Causal Inference

Rahul G. Krishnan

New York University

Lab 13, Dec 7, 2016

Outline

1. Questions on Piazza
2. Motivation
3. Causal Inference
   - Framework and Assumptions
   - Matching
   - G-formula & propensity score
   - Covariate Adjustment
   - Structural Equation Models & Causal Graphs
4. Conclusion

CRFs and MaxEnt

- **Question:** “Can we interpret CRFs on a given graph trained by MLE as Cond. MaxEnt models where the graph structure is encoded in the feature function?”
- In a Max.Ent model, the choice of statistics you choose to match in the set of your constraints (which pairwise etc) define your graph structure

Quality of Research Work

- **Question:** “To answer these questions, are there any deterministic measures? Or the judgement mainly comes from the research experience and intuition?”
- Experiments. Are they asking the right questions. When does their method excel/fail?
- Novelty is often about knowing prior work and using your internal model of how the authors’ idea might be adopted and used by others
- Reviewing papers is very subjective (NIPS Experiment)
- “57% of papers at NIPS would be rejected if one reran the conference review process (with a 95% confidence interval of 40-75%):”

Why should we care about causal inference?

- Algorithms are becoming more and more prevalent in our daily lives whether we like it or not
- AI for Starcraft or compiling daily email : fairly harmless
- Which drug for a critically-ill patient received?
- Length of a person's prison sentence?
- These are asking *causal* questions! Important to know the limitations of the algorithms.

Fairness in ML

- ProPublica: Machine Bias
- Code & Data for ProPublica Article
- Can we create algorithms that are transparent to inspection, fair and open to criticism
- Movement in ML: Fairness and Transparency in Machine Learning

Potential Outcomes Framework

- Each unit x_i has two potential outcomes $Y_0(x_i)$ (control outcome) and $Y_1(x_i)$ (treated outcome)
- We only observed *one* of the outcomes for x_i during training
- Individual Treatment Effect (for personalized medicine) $\mathbb{E}_{p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{p(Y_0|x_i)}[Y_0|x_i]$
- Average Treatment Effect (for drug effectiveness) $\mathbb{E}[Y_1 - Y_0]$

Assumptions

- No unmeasured confounding (aka ignorability, exchangeability)
  - Bad when the confounder affects treatment assignment and outcome
  - $(Y_0, Y_1) \perp\!\!\!\perp T | x$
- Common support (aka overlap, positivity)
  - If only males received no job training, and females did, then we would erroneously conclude that being female = jobs
  - $p(T = t | X = x) > 0 \forall t, x$

In class

- **Key Challenge:** Controlling for confounding!
- Why supervised learning for $p(y|x, t)$ isn't enough.
  - Can ignore t
  - High-dimensional x can be challenging

A simple causal graph

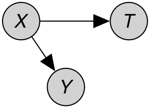


Figure: A simple causal graph that satisfies ignorability. T (Treatment), Y (Outcome), X (Features)

Method 1: Matching

- Define $d(\cdot, \cdot)$ a metric between x and $j(i) = \arg\min_{js.tt_j \neq t_i} d(x_j, x_i)$
- If treated, find closest control and vice versa
- $I\hat{T}E(x_i) = y_i - y_{j(i)}$ if i treated
- $I\hat{T}E(x_i) = y_{j(i)} - y_i$ if i control
- $A\hat{T}E = \frac{1}{n} \sum_i^n I\hat{T}E(x_i)$

Matching: Yay or Nay

- Interpretable (for small samples)
- Non-parametric (no model)
- Relies on metric d (could be misled)

A Technical Difficulty

- Matching created *artificial counterfactual* samples
- Estimate the average treatment effect from the (factual, matched counterfactual) tuples directly
- We cannot find a way to estimate $\mathbb{E}[Y_0]$ directly
- Never observe it since in our data we only observe Y_0 for patients who did not get treatment $T = 0$
- Can we get around this?

Method 2: Adjustment Formula - G formula

- Allows us to write the ATE as a function of quantities we can form emperical estimates for from data

$$
\begin{aligned}
&\mathbb{E}[Y_0](\text{cannot estimate from data}) \\
&= \mathbb{E}_{p(x)}\mathbb{E}_{p(Y_0|x)}[Y_0|x] \\
&= \mathbb{E}_{p(x)}\mathbb{E}_{p(Y_0|x)}[Y_0|x, T = 0] \\
&= \mathbb{E}_{p(x)}\mathbb{E}[Y_0|x, T = 0]
\end{aligned}
$$

- Similarly, $\mathbb{E}[Y_1] = \mathbb{E}_{p(x)}[Y_1|x, T = 1]$
- Both can be estimated from data
- ATE = $\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{p(x)}[Y_1|x, T = 1] - \mathbb{E}_{p(x)}\mathbb{E}[Y_0|x, T = 0]$

Are we there yet?

- ATE = $\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{p(x)}[Y_1|x, T = 1] - \mathbb{E}_{p(x)}\mathbb{E}[Y_0|x, T = 0]$
- Not quite.
- The issue is that our samples are biased (i.e we can only evaluate $\mathbb{E}_{p(x|T=1)}$ and not $\mathbb{E}_{p(x)}$
- How do we get around this?

Propensity Score

- IPTW (Inverse Probability of Treatment Weighted) Estimator
- **Key Idea:** Form a parametric estimate of $p(T|x)$
- Different factorizations of the joint via the chain rule: $p(x|T=1)p(T=1) = p(x)p(T=1|x)$
- Therefore use $p(x) = p(x|T=1)\frac{p(T=1)}{p(T=1|x)}$ to re-weight samples

What about now?

$$\text{Given: } p(x) = p(x|T=1)\frac{p(T=1)}{p(T=1|x)}$$

$$\mathbb{E}_{p(x|T=1)}[\underbrace{\frac{p(T=1)}{p(T=1|x)}}_{\text{Weighting: } w(x)}\mathbb{E}[Y_1|x,T=1]]$$

$$\mathbb{E}_{p(x)}[\mathbb{E}[Y_1|x,T=1]]$$

- Now, we have a way to estimate the ATE!
- What about ITE?

Method 3: Covariate Adjustment

- How do we estimate the individual treatment effect?
- Fit a model to approximate $f(x,t) \approx \mathbb{E}[Y_t | T = t, x]$
- AKA Response surface modeling
- $I\hat{T}E(x_i) = f(x_i, 1) - f(x_i, 0)$
- $A\hat{T}E = \frac{1}{n} \sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$
- If f is linear, then *ATE* is the parameter that modulates how the outcome behaves as a function of the treatment assignment

Covariate Adjustment: Yay or Nay?

- Model misspecification is a problem
- Allows use of fancier ML models possible for causal inference (at the cost of a less interpretable ATE)
- Can be upgraded with doubly robust estimators

Overview

- So far, we've talked about a very simplistic world with three random variables.
- What if we had many random variables and relationships between them?

1. Introduce structural equation models (causality among random variables)
2. SEMs equivalently written as causal graphs
3. How to estimate causal effects in a causal graph
4. What is a causal graph (I know. . . a bit backward. . . bear with me)
5. Can we identify an effect from a causal graph?

Structural Equation Models (SEMs)

- Method by Rubin to formalize causal influence between multiple random variables
- Lets look at an example:

$$
\begin{aligned}
z &\sim \mathcal{N}(0,1) \\
y &= z + 2 \\
x &= y + z + 12
\end{aligned}
$$

- Collection of stochastic and deterministic relationships between random variables
- Nicely captures the intuition for causality e.g if $y = 5$ then we set $y = 5$ above and that gives us a *new* set of equations between x, z

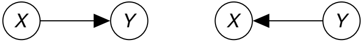
Do-Operator on graphs

- Graphical version of causal inference with SEMs
- We will assume that we have a causal graph G
- The do-operator is a combination of surgery on a graph G with probabilistic inference
- $p_G(Y|do(X=x)$
- Doing surgery on a graph G yields G'.
- G' is a subgraph of G with no edges from $pa(X) \rightarrow X$
- $p_G(Y|do(X=x) = p_{\hat{G}}(Y|X=x)$ involves inference on the resulting subgraph $\hat{G}$

Causal Graphs

- A causal graph is a Bayesian network but a Bayesian network need not be a causal graph
- Why? Because we use domain knowledge and intuition to pre-specify directions of causality

BN & Causal Graphs



- Intuitively: directionality of the edges encodes causal influence and consequently affects the result of causal query
- Formally, the two structures are *I-equivalent* but the result of do-calculus (from class, revisit this in a bit) yields different results

Setting up the graph

- Think hard to make sure you have captured the random variables of interest
- Talk to a domain expert to setup the edges correctly (Remember: no hidden confounders!)

Testing for Identifiability

- Given a causal graph G and a joint distribution over random variables
- Amongst the random variables we care about a particular query we will be asking of the graph
- **Key Question:** Is the causal effect identifiable from my data?
- Identifiable means that we can control for confounding
- Lets assume we want to estimate the effect of T causing Y
- The intuition is that adequate control variables will block paths between T and Y

Back Door Criterion

- Back-door criterion: The observed variables X (features) d-separate all paths between Y (outcome) and T (treatment assignment) that end with an arrow pointing to T
- In References, see document about Front-Door criterion

Procedure

Approximate pseudocode for causal inference

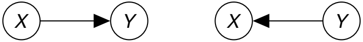
- **Setup graph:**
  - Make sure no unobserved confounders exist
  - Use domain knowledge & common sense to setup graph structure
  - Check if causal effect is identifiable: front-door and back-door criterion
- **Estimate Parameters:** Parameterize CPDs and estimate model parameters from data (might need inference for latent variables)
- **Estimate ITE with do-calculus** (might need inference on intervened graph)

Warning: Not exhaustive but should give you the general idea

Question: Unmeasured Confounders

- **Question:** “This means we can have a hidden factor that influences treatment outcome, as long as it does not influence treatment assignment, am I right?”
- In that example from the slides, the Back-Door criterion applies. Specifically, our post-treatment blood pressure is conditionally independent of our outcome given our covariates (age etc)

Identifying Causal Direction



- Which is the causal direction $X \rightarrow Y$ or $Y \rightarrow X$
- The underlying intuition is that the causal direction has an *easier* distribution to estimate from data (Janzing (2007), Hoyer et. al (2009))

References

- Causality, Judea Pearl (Book)
- Course Notes on Causality from Prof. Cosma Shalizi
- Datastories: Machine Bias with Jeff Larson
- Machine bias risk assessments in criminal sentencing