mu(w, x|\Pi) = E\left[Y_i(w)|X_i \in l(x|\Pi)\right]$$

- The Average Treatment Effect

$$\tau(x|\Pi) = E\left[Y_i(1) - Y_i(0)|X_i \in l(x|\Pi)\right]$$

$$= \mu(1, x|\Pi) - \mu(0, x|\Pi)$$

Honest Inference for Treatment Effects

- The estimated counterparts are

$$\hat{\mu}(w, x|S, \Pi) = \frac{1}{\#\left(\{i \in S_w : X_i \in l(x|\Pi)\}\right)} \sum_{i \in S_w : X_i \in l(x|\Pi)} Y_i^{obs} \tag{15}$$

$$\hat{\tau}(X, S, \Pi) = \hat{\mu}(1, x|S, \Pi) - \hat{\mu}(0, x|S, \Pi) \tag{16}$$

- Define the MSE for treatment effects as

$$MSE_\tau(S^{te}, S^{est}, \Pi) = \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ \left(\tau_i - \hat{\tau}(X_i|S^{est}, \Pi\right)^2 - \tau_i^2 \right\}$$

Honest Inference for Treatment Effects

Adapt $EMSE_\mu$ to estimate $EMSE_\tau$

$$-EMSE_\mu(\hat{S^{tr}}, S^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i|S^{tr}, \Pi) - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi} S^2_{S^{tr}}(l)$$

for HTE

$$-EMSE_\tau(\hat{S^{tr}}, S^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(X_i|S^{tr}, \Pi) - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi} \left(\frac{S^2_{S^{tr}_{treat}}(l)}{p} + \frac{S^2_{S^{tr}_{control}}(l)}{(1-p)}\right)$$

Observational Studies with Unconfoundedness

- Athey and Imbens (2016):

  "The proposed methods can be adapted to observational studies under the assumption of unconfoundedness. In that case we need to modify the estimates within leaves to remove the bias from simple comparisons of treated and control units. There is a large literature on methods for doing so, ..., for example, we can do so by propensity score weighting. Efficiency will improve if we renormalize the weights within each leaf and within the treatment and control group when estimating treatment effects"

Review & Next Steps

- Problem: we never observe t_i unlike prediction that we observe Y_i
- Causal Trees search for leaves with
  - HTE across leaves
  - precisely-estimated leaf effects
- Key is the honest Criterion
- Work well with RCTs
- With selection on observables, recommendation is propensity forests?
- Next class: Forests
- Questions? Questions about software?

Further Readings

- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public opinion quarterly, 76(3), 491-511.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.