Machine Learning & Causal Inference

Heterogeneous Treatment Effects

Susan Athey

Stanford GSB

ML For Policy

Machine learning helps us build more granular statistical models.

- Improved causal inference about the effect of a policy, program or intervention
  - More granual counterfactual predictions about what would have happened to individuals in the absence of the treatment
  - Better use of observables to control for confounders
- Better understanding of how and why a policy works, for continued improvement
- Personalized evaluation of how a policy works
- Personalized policy assignment rules

Reframing ML for Policy

- ML methods perform well in practice, but many do not have well-established statistical properties
- Unlike prediction, ground truth for causal parameters are not directly observed
- Need valid confidence intervals for many applications (A/B testing, drug trials); challenges include adaptive model selection and multiple testing

Machine Learning and Econometrics for Causal Inference

ML Themes (See Athey, "The Impact of ML on Economics"; Athey & Imbens, "ML Methods Economists Should Know About" for surveys)

- Data-driven model selection with very flexible functional forms
- Regularization: penalization, model averaging, subsampling
- **Goal:** goodness of fit in held-out test set, same distribution
- Similar goals as semi-parametric estimation, with better practical performance. Theory?

Contributions to Causal Inference

- ATE: Control for confounders (Post-Double Lasso (Belloni, Chernozhukov, and Hansen); Residual Balancing (Athey, Imbens and Wager); AIPW/Double ML (Chernozhukov et al.))
- IV: Select from many instruments (Chernozhukov et al)
- Panel data: Matrix factorization as alternative to DID, synthetic controls (Athey et al) and as part of structural models of consumer behavior (Athey, Blei, Ruiz; Athey et al)
- **CATE: Heterogeneous treatment effects, optimal policies, adaptive experiments**

Heterogeneous Treatment Effects

1. Estimating Impact of a Treatment

Applications:

- A/B Tests; Randomized Experiment
- Observational study of the impact of a system change or marketing change
- Effect of a drug

Goals:

- Systematically identify sub-populations and estimate treatment effects, valid inference
- Understand mechanisms
- Gain insight for developing further improvements to the treatment

2. Optimal Policies

Heterogeneous Treatment Effects

1. Estimating Imact of a Treatment

2. Optimal Policies

Customized to Subpopulations:

- $E[\tau \mid X_i \in S]$
- Ship a new algorithm only when set of conditions are met
- Decision/triage guidelines for doctors, police, judges, salespeople, etc.

Customized to each individual's observables:

- $E[\tau \mid X_i = x]$
- Online assignment of users to different experiences, ads, or ranking algorithms according to which is most effective
- Computer-assisted medicine, judicial decisions, finance decisions, etc.

Section 1

Heterogeneous Treatment Effects

Treatment Effect Heterogeneity

Literature:

- Identifying subgroups (Athey and Imbens, 2016) or other low-dimensional parameter estimates
- Testing for heterogeneity across all covariates (List, Shaikh, and Xu, 2016)
- Robustness to model specification (Athey and Imbens, 2015)
- Imai and Ratkovic (2013) analyze treatment effect heterogeneity with LASSO
- Conditional average treatment effect with theoretical guarantees (Wager and Athey, 2018; Athey, Tibshirani, and Wager, 2019)
- Identifying individuals with highest estimated treatment effects (Chernozhukov et al, 2018)
- Estimating optimal policies (Athey and Wager, 2016; Zhou, Athey and Wager 2018)
- Contextual bandits for policy learning and data collection (Langford et al (2016); Dimakopoulou, Zhou, Athey and Imbens (2018))
- Targeted ML (van der Laan, 2006) can be used as a semi-parametric approach to estimating treatment effect heterogeneity

ML Methods for Causal Inference: Treatment Effect Heterogeneity

- ML methods perform well in practice, but many do not have well established statistical properties
- Unlike prediction, ground truth for causal parameters not directly observed
- Need valid confidence intervals for many applications (AB testing, drug trials); challenges include adaptive model selection and multiple testing

Some themes of ML/CI research agenda:

- Either decompose problem into prediction and causal components; or build novel methods inspired by ML
- Sample splitting/cross-fitting to avoid spurious findings and to get consistency/asymptotic normality
- Build on insights from semi-parametric theory
- Use orthogonal moments to build in greater tolerance for slow convergence in estimation of nuisance parameters

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of:

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}^p$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Holland, 1986, Imbens and Rubin, 2015, Rosenbaum and Rubin, 1983, Rubin, 1974), we posit the existence of quantities $Y_i^{(0)}$ and $Y_i^{(1)}$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment ($W_i = 1$) or no treatment ($W_i = 0$).

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of:

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}^p$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Goal is to estimate the **conditional average treatment effect**

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]$$

NB: In experiments, we only get to see $Y_i = Y_i^{(W_i)}$.

The potential outcomes framework

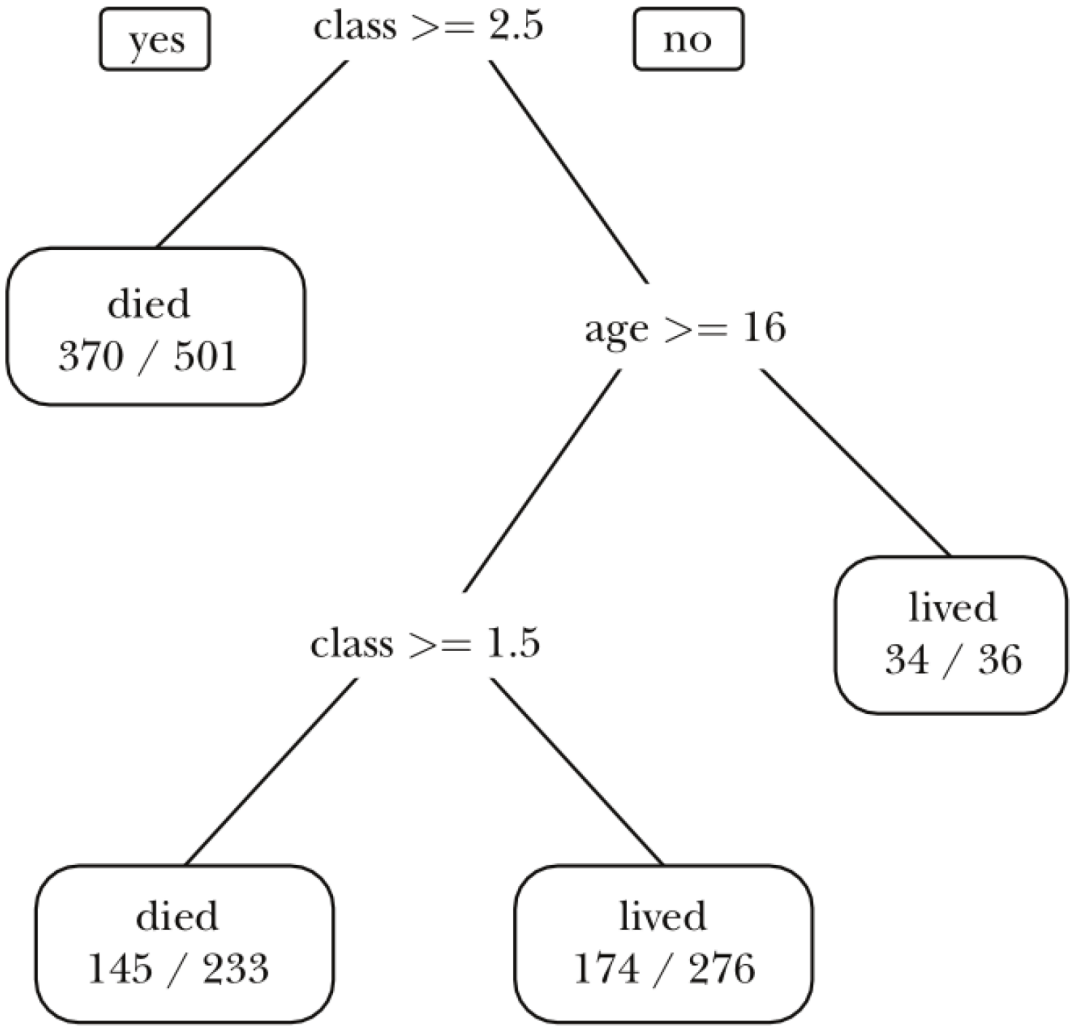
If we make no further assumptions, estimating $\tau(x)$ is not possible.

- Literature often assumes {**unconfoundedness**} **(Rosenbaum and Rubin, 1983)**

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

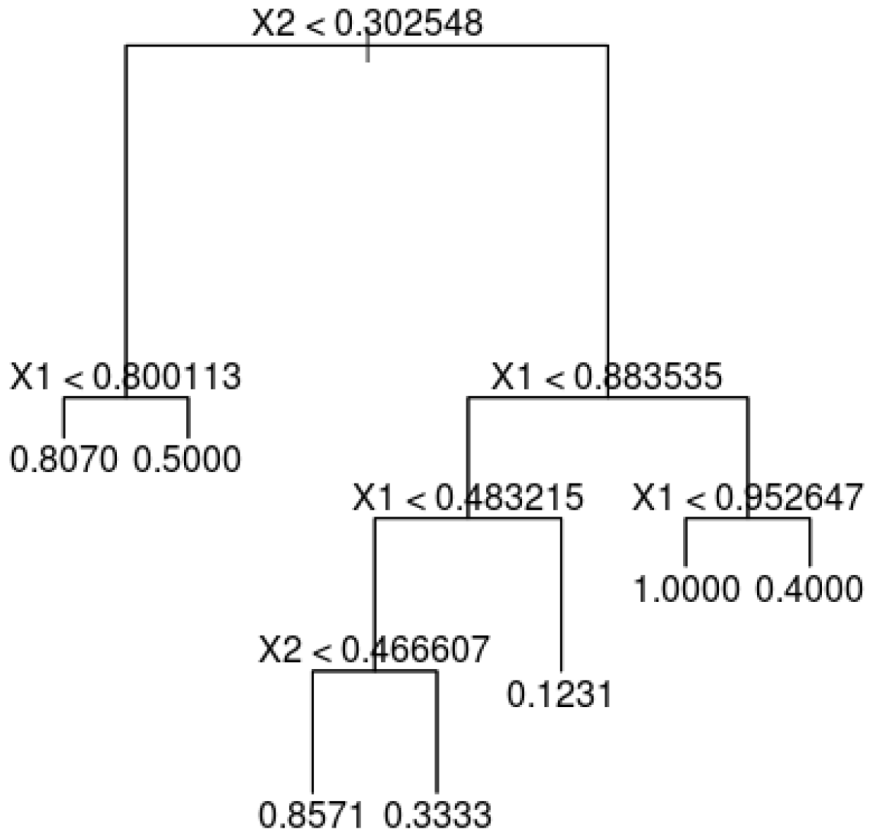
- When this assumption holds, methods based on matching or propensity score estimation are usually consistent.

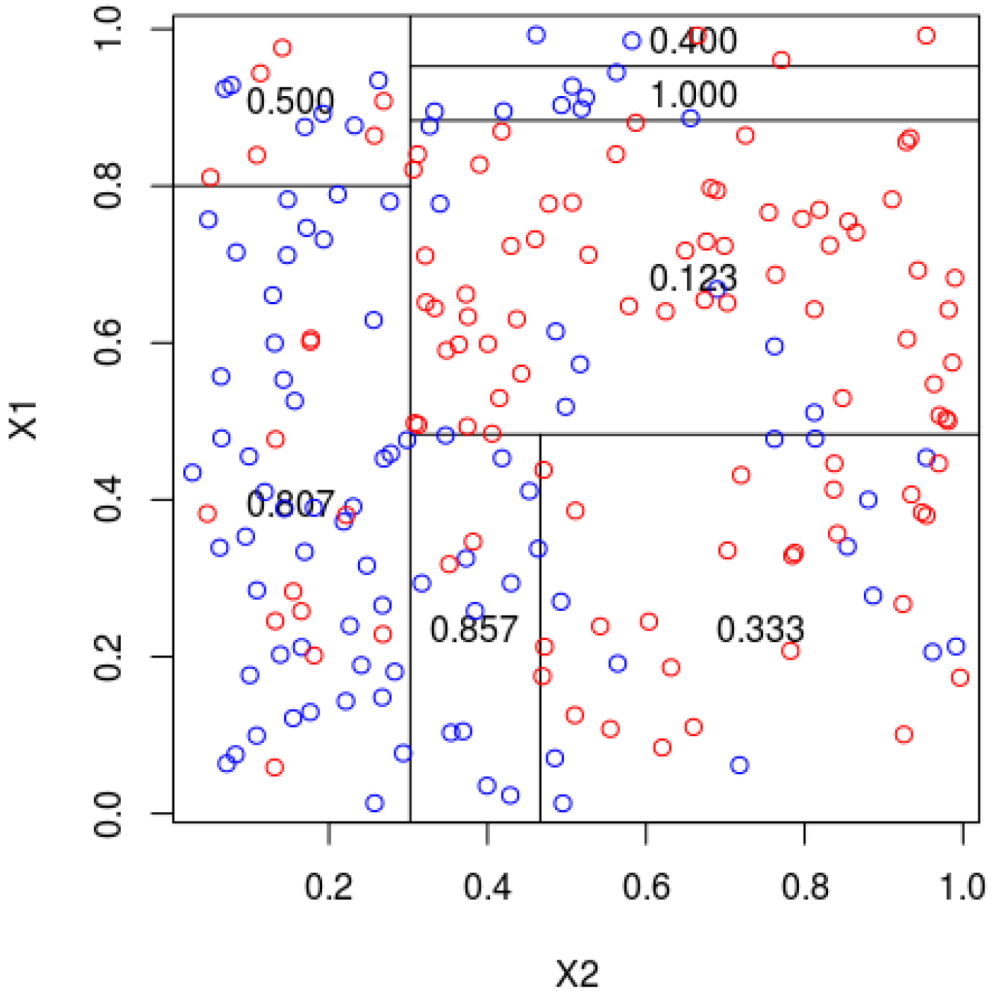
Regression Trees Refresher: Titanic Example



Source: Varian (2014)

Regression Trees: The Tree as a Partition





Causal Trees

Divide population into subgroups to minimize MSE in treatment effects

- Goal: report heterogeneity without pre-analysis plan but with valid confidence intervals
- Moving the goalposts: method defines estimand (treatment effects for subgroups) and generates estimates
- Solve over-fitting problem with sample splitting: choose subgroups in half the sample and estimate on other half

Challenges

- Objective function is infeasible: $\sum_i \left[(\tau_i - \hat{\tau}(X_i))^2\right]$
- Need to estimate objective to optimize for it rather than take a simple average of squared error $\sum_i \left[(Y_i - \hat{\mu}(X_i))^2\right]$
- Estimand is unstable

Approaches to CATE: Taxonomy from Athey and Imbens, 2016

- Build a single model ("single trees"): include treatment indicator as a covariate
  - May not split on W, estimate of treatment effect identically zero for some parts of tree. Later named "S-Learner."
- Build two separate models for treatment and control ("two trees"). Later named "T-Learner."
  - For a given value x, the neighborhood defined for treatment group might be totally different than control group, e.g. treated tree splits on gender and control tree splits on age. Then male, 60 yrs old, is in the "male" leaf in treatment group tree, and in the "over 50" leaf in the control group tree. Estimate of treatment effect heterogeneity compares treated males to older people in control group.
- "Transformed outcome" - transform the outcome and build a single prediction model. Later generalized (e.g. Chernozukov et al) to AIPW score as the outcome.
- Estimate the MSE criterion (preferred in Athey and Imbens, 2016)

Approaches to CATE: Transformed Outcome from Athey and Imbens, 2016

If we want to apply off-the-shelf prediction methods to estimate CATE, one approach is to transform the outcome. In particular, if p is the assignment probability:

Let $Y_i^* = Y_i/p$ if $W_i = 1$, and let $Y_i^* = -Y_i/(1-p)$ if $W_i = 0$.

It is straightforward to check that $\mathbb{E}[Y_i^*|X_i = x] = \tau(x)$. It is a noisy, but unbiased estimate of CATE.

Thus, if we first tranform the outcome, and then apply a prediction method to the tranformed outcome, we will get estimates of CATE.

This can be generalized in observational studies to weighting by an estimated propensity score, or use the AIPW score as an outcome.

Notation for Partitions and Leaf Effect Estimates

Three samples: model selection/tree construction: $\mathcal{S}^{tr}$, estimation sample for leaf effects $\mathcal{S}^{est}$, and a (hypothetical) test sample $\mathcal{S}^{te}$.

Given a partition Π,

$\hat{\tau}(X_i; \mathcal{S}^{est}, \Pi)$ is the sample average treatment effect in sample $\mathcal{S}^{est}$ for the leaf $\ell(X_i; \Pi)$ associated with covariates X_i:

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

Estimating the Mean-squared Error (MSE) Criterion

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

$$\begin{aligned} MSE(\mathcal{S}^{est}, \mathcal{S}^{te}) &= \frac{1}{n^{te}} \sum_{i \in \mathcal{S}^{te}} (\tau_i - \hat{\tau}(X_i; \mathcal{S}^{est}, \Pi))^2 \\ &= \frac{1}{n^{te}} \sum_{i \in \mathcal{S}^{te}} \left(\tau_i^2 - 2 \cdot \tau_i \cdot \hat{\tau}(X_i; \mathcal{S}^{est}, \Pi) + \hat{\tau}^2(X_i; \mathcal{S}^{est}, \Pi) \right) \end{aligned}$$

$$\begin{aligned} EMSE &= E_{\mathcal{S}^{est}, \mathcal{S}^{te}} \left[MSE(\mathcal{S}^{est}, \mathcal{S}^{te}) \right] \\ &= \mathbb{V}_{\mathcal{S}^{est}, X_i} \left[\hat{\tau}(X_i; \Pi, \mathcal{S}^{est}) \right] - E_{X_i} \left[\tau^2(X_i; \Pi) \right] + E[\tau_i^2] \end{aligned}$$

The last equality makes use of fact that estimates are unbiased in independent test sample. Can construct empirical estimates of each of these quantities except for the last which does not depend on Π and thus does not affect partition selection.

Causal Tree Algorithm

- Divide data into tree-building $\mathcal{S}^{tr}$ and estimation $\mathcal{S}^{est}$ samples
- Use a greedy algorithm to recursively partition covariate space $\mathcal{X}$ into a deep partition Π
  - At each node the split is selected as the one that minimizes our estimate of EMSE over all possible binary splits
  - Preserve minimum number of treated and control units in each child leaf
- Use cross-validation to select the depth d^* of the partition that minimizes an estimate of MSE of treatment effects, using left-out folds as proxies for the test set
- Select partition Π^* by pruning Π to depth d^*, pruning leaves that provide the smallest improvement in goodness of fit
- Estimate the treatment effects in each leaf of Π^* using the estimation sample $\mathcal{S}$