Causal Trees

Ciencia de Datos y Econometría Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 22, 2025

Causal inference summary

- Target :

$$ATE = E\left[Y_i(1) - Y_i(0)\right] = E\left[\tau_i\right]$$

- Identifying assumption:

$$\left(Y_i(0), Y_i(1)\right) \perp\!\!\!\perp D_i | X_i$$

$$0 < Pr(D = 1|X) < 1$$

What's a "good" prediction?

- Want our prediction to be "close," i.e. minimize the expected **mean squared error**:

$$\min_{f(x)} E\left[(y - f(x))^2 \middle| X = x \right]$$

Combining causal effects and ML: predicting heterogeneous treatment effects

- What is the effect of adds on client expenditure
  - for men vs. women?
  - for young vs. old?
  - etc....
- Why does it matter?

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

To explore heterogeneity by sex:

$$\begin{aligned} Y_i &= \tau^{female} D_i + \varepsilon_i, \qquad i : Female_i = 1 \\ Y_i &= \tau^{male} D_i + \varepsilon_i, \qquad i : Female_i = 0, \end{aligned}$$

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

To explore heterogeneity by sex:

$$\begin{aligned} Y_i &= \tau^{female} D_i + \varepsilon_i, \qquad i : Female_i = 1 \\ Y_i &= \tau^{male} D_i + \varepsilon_i, \qquad i : Female_i = 0, \end{aligned}$$

or, equivalently:

$$\begin{aligned} Y_i &= \tau^{male} D_i + \beta Female_i + \gamma D_i \times Female_i + \varepsilon_i \\ \tau^{female} &= \tau^{male} + \gamma. \end{aligned}$$

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

To explore heterogeneity by sex:

$$\begin{aligned} Y_i &= \tau^{female} D_i + \varepsilon_i, \qquad i : Female_i = 1 \\ Y_i &= \tau^{male} D_i + \varepsilon_i, \qquad i : Female_i = 0, \end{aligned}$$

or, equivalently:

$$\begin{aligned} Y_i &= \tau^{male} D_i + \beta Female_i + \gamma D_i \times Female_i + \varepsilon_i \\ \tau^{female} &= \tau^{male} + \gamma. \end{aligned}$$

More generally,

$$\begin{aligned} Y_i &= \tau D_i + X_i' \beta + D_i X_i' \gamma + \varepsilon_i, \\ \tau(x) &= \tau + x' \gamma \end{aligned}$$

Challenges with traditional heterogeneity analysis

$$Y_i = \tau D_i + X_i'\beta + D_i X_i'\gamma + \varepsilon_i$$

- Functional form: treatment effects may not vary linearly with X_i
- Curse of dimensionality: when X_i includes many variables, OLS impractical or infeasible
- False discovery rate.

Predicting outcomes vs. treatment effects

Predicting outcomes	Predicting treatment effects
Target: $\hat{y}(x) = E[Y_i \| X_i = x]$	Target: $\tau(x) = E[\tau_i \| X_i = x]$
Criterion:	Criterion:
$\min E\left[\left(\hat{y}(x) - Y_i\right)^2 \| X_i = x\right]$	$\min E\left[\left(\tau(x) - \tau_i\right)^2 \| X_i = x\right]$
Training data: $\{Y_i, X_i\}_{i=1}^{n}$	Training data: $\{\tau_i, X_i\}_{i=1}^{n}$

Predicting outcomes vs. treatment effects

Predicting outcomes	Predicting treatment effects
Target: $\hat{y}(x) = E[Y_i \| X_i = x]$	Target: $\tau(x) = E[\tau_i \| X_i = x]$
Criterion:	Criterion:
$\min E\left[(\hat{y}(x) - Y_i)^2 \| X_i = x\right]$	$\min E\left[(\tau(x) - \tau_i)^2 \| X_i = x\right]$
Training data: $\{Y_i, X_i\}_{i=1}^{n}$	Training data: $\{\tau_i, X_i\}_{i=1}^{n}$

Why is training data a problem for predicting treatment effecs?

Predicting outcomes vs. treatment effects

Predicting outcomes	Predicting treatment effects
Target: $\hat{y}(x) = E[Y_i \| X_i = x]$	Target: $\tau(x) = E[\tau_i \| X_i = x]$
Criterion:	Criterion:
$\min E\left[\left(\hat{y}(x) - Y_i\right)^2 \| X_i = x\right]$	$\min E\left[\left(\tau(x) - \tau_i\right)^2 \| X_i = x\right]$
Training data: $\{Y_i, X_i\}_{i=1}^n$	Training data: $\{\tau_i, X_i\}_{i=1}^n$

Why is training data a problem for predicting treatment effecs?

- Consequence: can't apply ML directly to predicting treatment effects; have to adapt them

Adapting ML to predict treatment effects

- Break it up:

$$\begin{aligned} E\left[\tau_i | X_i\right] &:= E\left[Y_i\left(1\right) - Y_i\left(0\right)|X_i\right] \\ &= E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right] \end{aligned}$$

(by what assumption?)

Adapting ML to predict treatment effects

- Break it up:

$$\begin{aligned} E\left[\tau_i|X_i\right] &:= E\left[Y_i\left(1\right)-Y_i\left(0\right)|X_i\right] \\ &= E\left[Y_i|X_i,D_i=1\right]-E\left[Y_i|X_i,D_i=0\right] \end{aligned}$$

  (by what assumption?)
- Adjust the criterion: (why?)

$$\min\sum_{i=1}^{n}\left(\tau\left(X_i\right)-\tau_i\right)^2 \iff \max\sum_{i=1}^{n}\tau\left(X_i\right)^2$$

Adapting ML to predict treatment effects

- Break it up:

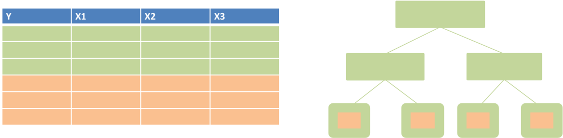
$$
\begin{aligned}
E\left[\tau_i | X_i\right] &:= E\left[Y_i\left(1\right) - Y_i\left(0\right) | X_i\right] \\
&= E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right]
\end{aligned}
$$

  (by what assumption?)

- Adjust the criterion: (why?)

$$
\min \sum_{i=1}^{n} \left(\tau\left(X_i\right) - \tau_i\right)^2 \iff \max \sum_{i=1}^{n} \tau\left(X_i\right)^2
$$

- Be honest: use one set of observations to select the tree structure, and another to generate predictions



Predicting treatment effects using ML: Summary

- Target:

$$CATE := \tau(x) = E[\tau_i | X_i = x]$$

- Key identifying assumption:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | X_i$$

$$0 < Pr(D = 1|X) < 1$$

- Estimation: Random Causal Forest
  - Grow decision trees on many bootstrapped samples
  - Choose splits using the training set to max $\sum_{i=1}^{n} \tau(X_i)^2$
  - Generate predictions in each leaf using the estimation set
  - Average predictions over the trees in the forest
- Go to R!