

Seattle Data Science Journal Club: Wager & Athey

Benjamin S. Skrainka

January 11, 2018

Seattle Data Science Journal Club

Building a platform where data scientists can discuss the latest and greatest in the field and network:

- A seminar series like in graduate school:
 - ▶ Discuss important papers with top Seattle data scientists
 - ▶ Remain current on latest ideas
 - ▶ Occurs every other month
- A speaker series:
 - ▶ Hear key data science thinkers
 - ▶ Occurs every other month + 1

This is a meetup to discuss the latest data science ideas:

- What topics interest you?
- Are there speakers you want to hear?
- What cadence?
- Do you want to help lead/organize this group?
- Let us know. . .

Wager & Athey (2017)

Wager & Athey (2017): *Estimation and Inference of Heterogeneous Treatment Effects Using Random Forest*

- Develop non-parametric *causal forest* to estimate causal effects:
 - ▶ Heterogeneous treatment effects
 - ▶ Extends on random forest
- First tools to perform valid statistical inference:
 - ▶ Asymptotically Gaussian
 - ▶ Pointwise consistent
 - ▶ Works for any random forest algorithm

Part of a new strand of literature uniting Econometrics and ML to estimate causal effects.

- ML:
 - ▶ Great for prediction
 - ▶ Great for large datasets
 - ▶ Poor for inference
- Econometrics/Applied Statistics/etc.:
 - ▶ Great for causality (e.g., Rubin)
 - ▶ Great for estimation and inference of causal effects
 - ▶ Poor for model selection and many features

Social science problems often consist of prediction + causal inference:

- Use ML for prediction, model selection, and robustness
- Extend to handle inference & estimation of causal effects

Data scientists often need to estimate the impact of a policy:

- Is feature X better than feature Y?
- Did our advertising work?

We can apply this literature to many problems we face, such as A/B testing

Real-world example: Ascarza (2016) *Retention futility: Targeting high risk customers might be ineffective*:

- Uses a similar method to measure heterogeneous response to churn intervention
- Computes optimal policy, which is counter to conventional wisdom

Some classics:

- Breiman (2004). *Random forests*
- Imbens & Rubin (2015). *Causal Inference*

Some recent papers:

- Athey & Imbens (2016). *Recursive partitioning for heterogeneous causal effects*
- Athey, Tibshirani, and Wager (2016). *Generalized random forests*
- Wager, Hastie, and Efron (2014). *Confidence intervals for random forests*

Paper tackles several problems:

- Gelman's "Garden of forking paths" – well-intentioned, ex-post data-driven hypothesis testing
 - ▶ Should pre-specify analysis plan
 - ▶ But, cannot anticipate all forms of heterogeneity ex-ante
- Optimal policy: must estimate treatment effect heterogeneity

Construct confidence intervals for estimates from modified random forest algorithm using several insights:

- Estimate treatment effects using RF to determine “nearby” observations
 - ▶ I.e., with correct splitting, each leaf should be (close to) a random experiment with nigh identical units
- Cross-validation for inference (*honest trees*)
- *Given a tree built on the training set, can use any valid method to estimate τ on test set*
- Prediction at individual and not leaf/group level (e.g., Athey & Imbens (2016))

Applied to decision trees (Athey & Imbens) and random forests (this paper).

Organization of the paper:

- 1 Prove consistency & asymptotic normality for a variant of RF
- 2 Prove infinitesimal jackknife consistent for aVar
- 3 Extend results to estimation of heterogeneous treatment effects in potential outcomes framework
- 4 Compare causal forest vs. k-NN using simulations

Notation: potential outcomes notation

Paper uses potential outcomes notation:

- Outcome is $Y_i(W_i)$ for individual i with treatment status W_i
- Treatment is $W_i \in \{0, 1\}$
- Want to measure causal effect, $\tau(x)$ at x

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

but we cannot observe both $Y_i(1)$ and $Y_i(0)$...

Prediction at x :

$$\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \cdot \sum_{\{i: X_i \in L(x)\}} Y_i$$

Let

$$\hat{\mu}(x|w) = \frac{1}{|\{i : W_i = w, X_i \in L(x)\}|} \cdot \sum_{\{i: W_i = w, X_i \in L(x)\}} Y_i$$

Estimation of treatment effect at x for a causal tree:

$$\hat{\tau}(x) = \hat{\mu}(x|w = 1) - \hat{\mu}(x|w = 0)$$

For a RF with B trees, $\hat{\tau}(x) = \frac{1}{B} \cdot \sum_{b=1}^B \hat{\tau}_b(x)$

Estimation of variance

Estimation of variance uses:

$$\hat{V}_{IJ}(x) = \frac{n-1}{n} \left(\frac{n}{n-s} \right)^2 \sum_{i=1}^n \text{Cov}_*[\hat{\tau}_b^*(x), N_{ib}^*],$$

where:

- $\text{Cov}_*[\cdot, \cdot]$ is over all trees $b = 1, \dots, B$
- N_{ib}^* indicates whether observation i is in tree b

Unconfoundedness:

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$
- As if a neighborhood is a randomized experiment

Overlap (probabilistic):

- $\epsilon < \mathbb{P}[W = 1 \mid X = x] < 1 - \epsilon$
- For large n , a neighborhood contains both treatments

Asymptotic normality and consistency require:

- Subsample size s scales appropriately
- *Honest* trees:
 - ▶ *Double-sample* trees use two samples:
 - ★ \mathcal{I} : used to estimate effects within each leaf
 - ★ \mathcal{J} : used to determine splits
 - ▶ *Propensity* trees:
 - ★ Ignore Y_i when computing splits
 - ★ Train classification tree for W_i
 - ★ Estimate leaf-level responses
 - ★ In tradition of propensity matching

Procedure 1: double-sample trees

Plan: split the sample and use one half to build tree and other to estimate:

- 1 Draw random subsample of size s without replacement
- 2 Split it into \mathcal{I} and \mathcal{J}
- 3 Grow tree via recursive partitioning:
 - ▶ Use any data in \mathcal{J} and only X or W in \mathcal{I}
 - ▶ Do not use Y in \mathcal{I}
- 4 Estimate $\hat{\tau}(x)$ using only \mathcal{I}

Note: must sample without replacement

Procedure 2: propensity trees

Use W_i to determine splits and Y_i to estimate τ :

- 1 Draw random subsample \mathcal{I} of size s without replacement
- 2 Train classification tree using \mathcal{I} using W_i as label and X_i as features
 - ▶ Must have $\geq k$ observations in each leaf for each treatment
 - ▶ Can optimize using Gini criterion, entropy, etc.
- 3 Estimate $\tau(x)$ on $L(x)$

Key definitions:

- **honest** tree:

- ▶ Double-sample tree: does not use Y_i in \mathcal{I} to choose splits
- ▶ Propensity tree: does not use Y_i to choose splits

- **random-split** tree:

- ▶ Marginalize over auxiliary randomness, $\xi \sim \Xi$ in RF
- ▶ At every split, $\pi/d < \mathbb{P}[\text{split along } j\text{-th feature}], \forall \pi \in (0, 1]$
- ▶ Note: ξ contains randomness for splitting features

Key definitions:

- **α -regular** $\forall \alpha > 0$ if:
 - ▶ standard case:
 - ★ At least α of training observations on each side of split
 - ★ Terminal nodes have at between k and $2k - 1$ observations
 - ▶ double-sample: \mathcal{I} satisfies above condition
- **symmetric**:
 - ▶ Output of predictor independent of order of training set

Results are theoretical with some confirmation via simulation:

- Theorems on asymptotic normality of mean and treatment effects
- Simulation experiments

Then they prove some theorems, given regularity conditions:

- Theorem 1: $\frac{\hat{\mu}_n - \mu(x)}{\sigma_n(x)} \underset{a}{\approx} N(0, 1)$
- Lemma 2: probability limit on $\text{diam}_j(L(x))$
- Theorem 3: $|\mathbb{E}[\hat{\mu}_n(x)] - \mu(x)| = \mathcal{O}(f(s, \alpha))$
- Theorem 11: causal forest has:
 - ▶ Predictions $\hat{\tau}(x)$ are consistent and asymptotically Gaussian and centered
 - ▶ Variance that is consistently estimated

Compare causal forest to k-NN:

- CF provides:
 - ▶ Superior matching
 - ▶ Stable MSE which is \ll MSE of k-NN
- CF coverage deteriorates for more than ≈ 10 features
- In some simulations, bias dominates variance for RF and causes uncentered CI

\Rightarrow need to improve control for bias, perhaps via better splitting rule

Some questions:

- Simulations use large enough n ?
- How much data needed to succeed?
- How much overlap needed to measure τ well?
- Performance vs. classical methods (propensity score, matching)?
- Why not split on entropy?
- How to identify lack of balance in leaves?
- Performance on real data?

Future meetings

Thanks for attending!

We will meet every month:

- $\text{month} \% 2 == 1 \Rightarrow$ discuss a paper
- $\text{month} \% 2 == 0 \Rightarrow$ listen to a speaker
- Next month's speaker: will announce next week