Introduction to Causal Machine Learning

Causal Forest

Anthony Strittmatter

References

- Huber (2023): "Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R", Chapter 5.4, online version.
- Athey and Imbens (2016): "Recursive Partitioning for Heterogeneous Causal Effects", Proceedings in the National Academy of Sciences, 113 (27), pp. 7353-7360, download.
- Wager and Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests", Journal of the American Statistical Association, 113 (523), pp. 1228-1242, download.

Why are Heterogeneous Effects Interesting?

- For most treatments, there is no specific reasons to assume that the effects are homogeneous
- If the effects are heterogeneous, there might be some individuals who benefit from a treatment and others may experience disadvantages from the treatment
- Investigating heterogeneous treatment effects enables us to understand how a treatment works
- Most treatments have a target group (e.g. disadvantaged youths are the target group of the Job Corps program). If there are two treatments with the same average effect, they might be evaluated differently, when one treatment mainly affects the target group while the other treatment affects non-targets

$\Rightarrow$ **Can we use ML to identify the relevant dimensions of effect heterogeneity in a data-driven way?**

Relevant Parameters

- Individual Causal Effects:

$$\delta_i = Y_i^1 - Y_i^0$$

- Conditional Average Treatment Effects (CATEs):

$$\delta(x) = E[Y_i^1 - Y_i^0 | X_i = x] = E[\delta_i | X_i = x]$$

Why Can't we use Off-the-Shelf Trees?

Out-of-Sample Mean-Squared-Error (MSE):

- Individual Causal Effect δ_i:

$$MSE_{\hat{\delta}} = E\left[(\hat{\delta}_i - \delta_i)^2\right] = \underbrace{E\left[(\hat{\delta}_i - E[\hat{\delta}_i])^2\right]}_{\text{Variance}} + \underbrace{E[\hat{\delta}_i - \delta_i]^2}_{\text{Squared Bias}}$$

- CATEs $\delta(x)$:

$$\begin{aligned} MSE_{\hat{\delta}(x)} &= E\left[(\hat{\delta}(x) - \delta(x))^2\right] \\ &= \underbrace{E\left[(\hat{\delta}(x) - E[\hat{\delta}(x)])^2\right]}_{\text{Variance}} + \underbrace{E[\hat{\delta}(x) - \delta(x)]^2}_{\text{Squared Bias}} \end{aligned}$$

$\rightarrow$ δ_i and $\delta(x)$ are unobservable

Preliminaries

- Univariate OLS regression:

$$Y_i = \alpha + \delta D_i + u_i$$

with

$$\delta = \frac{Cov(Y, D)}{Var(D)}$$

- Variance estimator:

$$Var(D) = \frac{1}{N} \sum_{i=1}^{N} (D_i - \bar{D})^2$$

- Covariance estimator:

$$Cov(Y, D) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})(D_i - \bar{D})$$

Causal Forest

- Consider the random effects model $Y_i = D_i \delta(X_i) + \epsilon_i$ with

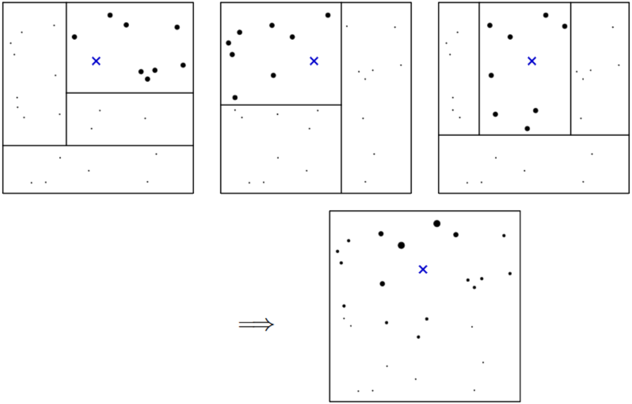
$$\delta(x) = Var(D_i | X_i = x)^{-1} Cov(Y_i, D_i | X_i = x)$$

- Estimator

$$\hat{\delta}(x) = \left(\sum_{i=1}^{N} \alpha_i(x)(D_i - \bar{D}_\alpha)^2 \right)^{-1} \sum_{i=1}^{N} \alpha_i(x)(D_i - \bar{D}_\alpha)(Y_i - \bar{Y}_\alpha)$$

with $\bar{D}_\alpha = \sum_{i=1}^{N} \alpha_i(x) D_i$ and $\bar{Y}_\alpha = \sum_{i=1}^{N} \alpha_i(x) Y_i$

Random Forest: Weighted Representation



Source: Athey, Tibshirani, Wager (2018)

Causal Forest

- The weights $\alpha_i(x)$ are obtained with the generalised causal forest algorithm
  - Weights for causal tree g:

$$\alpha_{ig}(x) = \frac{1\{X_i \in l_j(x, d, \pi_g)\}}{\sum_{i=1}^{N} 1\{X_i \in l_j(x, d, \pi_g)\}}$$

  - Causal forest weights:

$$\alpha_i(x) = \frac{1}{G} \sum_{i=1}^{G} \alpha_{ig}(x)$$

Reference: Athey, Tibshirani, and Wager (2018)

Algorithm

- Optimal optimisation criteria:

$$\max \Delta(C_1, C_2) = \frac{N_{C_1} N_{C_2}}{N_P} \left(\hat{\delta}_{C_1} - \hat{\delta}_{C_2}\right)^2$$

  $\rightarrow$ Requires the estimation of $\hat{\delta}_{C_1}$ and $\hat{\delta}_{C_2}$ at each candidate split

- Computational efficient optimisation algorithm
  (1) **Labelling step:** Calculate $\hat{\delta}_P$, $A_P = Var_P(D_i)$, and the pseudo-outcome

$$p_i = A_P^{-1}(D_i - \bar{D}_P)(Y_i - \bar{Y}_P - (D_i - \bar{D}_P)\hat{\delta}_P)$$

  (2) **Regression step:**

$$\max \tilde{\Delta}(C_1, C_2) = \frac{1}{N_{C_1}} \left(\sum_{i:X_i \in C_1} p_i\right)^2 + \frac{1}{N_{C_2}} \left(\sum_{i:X_i \in C_2} p_i\right)^2$$

  (3) Relabel child nodes to parent nodes and repeat (1) and (2) until stopping criteria of tree is reached
  (4) Calculate weights $\alpha_{ig}(x)$ and build forest by repeating (1)-(3) with different subsamples and covariates

Pseudo-Outcome

- Assume D_i is binary and randomly assigned and denote $p = \bar{D}_P = Pr(D_i = 1) = Pr(D_i = 1|C_j)$

$$
\begin{aligned}
E[p_i|C_j] =& E\left[\frac{D_i - p}{p(1-p)}(Y_i - \bar{Y}_P - (D_i - p)\hat{\delta}_P) \middle| C_j \right] \\
=& E\left[\frac{1}{p}(Y_i - \bar{Y}_P - (1-p)\hat{\delta}_P) \middle| C_j, D_i = 1 \right] p \\
& - E\left[\frac{1}{(1-p)}(Y_i - \bar{Y}_P + p\hat{\delta}_P) \middle| C_j, D_i = 0 \right] (1-p) \\
=& E\left[Y_i - \bar{Y}_P - (1-p)\hat{\delta}_P \middle| C_j, D_i = 1 \right] - E\left[Y_i - \bar{Y}_P + p\hat{\delta}_P \middle| C_j, D_i = 0 \right] \\
=& E\left[Y_i(1) - \bar{Y}_P \middle| C_j, D_i = 1 \right] - E\left[Y_i(0) - \bar{Y}_P \middle| C_j, D_i = 0 \right] - \hat{\delta}_P \\
=& E\left[Y_i(1) - Y_i(0) | C_j \right] - \hat{\delta}_P
\end{aligned}
$$

- Difference between approximated causal effect at child and parent node
- Pseudo-outcome is updated at each parent node

Local Centering

- Generalised causal forest is targeted to find maximum heterogeneity in pseudo-outcome
- But not specifically designed to account for selection into treatment (even though deep causal forests correct automatically to some extent for selection)
- Define centred variables

$$\tilde{Y}_i = Y_i - \hat{\mu}(X_i)$$

and

$$\tilde{D}_i = D_i - \hat{p}(X_i)$$

with $\hat{\mu}(x) = \hat{E}[Y_i|X_i = x]$ and $\hat{p}(x) = \hat{E}[D_i|X_i = x]$

- Apply generalised causal forest algorithm to $\tilde{Y}_i$ and $\tilde{D}_i$ instead of Y_i and D_i

$\rightarrow$ Orthogonalisation

Local Centering (cont.)

MSE from Simulation

Confounding	Heterogeneity	K	N	GCF	Centred GCF
No	No	10	800	0.85	0.87
No	No	10	1,600	0.58	0.59
No	No	20	800	0.92	0.93
No	No	20	1,600	0.52	0.52
Yes	No	10	800	1.12	0.27
Yes	No	10	1,600	0.80	0.20
Yes	No	20	800	1.17	0.17
Yes	No	20	1,600	0.95	0.11
Yes	Yes	10	800	1.92	0.91
Yes	Yes	10	1,600	1.51	0.62
Yes	Yes	20	800	1.92	0.93
Yes	Yes	20	1,600	1.55	0.57

Note: K is the number of covariates in the simulation

Source: Athey, Tibshirani, and Wager (2018)

Asymptotic Properties for Causal Forest

- Minimum subsample size S is scaled N^{β} with $\beta_{min} < \beta < 1$
- CATEs are consistent and asymptotically normal

$$(\hat{\delta}(x) - \delta(x))/\sqrt{Var(\delta(x))} \xrightarrow{d} N(0,1)$$

Reference: Wager and Athey (2018)

Inference

- Infinitesimal Jackknife
  - $Q_{ig} = 1$ when observation i is used to built tree g and $Q_{ig} = 0$ otherwise
  - Calculate the covariance $Cov(\hat{\delta}(x), Q_{ig})$ across all trees $g = 1, ..., G$
  - Variance estimator:

$$\hat{V}(x) = \frac{N-1}{N} \left(\frac{N}{N-S} \right)^2 \sum_{i=1}^{N} Cov(\hat{\delta}(x), Q_{ig})^2$$

  - $(N/(N-S)))^2$ is a finite sample correction for subsampling
  - $\hat{V}(x) \xrightarrow{p} Var(\delta(x))$

Reference: Wager and Athey (2018)

Advantages and Disadvantages of Causal Forest

Advantages:

- D_i can be binary or continuous
- Asymptotic properties for CATEs available
- Variance estimator available

Disadvantages:

- Best suited for experiments
- We have to assume that subsample sizes do not get too small, because otherwise asymptotic results break down