Classifying Treatment Responders Under Causal Effect Monotonicity

Nathan Kallus
ORIE and Cornell Tech, Cornell University
Heterogeneous Treatment Effect Estimation

<table>
<thead>
<tr>
<th>$X_{\text{Age}}$</th>
<th>$X_{\text{Weight}}$</th>
<th>$X_{\text{BMI}}$</th>
<th>$X_{\text{SysBP}}$</th>
<th>$T$ (Anticoagulant)</th>
<th>$Y$ (Hemorrhage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>106</td>
<td>31</td>
<td></td>
<td>Warfarin</td>
<td>1</td>
</tr>
<tr>
<td>54</td>
<td>89</td>
<td>26</td>
<td></td>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>130</td>
<td>38</td>
<td></td>
<td>None</td>
<td>1</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>

Fit CATE $\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$ to data on $X, T, Y$

E.g.:
Causal Forest (Wager & Athey ’17),
TARNet (Shalit et al. ’17),
...
Often Outcome is **Binary**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Outcome Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give anticoagulant</td>
<td>Hemorrhage?</td>
</tr>
<tr>
<td>Personalized discount</td>
<td>Buy?</td>
</tr>
<tr>
<td>Target job training</td>
<td>Employed in 6 months?</td>
</tr>
<tr>
<td>Homelessness prevention program</td>
<td>Re-enter?</td>
</tr>
<tr>
<td>Recidivism prevention program</td>
<td>Recidivate?</td>
</tr>
<tr>
<td>Support for minority CS students</td>
<td>Drop out?</td>
</tr>
</tbody>
</table>
Often We Want to Predict **Response**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Individual Label of Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>( Y(1) - Y(0) )</td>
</tr>
<tr>
<td>Give anticoagulant</td>
<td>Hemorrhage iff medicated</td>
</tr>
<tr>
<td>Personalized discount</td>
<td>Would buy iff discounted</td>
</tr>
<tr>
<td>Target job training</td>
<td>Would get job iff trained</td>
</tr>
<tr>
<td>Homelessness prevention program</td>
<td>Re-enter iff not targeted</td>
</tr>
<tr>
<td>Recidivism prevention program</td>
<td>Recidivate iff not targeted</td>
</tr>
<tr>
<td>Support for minority CS students</td>
<td>Drop out iff not targeted</td>
</tr>
</tbody>
</table>

Nathan Kallus

Classifying Treatment Responders Under Causal Effect Monotonicity
Classifying Responders: The Problem

- Each unit consists of
  - Features $X$
  - Potential outcomes $Y(1), Y(0) \in \{0, 1\}$
- “Non-responder” has $Y(0) = Y(1)$
  - Would’ve bought (or, not bought) \textit{regardless} of discount
  - Would’ve hemorrhaged (or, not) \textit{regardless} of anticoagulant
- “Responder” has $Y(1) = 1 > 0 = Y(0)$
  - Would’ve bought \textit{if and only if} offered discount
- $R = \mathbb{I}[Y(1) > Y(0)]$
- Ground truth \underline{NOT} observed in $X, T, Y$ data
- Want classifier $f : \mathcal{X} \rightarrow \{0, 1\}$ with small loss

\[
L_\theta(f) = \theta \mathbb{P}(\text{false positive}) + (1 - \theta) \mathbb{P}(\text{false negative})
= \theta \mathbb{P}(f(X) = 1, R = 0)
+ (1 - \theta) \mathbb{P}(f(X) = 0, R = 1).
\]
Monotonicity

- **Monotone treatment response assumption:**

\[ Y(1) \geq Y(0) \]

- Discount never causes a would-be buyer to *not* buy
- Job training never causes someone to *not* get employed?

\[ R = Y(1) - Y(0) \in \{0, 1\} \]

So,
\[ P(R = 1 | X) = \tau(X) = E[Y(1) - Y(0) | X] \]

\[ f(X) = I[\tau(X) \geq \theta] \text{ minimizes } L_{\theta}(f) \]

Can take plug-in approach using any CATE estimator \( \hat{\tau} \)

**Question:** any value to a direct classification approach?
Monotonicity

- **Monotone treatment response assumption:**

  \[ Y(1) \geq Y(0) \]

- Discount never causes a would-be buyer to *not* buy
- Job training never causes someone to *not* get employed?

- Under monotonicity, \( R = Y(1) - Y(0) \in \{0, 1\} \)
- So,

  \[
  \mathbb{P}(R = 1 \mid X) = \tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]
  \]

- \( f(X) = \mathbb{I}[\tau(X) \geq \theta] \) minimizes \( L_\theta(f) \)
- Can take plug-in approach using any CATE estimator \( \hat{\tau} \)
- Question: any value to a direct classification approach?
For simplicity, consider completely randomized data with \( \mathbb{P}(T = 1) = 0.5 \).

Let \( Z = \mathbb{I}[Y = T] \) (observable!)

- \( R = 1 \implies Z = 1 \)
- \( R = 0 \implies Z \sim \text{Bernoulli}(0.5) \)

\( Z \) is like a corrupted observation of \( R \)

- Seeing \( Z = 0 \) is more informative about \( R \)

Using \( Z \) as a surrogate label for \( R \) leads to new direct approaches to the classification problem

- Two instantiations of this are RespSVM, RespNet
Empirical Results: Synthetic

The true label $R$

The observable label $Z$

$T = +1$

$T = 0$
Empirical Results: Synthetic

Linear responder classification boundary

Spherical responder classification boundary

$\text{Nathan Kallus}$

Classifying Treatment Responders Under Causal Effect Monotonicity
Empirical Results: Census Data

- Predict whether the sex-at-birth of mother’s first two kids being the same influences her decision to have a third
  - Follows data construction by Angrist & Evans ’96
  - Covariates: ethnicity of mother and father; their ages at marriage, at census, at 1st kid, and at 2nd kid, year of marriage, and education level

<table>
<thead>
<tr>
<th>Method</th>
<th>$L_\theta$ (in 0.01)</th>
<th>% 1st</th>
<th>% 2nd</th>
<th>% 3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>RespSVM lin</td>
<td>49 ± 2.7</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RespLR-gen</td>
<td>57 ± 2.4</td>
<td></td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>RespLR-disc</td>
<td>58 ± 2.3</td>
<td></td>
<td></td>
<td>2%</td>
</tr>
<tr>
<td>LR</td>
<td>58 ± 2.3</td>
<td></td>
<td></td>
<td>92%</td>
</tr>
<tr>
<td>RF</td>
<td>58 ± 2.3</td>
<td></td>
<td></td>
<td>6%</td>
</tr>
</tbody>
</table>
Thank you!

Poster: Today 6:30pm @ Pacific Ballroom #74