Lecture 24: Causal Trees (Cont.)

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

November 5, 2020

Agenda

Treatment Effects

- 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

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

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

y eso sí puedo

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

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

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

↳ Alta o Baja Educacion

Heterogeneous Treatment Effects

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

Heterogeneous Treatment Effects

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

$$Y_i(1), Y_i(0) \perp W_i \mid X_i \tag{3}$$

  - 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

$$\forall\, x \in \operatorname{supp}(X), \qquad 0 < P\left(W = 1 \mid X = x\right) < 1 \tag{4}$$

  - 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.

Causal Tree: Theory

- Work well in RCTs
- Issue: we do not observe the ground truth → τ_i
- Honest estimation (Innovation):
  - One sample to choose partition → construir el arbol
  - 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

Trees

- A simple tree

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

$$MSE_1 = \frac{1}{N}\sum(Y_i - \bar{Y}_{j:x_j \in l(x_i|\Pi)})^2$$

All observations

$X_i < c_1$ $\quad$ $X_i \geq c_2$

ℓ_1 $\quad$ ℓ_2

μ_1 μ_2 ℓ_1 ℓ_2 c_1

- Partition $\Pi \in P$

$$\{l_1 = \{x_i : x_i < c_1\}, l_2 = \{x_i : x_i \geq c_2\}\} \tag{5}$$

- Prediction is

$$\hat{\mu}(x) = \bar{Y}_{j:x_j \in l(x_i|\Pi)} \tag{6}$$

The Honest Target: Athey and Imbens Innovation

MSE

est

te

Adjustment algebraic

- Given a partition Π define

$$MSE_\mu(S^{te}, S^{est}, \Pi) = \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ \left(Y_i - \hat{\mu}(X_i, S^{est}, \Pi) \right)^2 - Y_i^2 \right\} \tag{7}$$

$= y - \hat{f}$

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

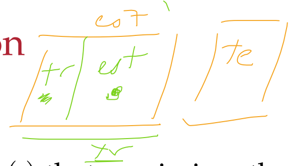
$$EMSE_\mu(\Pi) = E_{S^{te}, S^{est}} \left[MSE_\mu(S^{te}, S^{est}, \Pi) \right] \tag{8}$$

est

te

independent

The Honest Target: Athey and Imbens Innovation

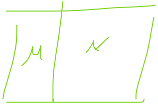


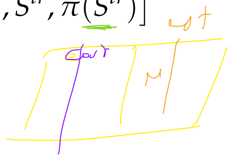
- The ultimate goal is to construct and assess an algorithm $\pi(.)$ that maximizes the honest criterion

$$max\, Q^H(\pi) = -E_{S^{te},S^{est},S^{tr}}\left[MSE_\mu(S^{te},S^{est},S^{tr},\pi(S^{tr})\right] \tag{9}$$

- In CART the target is different (adaptive target)

$$max\, Q^C(\pi) = -E_{S^{te},S^{tr}}\left[MSE_\mu(S^{te},S^{tr},\pi(S^{tr})\right] \tag{10}$$





The Honest Criterion

Random variable

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

from est

control group

estimate group

The Honest Criterion

te est

$y - \hat{f}$

for indep

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

$$Q = -EMSE_\mu(\Pi) = -E_{S^{te},S^{est}}\left[\left(Y_i - \hat{\mu}(X_i, S^{est}, \Pi)\right)^2 - Y_i^2\right] \tag{12}$$

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

add 0

$(a-b)^2 = a^2 - 2ab + b^2$

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

σ^2

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

B^2

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

A B

$cov(A,B) = 0$

Heterogeneous Treatment Effects

$y - f$

$Y_i, X_i \in S$

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

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

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

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

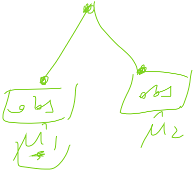
$$= -E_{(Y_i,X_i),S^{est}}\left[\mu^2(X_i,\Pi)\right] - E_{X_i,S^{est}}\left[V(\hat{\mu}(X_i,S^{est},\Pi))\right]$$

Bias + Var

The Honest Criterion

$y - \hat{y}$

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



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

$V(\bar{x}) = \frac{\sigma^2}{n}$

$\hat{V}(\bar{x}) = \frac{S^2}{n}$

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

assume
$p_l = p \ \forall l$

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

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

The Honest Criterion

$E(\hat{\mu}|x,\Pi) = \mu$

$(E(\hat{\mu}|x,\Pi))^2 = \mu^2$

$\hat{\mu}^2$

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

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

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

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

The Honest Criterion

tr | est

- Finally

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

$$= \underbrace{\frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(x|S^{tr}, \Pi)}_{\text{convencional de CART}} - \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi} S^2_{S^{tr}}(l)}_{\text{penaliza incertidumbre adentro de la hoja}}$$

si no usamos honest estimation

$MSE(S^{tr}, S^{tr}, \Pi)$

HW

so

honesty

$\widehat{MSE}(S^{tr}, S^{est}, \Pi)$

Honest Inference for Treatment Effects

$y_i - \delta$

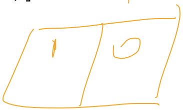
- Given a tree Π, define for all x and both treatment levels w the population average outcome

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

- The Average Treatment Effect

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

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



$\tau(x | x_i \in l) = \mu_1 - \mu_0$

Honest Inference for Treatment Effects

- The estimated counterparts are

$$\hat{\mu}(w, x|S, \Pi) = \frac{1}{\#\left(\{i \in S_w : X_i \in l(x|\Pi)\}\right)} \sum_{i \in S_w : X_i \in l(x|\Pi)} Y_i^{obs} \tag{13}$$

$$\hat{\tau}(X, S, \Pi) = \hat{\mu}(1, x|S, \Pi) - \hat{\mu}(0, x|S, \Pi) \tag{14}$$

- Define the MSE for treatment effects as

$y_i - \hat{f} - y_i^2$

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

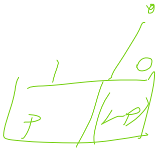
no es observado τ

no vemos la ground truth

Honest Inference for Treatment Effects

$-y^2$

$-\tau^2$



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

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

for HTE

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

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"

Causal Forests

- Trees can be ~~noise~~ noisy. We can use forests   Tr | Est
  - Draw a sample bootstrap of size s
  - Split the sample into *Tr* and *Est*
  - Use *Tr* to grow the tree
  - Use *Est* to estimate the leaf-specific effects
- Advantages
  - Consistent for $\tau(x)$
  - Asymptotically Normal
  - “Auto” search for HTE
- Disadvantage
  - Sample splitting (noisier estimates)

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: Causal forests demo
- 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.
- Lundberg, I (2017). Causal forests. A tutorial in high dimensional causal inference. Mimeo
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.