Data-efficient causal effect estimation

Adith Swaminathan
adswamin@microsoft.com

Joint work with Maggie Makar (MIT) and Emre Kıcıman (MSR AI)

Brown TRIPODS
1.16.2019
1. Improve ML applications using Causal Reasoning

2. Use ML tools to perform Causal Inference
1. Causal Reasoning -> ML

“Use logs collected from interactive systems to evaluate and train new interaction policies”

The data we collect from interactive systems...

Simple pragmatic fixes to address confounding!
Model the propensity of clicks on documents to de-bias training set of learning-to-rank models
Pointers to recent results

“IPW fixes collaborative filtering for recommendations”
[Schnabel et al, ICML’16]

“Similar IPW-like ideas massively improve learning-to-rank for search”
[Joachims et al, WSDM’17 Best Paper]

“Important to reason about variance of IPW for counterfactual learning”
[Swaminathan & Joachims, ICML’15]

“We can do much better than IPW for structured treatments (slates)”
[Swaminathan et al, NIPS’17]

“Self-normalized estimators are better to use in these applications”
[Swaminathan & Joachims, NIPS’15]

“These techniques complement deep learning”
[Joachims et al, ICLR’18]

“IPW fixes collaborative filtering for recommendations”
1. Improve ML applications using Causal Reasoning

2. Use ML tools to perform Causal Inference
2. ML -> Causal Reasoning

“Data efficient treatment effect estimation”

[AAAI’19]

Representation learning + Causal inference = Bias-Variance Trade-off?
Problem Setting

Will my patient’s blood pressure increase if I put her on medication A?

Challenges
- A question of causal nature
- Limited data at test time
Individual Treatment Effect (ITE)

- Estimate the causal effect of an intervention: if $t$ changes, how does the outcome $Y_t$ change?

- Target for estimation: $Y_1 - Y_0$

- Target is unobserved: the fundamental problem of causal inference

$$ITE: \tau(x) = \mathbb{E}_{Y_1 \sim \text{Pr}(Y_1|x)}[Y_1] - \mathbb{E}_{Y_0 \sim \text{Pr}(Y_0|x)}[Y_0]$$
ITE estimation from obs. data

Two functions:

- Adjustment for Confounding
  - Confounders

- Estimation of heterogeneity
  - Effect modifiers
Confounders vs. Effect modifiers

Average treatment effect
Confounders vs. Effect modifiers

![Graph showing the relationship between confounders and outcome](image)

- **t = 0, z = 0 or 1**
- **t = 1, z = 0**
- **t = 1, z = 1**
Data efficient ITE estimation

ITE discovery

Adjustment for Confounding

Estimation of heterogeneity

ITE Prediction
Insight

Leverage the difference between tasks at training and test time to reduce data collection burden at test time.

\[ \{x_i, t_i, y_i\}_{i=1}^N \rightarrow A_b \rightarrow \{\hat{\tau}_i\}_{i=1}^N \]

\[ \{x_i, \hat{\tau}_i\}_{i=1}^N \rightarrow A_d \rightarrow \{\tilde{\tau}_i\}_{i=1}^N \]
Why Trees?
Trees identify the most important axes of heterogeneity
Trees can be traversed till querying ability is exhausted
Different individuals → different queries
Algorithm: DEITEE

Data Efficient Individual Treatment Effect Estimator

\[
\{x_i, t_i, y_i\}_{i=1}^{N} \rightarrow \text{ITE Discovery: Base model} \rightarrow \{\hat{\tau}_i\}_{i=1}^{N}
\]

\[
\{x_i, \hat{\tau}_i\}_{i=1}^{N} \rightarrow \text{ITE Prediction DEITEE model} \rightarrow \{\tilde{\tau}_i\}_{i=1}^{N}
\]
Experiments: Synthetic

- Data: ACIC’17 simulated data ("semi-synthetic")
  \[ N=5k; \ d=58 \]
- Base models: BART and GRF
- Benchmarks: Train BART/GRF with feature regularization

- Evaluation: (1) Accuracy relative to true ITE;
  (2) Number of features queried
DEITEE: Features queried

![Bar chart comparing DEITEE-GRF and DEITEE-Bart](chart.png)
DEITEE doesn’t sacrifice accuracy

![Graph showing median squared error vs. features collected for different models: RT-GRF, DEITEE-Bart, DEITEE-GRF, DEITEE-Oracle, RT-BART. The graph illustrates that DEITEE does not sacrifice accuracy even when features are collected.]
Experiment on real data

What is the effect of mother’s habits on newborn’s health?
1989 MA singleton births (CDC) N=90k; d=77

<table>
<thead>
<tr>
<th>Mother’s habit (treatment)</th>
<th>Mean Absolute Error relative to proxy ITE</th>
<th>Mean number of DEITEE features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenatal care</td>
<td>580.20</td>
<td>BART</td>
</tr>
<tr>
<td>Smoking</td>
<td>587.62</td>
<td>15.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alcohol?</th>
<th>HS Education</th>
<th>Age? Health risks?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Married?</th>
<th>Age?</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
Conclusions

• DEITEE reduces the number of features required to estimate individual causal effects
  - Leverage difference between ITE discovery and ITE prediction

• Ongoing: Careful analysis of distillation error; guarantees on effect modifier discovery

• Need: Good robust method for model selection
Thanks!

Confounding

ITE Discovery

Heterogeneity

ITE Prediction

adswamin@microsoft.com