Machine Learning Methods for Causal Effects

Susan Athey, Stanford University
Guido Imbens, Stanford University

Introduction

Supervised Machine Learning v. Econometrics/Statistics Lit. on Causality

Supervised ML

- Well-developed and widely used nonparametric prediction methods that work well with big data
  - Used in technology companies, computer science, statistics, genomics, neuroscience, etc.
  - Rapidly growing in influence
- Cross-validation for model selection
- Focus on prediction and applications of prediction
- Weaknesses
  - Causality (with notable exceptions, including those attending this conference)

Econometrics/Soc Sci/Statistics

- Formal theory of causality
  - Potential outcomes method (Rubin) maps onto economic approaches
- "Structural models" that predict what happens when world changes
  - Used for auctions, anti-trust (e.g. mergers) and business decision-making (e.g. pricing)
- Well-developed and widely used tools for estimation and inference of causal effects in exp. and observational studies
  - Used by social science, policy-makers, development organizations, medicine, business, experimentation
- Weaknesses
  - Non-parametric approaches fail with many covariates
  - Model selection unprincipled

A Research Agenda

Problems

- Many problems in social sciences entail a combination of prediction and causal inference
- Existing ML approaches to estimation, model selection and robustness do not directly apply to the problem of estimating causal parameters
- Inference more challenging for some ML methods

Proposals

- Formally model the distinction between causal and predictive parts of the model and treat them differently for both estimation and inference
  - Abadie, Athey, Imbens and Wooldridge (2014, under review)
- Develop new estimation methods that combine ML approaches for prediction component of models with causal approaches
  - Today's paper, Athey-Imbens (WIP)
- Develop new approaches to cross-validation optimized for causal inference
  - Today's paper, Athey-Imbens (WIP)
- Develop robustness measures for causal parameters inspired by ML
  - Athey-Imbens (*AER* 2015)

Model for Causal Inference

- For causal questions, we wish to know what would happen if a policy-maker changes a policy
  - Potential outcomes notation:
    - $Y_i(w)$ is the outcome unit i would have if assigned treatment w
    - For binary treatment, treatment effect is $\tau_i = Y_i(1) - Y_i(0)$
  - Administer a drug, change minimum wage law, raise a price
  - Function of interest: mapping from alt. CF policies to outcomes
  - Holland: Fundamental Problem of Causal Inference
    - We do not see the same units at the same time with alt. CF policies

- Units of study typically have fixed attributes x_i
  - These would not change with alternative policies
  - E.g. we don't contemplate moving coastal states inland when we change minimum wage policy

Inference for Causal Effects v. Attributes: Abadie, Athey, Imbens & Wooldridge (2014)

Approach

- Formally define a population of interest and how sampling occurs
- Define an estimand that answers the economic question using these objects (effects versus attributes)
- Specify: "What data are missing, and how is the difference between your estimator and the estimand uncertain?"
  - Given data on 50 states from 2003, we know with certainty the difference in average income between coast and interior
  - Although we could contemplate using data from 2003 to estimate the 2004, difference this depends on serial correlation within states, no direct info in cross-section

Application to Effects v. Attributes in Regression Models

- Sampling: Sample/population does not go to zero, finite sample
- Causal effects have missing data: don't observe both treatments for any unit
- Huber-White robust standard errors are conservative but best feasible estimate for causal effects
- Standard errors on fixed attributes may be much smaller if sample is large relative to population
  - Conventional approaches take into account sampling variance that should not be there

Robustness of Causal Estimates
Athey and Imbens (AER, 2015)

- General nonlinear models/estimation methods
- Causal effect is defined as a function of model parameters
  - Simple case with binary treatment, effect is $\tau_i = Y_i(1) - Y_i(0)$
- Consider other variables/features as "attributes"
- Proposed metric for robustness:
  - Use a series of "tree" models to partition the sample by attributes
    - Simple case: take each attribute one by one
  - Re-estimate model within each partition
  - For each tree, calculate overall sample average effect as a weighted average of effects within each partition
  - This yields a set of sample average effects
  - Propose the standard deviation of effects as robustness measure
- 4 Applications:
  - Robustness measure better for randomized experiments, worse in observational studies

Machine Learning Methods for Estimating Heterogeneous Causal Effects

Susan Athey and Guido Imbens

Motivation I: Experiments and Data-Mining

- Concerns about ex-post "data-mining"
  - In medicine, scholars required to pre-specify analysis plan
  - In economic field experiments, calls for similar protocols
- But how is researcher to predict all forms of heterogeneity in an environment with many covariates?
- Goal:
  - Allow researcher to specify set of potential covariates
  - Data-driven search for heterogeneity in causal effects with valid standard errors

Motivation II: Treatment Effect Heterogeneity for Policy

- Estimate of treatment effect heterogeneity needed for optimal decision-making
- This paper focuses on estimating treatment effect as function of attributes directly, not optimized for choosing optimal policy in a given setting
- This "structural" function can be used in future decision-making by policy-makers without the need for customized analysis

Preview

- Distinguish between causal effects and attributes
- Estimate treatment effect heterogeneity:
  - Introduce estimation approaches that combine ML prediction & causal inference tools
- Introduce and analyze new cross-validation approaches for causal inference
- Inference on estimated treatment effects in subpopulations
  - Enabling post-experiment data-mining

Regression Trees for Prediction

Data

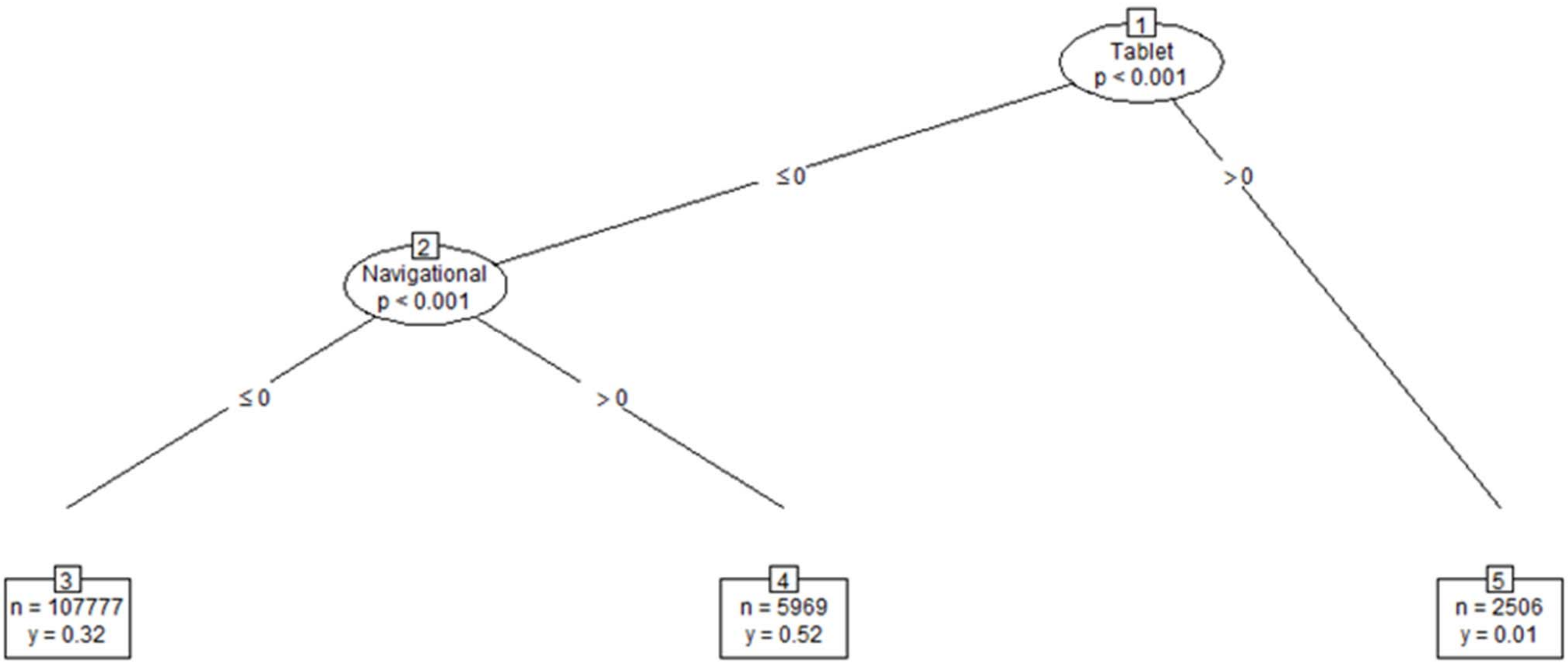
- Outcomes Y_i, attributes X_i.
- Support of X_i is $\mathcal{X}$.
- Have training sample with independent obs.
- Want to predict on new sample
- Ex: Predict how many clicks a link will receive if placed in the first position on a particular search query

Build a "tree":

- Partition of $\mathcal{X}$ into "leaves" $\mathcal{X}_j$
- Predict Y conditional on realization of X in each region $\mathcal{X}_j$ using the sample mean in that region
- Go through variables and leaves and decide whether and where to split leaves (creating a finer partition) using in-sample goodness of fit criterion
- Select tree complexity using cross-validation based on prediction quality

Regression Tree Illustration

Outcome: CTR for position 1 in subsample of
Bing search queries from 2012
(sample is non-representative)



Regression Trees for Prediction: Components

1. Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{Y}_i$: sample mean of Y_i within leaf
   C. Set of candidate estimators C: correspond to different specifications of how tree is split
2. Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

      $$Q^{is} = \text{-MSE (Mean Squared Error)} = -\frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2$$

   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda \times \#$ leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: $Q^{os} = -MSE$

Using Trees to Estimate Causal Effects

Model:

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & if\ \ W_i = 1, \\ Y_i(0) & otherwise. \end{cases}$$

- Suppose random assignment of W_i
- Want to predict individual *i*'s treatment effect
  - $\tau_i = Y_i(1) - Y_i(0)$
  - This is not observed for any individual
  - Not clear how to apply standard machine learning tools
- Let

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

Using Trees to Estimate Causal Effects

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

- Approach 1: Analyze two groups separately
  - Estimate $\hat{\mu}(1,x)$ using dataset where $W_i = 1$
  - Estimate $\hat{\mu}(0,x)$ using dataset where $W_i = 0$
  - Use propensity score weighting (PSW) if needed
  - Do within-group cross-validation to choose tuning parameters
  - Construct prediction using
    $$\hat{\mu}(1,x) - \hat{\mu}(0,x)$$
- Approach 2: Estimate $\mu(w,x)$ using tree including both covariates
  - Include PS as attribute if needed
  - Choose tuning parameters as usual
  - Construct prediction using
    $$\hat{\mu}(1,x) - \hat{\mu}(0,x)$$
  - Estimate is zero for *x* where tree does not split on *w*

- Observations
  - Estimation and cross-validation not optimized for goal
  - Lots of segments in Approach 1: combining two distinct ways to partition the data

- Problems with these approaches
  1. Approaches not tailored to the goal of estimating treatment effects
  2. How do you evaluate goodness of fit for tree splitting and cross-validation?
     - $\tau_i = Y_i(1) - Y_i(0)$ is not observed and thus you don't have ground truth for any unit

Literature

Approaches in the spirit of single tree and two trees

- Beygelzimer and Langford (2009)
  - Analogous to "two trees" approach with multiple treatments; construct optimal policy
- Dudick, Langford, and Li (2011)
  - Combine inverse propensity score method with "direct methods" (analogous to single tree approach) to estimate optimal policy
- Foster, Tailor, Ruberg, *Statistics and Medicine* (2011)
  - Estimate $\mu(w,x)$ using random forests, define $\hat{\tau}_i = \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i)$, and do trees on $\hat{\tau}_i$.
- Imai and Ratkovic (2013)
  - In context of randomized experiment, estimate $\mu(w,x)$ using lasso type methods, and then $\hat{\tau}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$.

Estimating treatment effects directly at leaves of trees

- Su, Tsai, Wang, Nickerson, Li (2009)
  - Do regular tree, but split if the t-stat for the treatment effect difference is large, rather than when the change in prediction error is large.
- Zeileis, Hothorn, and Hornick (2005)
  - "Model-based recursive partitioning": estimate a model at the leaves of a tree. In-sample splits based on prediction error, do not focus on out of sample cross-validation for tuning.
- None of these explore cross-validation based on treatment effect.

Proposed Approach 3: Transform the Outcome

- Suppose we have 50-50 randomization of treatment/control
  - Let $Y_i^* = \begin{cases} 2Y_i & if\ W_i = 1 \\ -2Y_i & if\ W_i = 0 \end{cases}$
  - Then $E[Y_i^*] = 2 \cdot \left(\frac{1}{2}E[Y_i(1)] - \frac{1}{2}E[Y_i(0)]\right) = E[\tau_i]$

- Suppose treatment with probability p_i
  - Let $Y_i^* = \frac{W_i - p}{p(1-p)} Y_i = \begin{cases} \frac{1}{p} Y_i & if\ W_i = 1 \\ -\frac{1}{1-p} Y_i & if\ W_i = 0 \end{cases}$
  - Then $E[Y_i^*] = \left(p\frac{1}{p}E[Y_i(1)] - (1-p)\frac{1}{1-p}E[Y_i(0)]\right) = E[\tau_i]$

- Selection on observables or stratified experiment
  - Let $Y_i^* = \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))} Y_i$
  - Estimate $\hat{p}(x)$ using traditional methods

Causal Trees:
Approach 3 (Conventional Tree, Transformed Outcome)

1. Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{\tau}_i^*$: sample mean of Y_i^* within leaf
   C. Set of candidate estimators *C*: correspond to different specifications of how tree is split
2. Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

      $$Q^{is} = \text{-MSE (Mean Squared Error)} = -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^* - Y_i^*)^2$$

   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda \times \#$ leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: Q^{os} = -MSE

Critique of Proposed Approach 3: Transform the Outcome

$$Y_i^* = \frac{W_i - p}{p(1-p)} Y_i = \begin{cases} \frac{1}{p} Y_i & if\ W_i = 1 \\ -\frac{1}{1-p} Y_i & if\ W_i = 0 \end{cases}$$

- Within a leaf, sample average of Y_i^* is not most efficient estimator of treatment effect
  - The proportion of treated units within the leaf is not the same as the overall sample proportion
- This motivates Approach 4: use sample average treatment effect in the leaf

Causal Trees:
Approach 4 (Causal Tree, Version 1)

1. Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{\tau}_i^{CT}$: sample average treatment effect within leaf (w/ PSW)
   C. Set of candidate estimators *C*: correspond to different specifications of how tree is split
2. Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

      Q^{is} = -MSE (Mean Squared Error) $= -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{CT} - Y_i^*)^2$
   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda \times$ # leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: Q^{os} = -MSE

Designing a Goodness of Fit Measure: What are other alternatives?

- Goodness of fit (infeasible):

$$Q^{\text{infeas}}(\hat{\tau}) = -\mathbb{E}[(\tau_i - \hat{\tau}(X_i))^2]$$

- Expanding, we have:

$$Q^{\text{infeas}}(\hat{\tau}) = -\mathbb{E}\left[\tau_i^2\right] - \mathbb{E}[\hat{\tau}^2(X_i)] + 2\,\mathbb{E}[\hat{\tau}(X_i)\cdot\tau_i]$$

  - First term doesn't depend on $\hat{\tau}$, thus irrelevant for comparing candidate estimators
  - Second term is straightforward to calculate given $\hat{\tau}$.
  - Third expectation:

$$\mathbb{E}[\hat{\tau}(X_i)\cdot\tau_i] = \mathbb{E}[\hat{\tau}(X_i)\cdot Y_i(1) - \hat{\tau}(X_i)\cdot Y_i(0)],$$

  - Effect of treatment on (alt) transformed outcome: $\tilde{Y}_i = Y_i \cdot \hat{\tau}(X_i)$.
  - Can be estimated. (Unusual to estimate fit measure.)
    - One alternative: matching. For computational reasons, we currently only use this to compare different overall approaches.

Estimating the In Sample Goodness of Fit Measure

- For tree splitting/comparing nested trees:

$$\mathbb{E}[\hat{\tau}(X_i)\cdot\tau_i] = \sum_j \mathbb{E}\left[\hat{\tau}(X_i)\cdot\tau_i | X_i \in S_j\right]\Pr(X_i \in S_j)$$

To estimate this, use fact that $\hat{\tau}(x_i)$ is constant within a segment, and is an estimate of $\mathbb{E}\left[\tau_i | X_i \in s_j(x_i)\right]$:

$$= \frac{1}{N}\sum_i \hat{\tau}^2(x_i)$$

- This motivates $Q^{is,sq}(\hat{\tau}) = \frac{1}{N}\sum_i \hat{\tau}^2(x_i)$
- Rewards variance of estimator (all candidates constrained to have same mean, and accurate mean on every segment)
- In expectation, but not in finite samples, compares alternative estimators the same as using $-\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{CT} - Y_i^*)^2$

Causal Trees:
Approach 5 (Modified Causal Tree)

1. Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{\tau}_i^{MCT}$: sample average treatment effect within leaf
   C. Set of candidate estimators *C*: correspond to different specifications of how tree is split
2. Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

$$Q^{is} = -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{MCT})^2$$

   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda$ x # leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: Q^{os} = -MSE=$-\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{MCT} - Y_i^*)^2$

Comparing “Standard” and Causal Approaches

- They will be more similar
  - If treatment effects and levels are highly correlated
- Two-tree approach
  - Will do poorly if there is a lot of heterogeneity in levels that is unrelated to treatment effects
  - Will do well in certain specific circumstances, e.g.
    - Control outcomes constant in covariates
    - Treatment outcomes vary with covariates

- How to compare approaches?

1. Oracle (simulations)
2. Transformed outcome goodness of fit
3. Use matching to estimate infeasible goodness of fit

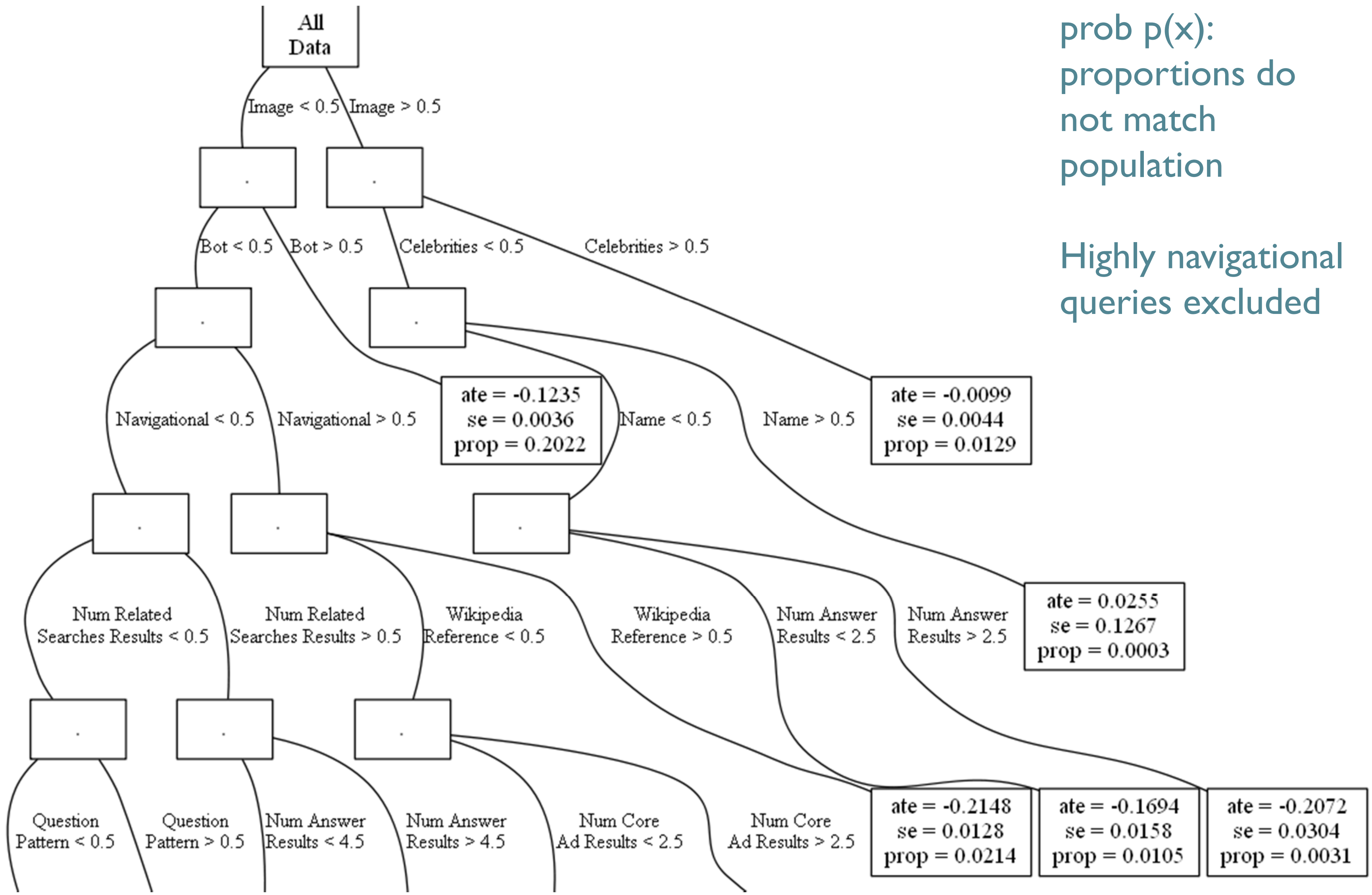
Inference

- Attractive feature of trees:
  - Can easily separate tree construction from treatment effect estimation
  - Tree constructed on training sample is independent of sampling variation in the test sample
  - Holding tree from training sample fixed, can use standard methods to conduct inference within each leaf of the tree on test sample
    - Can use any valid method for treatment effect estimation, not just the methods used in training
  - For observational studies, literature (e.g. Hirano, Imbens and Ridder (2003)) requires additional conditions for inference
    - E.g. leaf size must grow with population

Problem: Treatment Effect Heterogeneity in Estimating Position Effects in Search

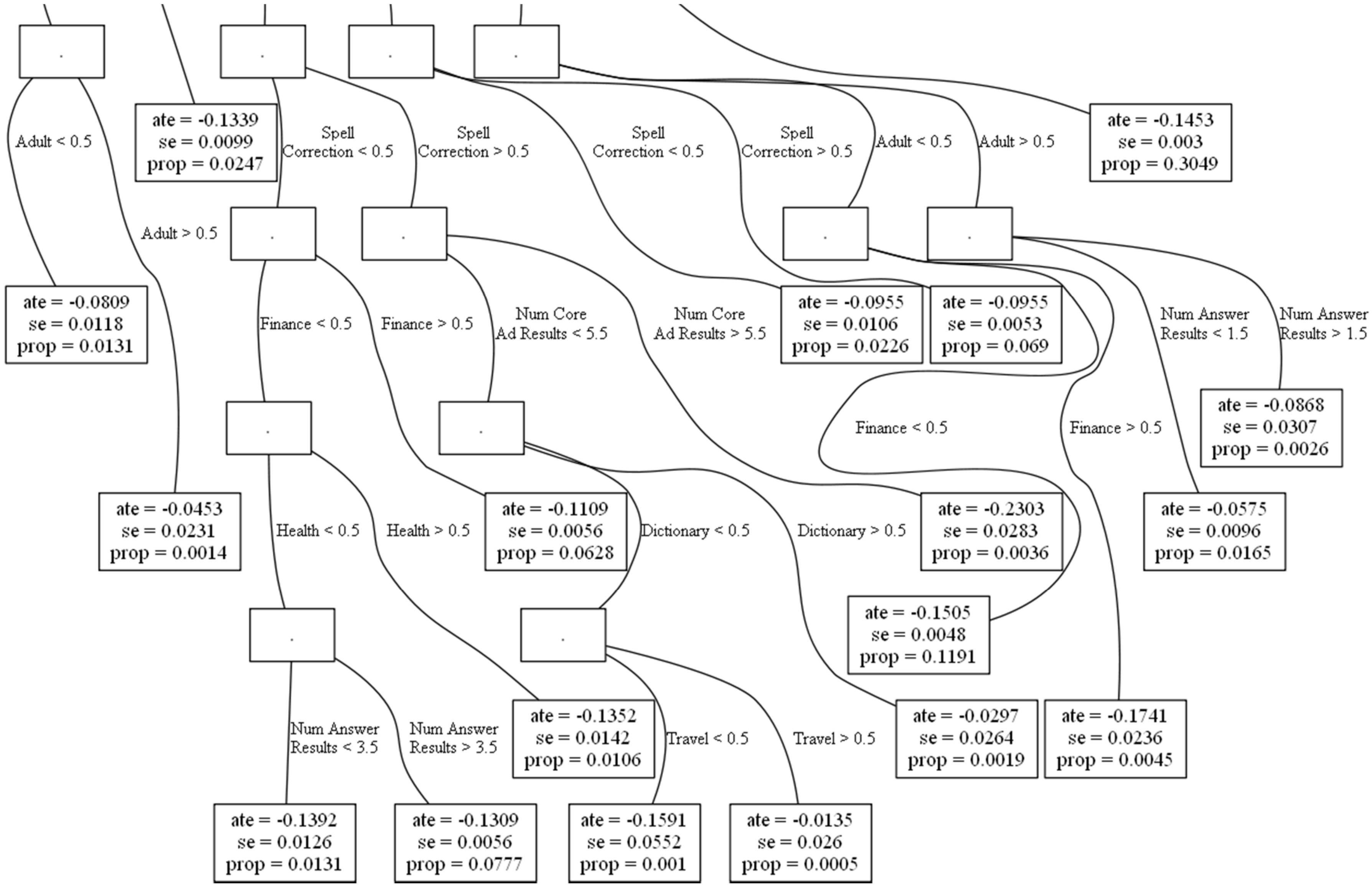
- Queries highly heterogeneous
  - Tens of millions of unique search phrases each month
  - Query mix changes month to month for a variety of reasons
  - Behavior conditional on query is fairly stable
- Desire for segments.
  - Want to understand heterogeneity and make decisions based on it
  - “Tune” algorithms separately by segment
  - Want to predict outcomes if query mix changes
    - For example, bring on new syndication partner with more queries of a certain type

Search Experiment Tree: Effect of Demoting Top Link (Test Sample Effects)



Some data excluded with prob p(x): proportions do not match population

Highly navigational queries excluded


Adult < 0.5
Adult > 0.5
ate = -0.1339
se = 0.0099
prop = 0.0247
Spell Correction < 0.5
Spell Correction > 0.5
Spell Correction < 0.5
Spell Correction > 0.5
Adult < 0.5
Adult > 0.5
ate = -0.1453
se = 0.003
prop = 0.3049
ate = -0.0809
se = 0.0118
prop = 0.0131
Finance < 0.5
Finance > 0.5
Num Core Ad Results < 5.5
Num Core Ad Results > 5.5
ate = -0.0955
se = 0.0106
prop = 0.0226
ate = -0.0955
se = 0.0053
prop = 0.069
Num Answer Results < 1.5
Num Answer Results > 1.5
ate = -0.0868
se = 0.0307
prop = 0.0026
Finance < 0.5
Finance > 0.5
ate = -0.0453
se = 0.0231
prop = 0.0014
Health < 0.5
Health > 0.5
ate = -0.1109
se = 0.0056
prop = 0.0628
Dictionary < 0.5
Dictionary > 0.5
ate = -0.2303
se = 0.0283
prop = 0.0036
ate = -0.0575
se = 0.0096
prop = 0.0165
ate = -0.1505
se = 0.0048
prop = 0.1191
Num Answer Results < 3.5
Num Answer Results > 3.5
ate = -0.1352
se = 0.0142
prop = 0.0106
Travel < 0.5
Travel > 0.5
ate = -0.0297
se = 0.0264
prop = 0.0019
ate = -0.1741
se = 0.0236
prop = 0.0045
ate = -0.1392
se = 0.0126
prop = 0.0131
ate = -0.1309
se = 0.0056
prop = 0.0777
ate = -0.1591
se = 0.0552
prop = 0.001
ate = -0.0135
se = 0.026
prop = 0.0005

Use Test Sample for Segment Means & Std Errors to Avoid Bias

Variance of estimated treatment effects in training sample 2.5 times that in test sample

Test Sample			Training Sample		
Treatment Effect	Standard Error	Proportion	Treatment Effect	Standard Error	Proportion
-0.124	0.004	0.202	-0.124	0.004	0.202
-0.134	0.010	0.025	-0.135	0.010	0.024
-0.010	0.004	0.013	-0.007	0.004	0.013
-0.215	0.013	0.021	-0.247	0.013	0.022
-0.145	0.003	0.305	-0.148	0.003	0.304
-0.111	0.006	0.063	-0.110	0.006	0.064
-0.230	0.028	0.004	-0.268	0.028	0.004
-0.058	0.010	0.017	-0.032	0.010	0.017
-0.087	0.031	0.003	-0.056	0.029	0.003
-0.151	0.005	0.119	-0.169	0.005	0.119
-0.174	0.024	0.005	-0.168	0.024	0.005
0.026	0.127	0.000	0.286	0.124	0.000
-0.030	0.026	0.002	-0.009	0.025	0.002
-0.135	0.014	0.011	-0.114	0.015	0.010
-0.159	0.055	0.001	-0.143	0.053	0.001
-0.014	0.026	0.001	0.008	0.050	0.000
-0.081	0.012	0.013	-0.050	0.012	0.013
-0.045	0.023	0.001	-0.045	0.021	0.001
-0.169	0.016	0.011	-0.200	0.016	0.011
-0.207	0.030	0.003	-0.279	0.031	0.003
-0.096	0.011	0.023	-0.083	0.011	0.022
-0.096	0.005	0.069	-0.096	0.005	0.070
-0.139	0.013	0.013	-0.159	0.013	0.013
-0.131	0.006	0.078	-0.128	0.006	0.078

Conclusions

- Key to approach
  - Distinguish between causal and predictive parts of model

- "Best of Both Worlds"
  - Combining very well established tools from different literatures
  - Systematic model selection with many covariates
  - Optimized for problem of causal effects
    - In terms of tradeoff between granular prediction and overfitting
  - With valid inference
  - Easy to communicate method and interpret results
    - Output is a partition of sample, treatment effects and standard errors

- Important application
  - Data-mining for heterogeneous effects in randomized experiments