CAUSAL INFERENCE AT HULU

Allen Tran

July 17, 2016

Hulu

Introduction

Most interesting business questions at Hulu are **causal**

Business: what would happen if we did *x* instead of *y*?

- dropped prices for risky subs
- halved our AdWords marketing spend
- bought some piece of Content

Causal inference is not supervised learning

Supervised learning model ($f : X \to Y$)

- Optimized for: $f(x) \approx y$
- test-train paradigm works
- regularization works

Causal Inference model ($f : X, T \to Y$)

- Optimized to identify treatment effect: $f(x, T = 1) - f(x, T = 0)$
- there is no test set, need to lean on statistical theory
- naive regularization is a horrible idea

Sacrifice predictive performance to identify treatment effect

Roadmap

1. Potential outcomes framework/treatment effects
2. What we do at Hulu
3. Modern ML approaches to causality (we should do this)

Goal: Convince us to think less supervised learning, more causal inference

Note: focus will be on observation studies, not designing randomized trials

Causal inference = identifying Treatment effects

Let x_i describe some features (a user), and two potential outcomes, $Y_1(x_i)$ and $Y_0(x_i)$.

$Y_k(x|T=k)$ is observed, $Y_k(x|T=1-k)$ is the unobserved counterfactual.

- individual treatment effect

$$ITE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}\left(Y_1(x_i)\right) - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}\left(Y_0(x_i)\right) \tag{1}$$

- heterogeneous treatment effect

$$HTE(X) = \mathbb{E}_{Y_1 \sim p(Y_1|x \in X)}\left(Y_1(x)\right) - \mathbb{E}_{Y_0 \sim p(Y_0|x \in X)}\left(Y_0(x)\right) \tag{2}$$

- average treatment effect

$$ATE = \mathbb{E}\left(Y_1(x) - Y_0(x)\right) = \mathbb{E}_{x \sim p(x)}\left(ITE(x)\right) \tag{3}$$

Assumptions

Ignorability: assignment to treatment/control effectively random, conditioning on x

$$(Y_0, Y_1) \perp\!\!\!\perp T \quad | \quad x \tag{4}$$

or equivalently,

$$p(Y_0, Y_1, T|x) = p(Y_0, Y_1|x)p(T|x) \tag{5}$$

Overlap/Common support: treatment/control across all units

$$p(t|x) > 0 \quad \forall t, x \tag{6}$$

Estimating Average Treatment Effects

First part of ATE

$$\mathbb{E}_{x,Y_1,T}\left(Y_1(x)\right) = \mathbb{E}_x\left(\mathbb{E}_{Y_1,T|x}Y_1(x,T)\right) \tag{7}$$

$$= \mathbb{E}_x\left(\mathbb{E}_{Y_1|x}Y_1(x,T=1)\right) \tag{8}$$

Term inside parentheses is observed.

$$ATE = \mathbb{E}_x\left(\mathbb{E}_{Y_1|x}Y_1(x,T=1) - \mathbb{E}_{Y_0|x}Y_0(x,T=0)\right) \tag{9}$$

But x_i units we observe are distributed $p(x_i|T=1)$ and $p(x_i|T=0)$, not $p(x)$.

Randomized trial, $x \perp\!\!\!\perp T$, difference in means is sufficient.

Estimating causal effects in practice

In order of simplicity

1. Matching
2. Propensity score/reweighting
3. Covariate adjustment (current strategy at Hulu - incremental retention, portfolio, sub churn)
4. Modern ML methods designed for causal inference

Covariate adjustment

Supervised learning with treatment variable as a feature

$$ATE = \frac{1}{N}\sum_i f(x_i, T=1) - f(x_i, T=0) \tag{10}$$

Problems

- No goodness of fit metrics. Remember, we are fitting treatment effects, not Y
- Treatment effects wrong if model is misspecified (e.g unconfoundedness does not hold)
  - unobserved demographics behind watch/did not watch (treatment) in incremental retention model
  - specific types of shows (unobserved) get more marketing (treatment) in MMM model
- Regularization? May regularize away our treatment/potentially harmful

Casual Trees/Forests

Athey & Imbens (2015)

Directly estimate treatment effects via decision trees. Shown to be consistent estimator of treatment effect (i.e is a “correct” model).

Regular decision tree

1. Estimator (of outcome): mean of **outcome** in leaf

$$\hat{y}_l = \frac{1}{N_l} \sum_{i \in l} Y_i \tag{11}$$

2. Splitting criterion: metrics (information gain/Gini/variance reduction) for **outcome** + complexity penalty
3. Score: assess estimates of predicted **outcome** on test set, $\hat{y}_i$ against y_i

Casual Trees/Forests: Transforming Outcomes

Transform outcome Y_i. Let $p_i = p(T = 1|x_i)$

$$Y_i^* = \frac{T_i - p_i}{1 - p_i} Y_i \tag{12}$$

Expectation of transformed outcome is average treatment effect!

$$\begin{aligned}
\mathbb{E}_{x,T}\left(Y_i^*\right) &= \mathbb{E}_x\left(\mathbb{E}_T\left(Y_i^*\right)\right) & (13)\\
&= \mathbb{E}_x\left(p_i \mathbb{E}\left(Y_i^* | T = 1, x\right) + (1 - p_i)\mathbb{E}\left(Y_i^* | T = 0, x\right)\right) & (14)\\
&= \mathbb{E}_x\left(p_i Y_1(x_i) - (1 - p_i) Y_0(x_i)\right) & (15)\\
&= ATE & (16)
\end{aligned}$$

Casual Trees/Forests: Splitting

Ideal but infeasible criterion, compare treatment effect to ground truth

$$Q(\hat{\tau}) = \mathbb{E}\left((\tau_i - \hat{\tau}_i)^2\right) \tag{17}$$

But decompose this:

$$Q(\hat{\tau}) = \mathbb{E}\left(\tau)i^2\right) + \mathbb{E}\left(\hat{\tau}_i{}^2\right) - 2\mathbb{E}\left(\tau_i\hat{\tau}_i\right) \tag{18}$$

First term doesn't involve τ_i so can ignore. Second term is sample mean of estimator. Third term can be calculated (after more algebra)

$$\begin{aligned}\mathbb{E}\left(\tau_i\hat{\tau}_i\right) &= \mathbb{E}_\ell\left(\mathbb{E}\left(\tau_i\hat{\tau}_i|x_i \in \ell\right)\right)\\ &= \mathbb{E}_\ell\left(\hat{\tau}_l\mathbb{E}\left(\tau_i|x_i \in \ell\right)\right)\\ &= \mathbb{E}\left(\hat{\tau}_i{}^2\right)\end{aligned}$$

Hence, criterion to continue splitting is

$$Q(\hat{\tau}) = -\mathbb{E}\left(\hat{\tau}_i{}^2\right) \tag{19}$$

Rewards variance in estimates of treatment effects. (Compare to limiting case of no splits, every x_i has the same treatment effect).

Causal Tree/Forests: Implementation

1. Estimator (of **treatment effect**): $\hat{\tau}_i$
   Mean of transformed outcome (with propensity score weighting as treatment/control distributions in leaf are not representative).

$$\hat{\tau}_i = \frac{1}{N_l} \sum_{i \in \ell} Y_i^* \frac{1}{p(T = 1|x_i)} \tag{20}$$

2. Splitting criterion: MSE on **treatment effect** + complexity of number of leaves

$$\frac{1}{N} \sum_i \hat{\tau}_i^2 + \lambda \cdot n_{\text{leaves}} \tag{21}$$

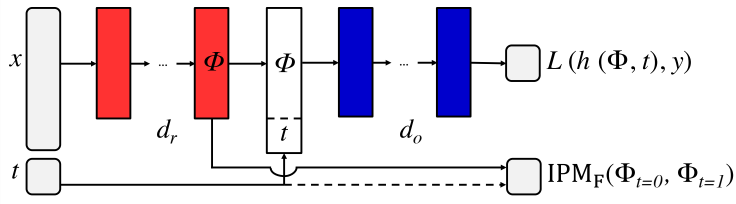
3. Score: out of sample MSE on **treatment effect**

$$\frac{1}{N_{\text{test}}} \sum_{i \in \text{test}} (\hat{\tau}_i - Y_i^*)^2 \tag{22}$$

Deep Learning/Regularizing Representations

Shalit, Johansson & Sontag (2016)

- Learn representations for x, $\Phi(x)$
- Merge representations and treatment to output prediction
- Measure difference in distribution of representations for treatment and control



Loss function

$$\min_{\Phi,h} \frac{1}{N} \sum_{i=1} L\left(h(\Phi(x_i), T_i), Y_i\right) + \alpha \cdot IPM_F\left(\{\Phi(x_i)\}_{i|T_i=0}, \{\Phi(x_i)\}_{i|T_i=1}\right) \quad (23)$$

- IPM_F is some distance metric over distributions
- Penalize differences between treatment and control group
  - mitigates problem that treatment is not randomly assigned (incremental retention: different demographics watch different shows)
  - regularization in the right place
- Theoretical result: if IPM_F is Wasserstein or Max-Mean Discrepancy, loss function is bound on true counterfactual error

Dealing with unconfoundedness/missing variables

What if we know that we have some unobserved variable that affects outcome?

Strategies

1. Causal graphs (Judea Pearl et al.) - assess whether it is a problem and potential fixes
2. Instrumental variables - find "instruments" that induce random-like properties into our x variables

Causal graphs

Bishop, Chapter 8: Pattern Recognition and Machine Learning

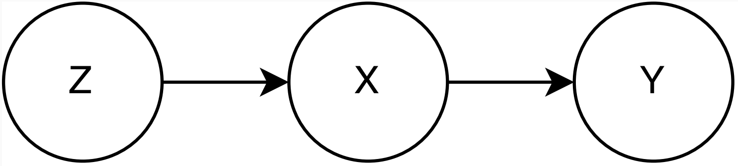
Represent casual effects via directed graphs.

Theoretical results ("d-separation"): infers conditional independence from directed graphs. We want

$$Y \perp\!\!\!\perp Z \quad | \quad X \tag{24}$$

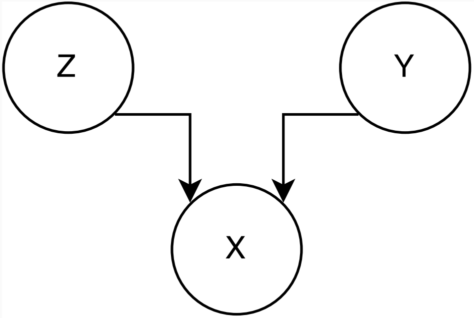
where Y is our outcome, Z is unobserved and X is observed

Causal Graphs: Head-to-tail



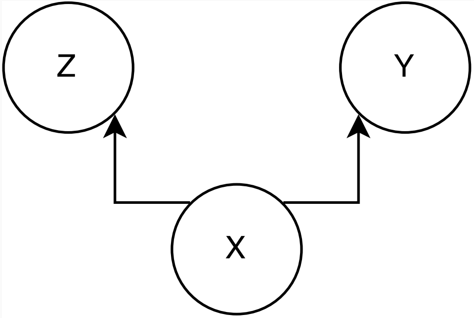
Observing X “blocks” the path from Z to Y, making Z independent of Y, conditional on X.

Causal Graphs: Head-To-Head



Observing X “blocks” the path from Z to Y, making Z independent of Y, conditional on X.

Causal Graphs: Tail-to-Tail

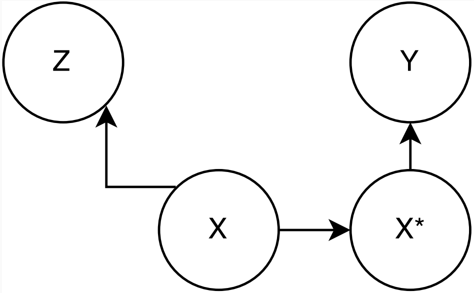


Observing *X* "**unblocks**" the path from *Z* to *Y*, making *Z* dependent of *Y*, conditional on *X*.

We can **not** use *X* as a dependent variable

Causal Graphs: Fixing Tail-to-Tail

We can fix the prior situation, by finding something else to condition on.



Conditioning on X^* "**blocks**" the path from Z to Y, making Z dependent of Y, conditional on X.

Instrumental Variables

Often (nearly all observational data) our X variable is correlated with unobserved variables making it impossible to identify the effect of X alone.

Idea: extract the random component on X uncorrelated with those other variables.

Instrumental Variables: Implementation

Suppose you have an instrument Z that is correlated with X but not Y.

Two Stage Least Squares

1. OLS of X on Z gets you X_{IV}
2. Second stage OLS of X_{IV} on Y

But linear regression is not very powerful, how to improve?

- What if I want to use a non-linear estimator (deep learning)?

Instrumental Variables: Examples

Note: there is a whole branch of economics dedicated to finding super fun/crazy instruments

- Effect of education on wages: education correlated with unobserved ability. Use exogenous districting/local government policies that force more/less education on people
- Watching show A on retention: but watch behavior correlated with unobserved demographics/preferences. Induce (some) randomness in reco strategy (change probability of exposure) and use exposure probabilities as instruments.

Conclusion

A lot of our questions are causal

- We try to answer these causal questions in a supervised learning heavy manner
- Supervised learning/covariate adjustment is not great at causal questions
- There are alternatives that make use of modern ML algorithms/big data and directly identify treatment effects