Machine Learning for Healthcare: Causal inference

David Sontag
Clinical Machine Learning Group
MIT
Does gastric bypass surgery prevent onset of diabetes?

- Yesterday we used machine learning for early detection of Type 2 diabetes
- Health system doesn’t want to know how to predict diabetes – they want to know how to prevent it
- Gastric bypass surgery is the highest negative weight (9th most predictive feature)
  - Does this mean it would be a good intervention?
What is the likelihood this patient, with breast cancer, will survive 5 years?

- Such predictive models widely used to stage patients. Should we initiate treatment? How aggressive?
- What could go wrong if we trained to predict survival, and then used to guide patient care?

A long survival time may be because of treatment!
What treatment should we give this patient?

- People respond differently to treatment
- Goal: use data from other patients and their journeys to guide future treatment decisions
- What could go wrong if we trained to predict (past) treatment decisions?

  “David”  ➔  Treatment A
  “John”  ➔  Treatment B
  “Juana”  ➔  Treatment A

Best this can do is match current medical practice!
Does smoking cause lung cancer?

- Doing a randomized control trial is unethical
- Could we simply answer this question by comparing \( \text{Pr}(\text{lung cancer} | \text{smoker}) \) vs \( \text{Pr}(\text{lung cancer} | \text{nonsmoker}) \)?
- No! Answering such questions from observational data is difficult because of *confounding*
To properly answer, need to formulate as causal questions:

- **Patient,** $X$
- **Intervention,** $T$
- **Outcome,** $Y$

(including all confounding factors)

(e.g. medication, procedure)

---

*High dimensional*  
*Observational data*
Potential Outcomes Framework (Rubin-Neyman Causal Model)

• Each unit (individual) $x_i$ has two potential outcomes:
  – $Y_0(x_i)$ is the potential outcome had the unit not been treated: “control outcome”
  – $Y_1(x_i)$ is the potential outcome had the unit been treated: “treated outcome”

• Conditional average treatment effect for unit $i$:
  $CATE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$

• Average Treatment Effect:
  $ATE := \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{x \sim p(x)}[CATE(x)]$
Potential Outcomes Framework
(Rubin-Neyman Causal Model)

• Each unit (individual) $x_i$ has two potential outcomes:
  – $Y_0(x_i)$ is the potential outcome had the unit not been treated: “control outcome”
  – $Y_1(x_i)$ is the potential outcome had the unit been treated: “treated outcome”

• Observed factual outcome:
  $$y_i = t_i Y_1(x_i) + (1 - t_i)Y_0(x_i)$$

• Unobserved counterfactual outcome:
  $$y_{i}^{CF} = (1 - t_i)Y_1(x_i) + t_i Y_0(x_i)$$
“The fundamental problem of causal inference”

We only ever observe one of the two outcomes
Example – Blood pressure and age

\[ y = \text{blood\_pres}. \]

\[ x = \text{age} \]

\[ Y_1(x) \]

\[ Y_0(x) \]
Blood pressure and age

\[ y = \text{blood\_pres.} \]

\[ y = x + \gamma(x) \]

\[ CATE(x) \]

\[ Y_1(x) \]

\[ Y_0(x) \]

\[ x = \text{age} \]
Blood pressure and age

\[ y = \text{blood\_pres}. \]

\[ x = \text{age} \]

\[ Y_1(x) \]

\[ Y_0(x) \]

\[ ATE \]
Blood pressure and age

\[ y = \text{blood\_pres}. \]

\[ Y_1(x) \]

\[ Y_0(x) \]

- Red: Treated
- Blue: Control

\[ x = \text{age} \]
Blood pressure and age

\[ y = \text{blood\_pres}. \]

\[ x = \text{age} \]

\[ Y_1(x) \]

\[ Y_0(x) \]

- Treated
- Control
- Counterfactual treated
- Counterfactual control
<table>
<thead>
<tr>
<th>(age, gender, exercise, treatment)</th>
<th>Observed sugar levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>(45, F, 0, A)</td>
<td>6</td>
</tr>
<tr>
<td>(45, F, 1, B)</td>
<td>6.5</td>
</tr>
<tr>
<td>(55, M, 0, A)</td>
<td>7</td>
</tr>
<tr>
<td>(55, M, 1, B)</td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 0, B)</td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 1, A)</td>
<td>7.5</td>
</tr>
<tr>
<td>(75, M, 0, B)</td>
<td>9</td>
</tr>
<tr>
<td>(75, M, 1, A)</td>
<td>8</td>
</tr>
</tbody>
</table>

(Example from Uri Shalit)
<table>
<thead>
<tr>
<th>(age, gender, exercise)</th>
<th></th>
<th></th>
<th>Observed sugar levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>(45, F, 0)</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>(45, F, 1)</td>
<td></td>
<td></td>
<td>6.5</td>
</tr>
<tr>
<td>(55, M, 0)</td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>(55, M, 1)</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 0)</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 1)</td>
<td></td>
<td></td>
<td>7.5</td>
</tr>
<tr>
<td>(75, M, 0)</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>(75, M, 1)</td>
<td></td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

*(Example from Uri Shalit)*
<table>
<thead>
<tr>
<th>(age, gender, exercise)</th>
<th>$Y_0$: Sugar levels <em>had they received</em> medication A</th>
<th>$Y_1$: Sugar levels <em>had they received</em> medication B</th>
<th>Observed sugar levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>(45, F, 0)</td>
<td>6</td>
<td>5.5</td>
<td>6</td>
</tr>
<tr>
<td>(45, F, 1)</td>
<td>7</td>
<td>6.5</td>
<td>6.5</td>
</tr>
<tr>
<td>(55, M, 0)</td>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>(55, M, 1)</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 0)</td>
<td>8.5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>(65,F, 1)</td>
<td>7.5</td>
<td>7</td>
<td>7.5</td>
</tr>
<tr>
<td>(75,M, 0)</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>(75,M, 1)</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

(Example from Uri Shalit)
<table>
<thead>
<tr>
<th>(age, gender, exercise)</th>
<th>Sugar levels</th>
<th>Sugar levels</th>
<th>Observed sugar levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>had they received</em> medication A</td>
<td><em>had they received</em> medication B</td>
<td></td>
</tr>
<tr>
<td>(45, F, 0)</td>
<td>6</td>
<td>5.5</td>
<td>6</td>
</tr>
<tr>
<td>(45, F, 1)</td>
<td>7</td>
<td><strong>6.5</strong></td>
<td><strong>6.5</strong></td>
</tr>
<tr>
<td>(55, M, 0)</td>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>(55, M, 1)</td>
<td>9</td>
<td><strong>8</strong></td>
<td><strong>8</strong></td>
</tr>
<tr>
<td>(65, F, 0)</td>
<td>8.5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 1)</td>
<td><strong>7.5</strong></td>
<td>7</td>
<td><strong>7.5</strong></td>
</tr>
<tr>
<td>(75, M, 0)</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>(75, M, 1)</td>
<td><strong>8</strong></td>
<td>7</td>
<td><strong>8</strong></td>
</tr>
</tbody>
</table>

\[
\text{mean}(\text{sugar} | \text{medication B}) - \text{mean}(\text{sugar} | \text{medication A}) = ?
\]

\[
\text{mean}(\text{sugar} | \text{had they received B}) - \text{mean}(\text{sugar} | \text{had they received A}) = ?
\]

(Example from Uri Shalit)
<table>
<thead>
<tr>
<th>(age, gender, exercise)</th>
<th>Sugar levels had they received medication A</th>
<th>Sugar levels had they received medication B</th>
<th>Observed sugar levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>(45, F, 0)</td>
<td>6</td>
<td>5.5</td>
<td>6</td>
</tr>
<tr>
<td>(45, F, 1)</td>
<td>7</td>
<td><strong>6.5</strong></td>
<td>6.5</td>
</tr>
<tr>
<td>(55, M, 0)</td>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>(55, M, 1)</td>
<td>9</td>
<td><strong>8</strong></td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 0)</td>
<td>8.5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>(65, F, 1)</td>
<td><strong>7.5</strong></td>
<td>7</td>
<td>7.5</td>
</tr>
<tr>
<td>(75, M, 0)</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>(75, M, 1)</td>
<td><strong>8</strong></td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

mean(sugar | medication B) – mean(sugar | medication A) = 7.875 - 7.125 = 0.75

mean(sugar | had they received B) – mean(sugar | had they received A) = 7.125 - 7.875 = -0.75

(Example from Uri Shalit)
Typical assumption – no unmeasured confounders

\( Y_0, Y_1 \): potential outcomes for control and treated

\( x \): unit covariates (features)

\( T \): treatment assignment

We assume:

\[
(Y_0, Y_1) \perp T \mid x
\]

The potential outcomes are independent of treatment assignment, conditioned on covariates \( x \)
Typical assumption – common support

$Y_0, Y_1$: potential outcomes for control and treated
$x$: unit covariates (features)
$T$: treatment assignment

We assume:

$$p(T = t | X = x) > 0 \forall t, x$$
Two common approaches for counterfactual inference

Covariate adjustment
Propensity scores
Covariate adjustment (parametric g-formula)

• Explicitly model the relationship between treatment, confounders, and outcome

• If no unmeasured confounders, the expected causal effect of $T$ on $Y$ (given $x$) is given by:
  \[\text{CATE}(x) = \mathbb{E}[Y|T = 1, x] - \mathbb{E}[Y|T = 0, x]\]

• Fit a model $f(x, t) \approx \mathbb{E}[Y|T = t, x]$

\[\text{CATE}(x_i) = f(x_i, 1) - f(x_i, 0)\]
Covariate adjustment (parametric g-formula)

• Explicitly model the relationship between treatment, confounders, and outcome

• If no unmeasured confounders, the expected causal effect of $T$ on $Y$ (given $x$) is given by: 
  \[
  \text{CATE}(x) = \mathbb{E}[Y|T = 1, x] - \mathbb{E}[Y|T = 0, x]
  \]

• Fit a model $f(x, t) \approx \mathbb{E}[Y|T = t, x]$

  \[
  \overline{ATE} = \frac{1}{n} \sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)
  \]
\[ x_1 + x_2 + \cdots + x_d + T \]

Regression model

\[ f(x, T) \]

Outcome

Covariates (Features)

\[ x_1, x_2, \ldots, x_d, T \]
Parameter of interest

\[ f(x, T) \]

Nuisance Parameters

\[ x_1, x_2, \ldots, x_d \]

Regression model

Outcome

\[ y \]
Example of how covariate adjustment fails when there is no overlap

\[ y = \text{blood-pres}. \]

\[ Y_1(x) \]

\[ Y_0(x) \]

- Treated
- Control

\[ x = \text{age} \]
Covariate adjustment with linear models

• Assume that:

\[
Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t
\]
\[
\mathbb{E}[\epsilon_t] = 0
\]

• Then:

\[
CATE(x) = \mathbb{E}[Y_1(x) - Y_0(x)]
\]
Covariate adjustment with linear models

• Assume that:

\[ Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t \]
\[ \mathbb{E}[\epsilon_t] = 0 \]

• Then:

\[ CATE(x) := \mathbb{E}[Y_1(x) - Y_0(x)] = \mathbb{E}[(\beta x + \gamma + \epsilon_1) - (\beta x + \epsilon_0)] = \gamma \]

\[ ATE := \mathbb{E}_{p(x)}[CATE(x)] = \gamma \]
Covariate adjustment with linear models

- Assume that:

\[ Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t \]
\[ \mathbb{E}[\epsilon_t] = 0 \]

**ATE**: \[ \mathbb{E}_{p(x)}[CATE(x)] = \gamma \]

- For causal inference, need to estimate \( \gamma \) well, not \( Y_t(x) \) - **Identification, not prediction**

- **Major difference between ML and statistics**
What happens if true model is not linear?

• True data generating process, \( x \in \mathbb{R} \):
  \[ Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2 \]
  \[ ATE = \mathbb{E}[Y_1 - Y_0] = \gamma \]

• Hypothesized model:
  \[ \hat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t \]

\[
\hat{\gamma} = \gamma + \delta \left( \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]} \right)
\]

Depending on \( \delta \), can be made to be arbitrarily large or small!
Covariate adjustment with non-linear models

- Random forests and Bayesian trees

- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012)

- Neural networks
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016),
  Lopez-Paz et al. (2016)
Example: Gaussian processes

Separate treated and control models

Joint treated and control model

Figures: Vincent Dorie & Jennifer Hill
Example: Neural networks

Two common approaches for counterfactual inference

Covariate adjustment
Propensity scores
Propensity scores

• Tool for estimating ATE

• Imagine that we had data from a randomized control trial (RCT). Then we could simply estimate the ATE using:

\[
\frac{1}{n_1} \sum_{i \text{ s.t.} T_i=1} Y_i - \frac{1}{n_0} \sum_{i \text{ s.t.} T_i=0} Y_i
\]

• Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples
Inverse propensity score re-weighting

\[ p(x|t = 0) \neq p(x|t = 1) \cdot w_1(x) \]

- reweighted control
- reweighted treated

\[ x_1 = \text{age} \]

\[ x_2 = \text{Charlson comorbidity index} \]

- Treated
- Control
Propensity score

- Propensity score: $p(T = 1|x)$, using machine learning tools
- Samples re-weighted by the inverse propensity score of the treatment they received
Propensity scores – algorithm

*Inverse probability of treatment weighted estimator*

How to calculate ATE with propensity score for sample \((x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\)

1. Use any ML method to estimate \(\hat{p}(T = t|x)\)

2. \[
A\hat{TE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i = 1} y_i \hat{p}(t_i = 1|x_i) - \frac{1}{n} \sum_{i \text{ s.t. } t_i = 0} y_i \hat{p}(t_i = 0|x_i)
\]
Propensity scores – algorithm

*Inverse probability of treatment weighted estimator*

How to calculate ATE with propensity score for sample \((x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\)

1. Randomized trial \(p(T = t|x) = 0.5\)

2. \[
\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i = 1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i = 0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}
\]
Propensity scores – algorithm

*Inverse probability of treatment weighted estimator*

How to calculate ATE with propensity score for sample \((x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\)

1. Randomized trial \(p(T = t|x) = 0.5\)

2. \[
\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i = 1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i = 0} \frac{y_i}{0.5} =
\]
Propensity scores – algorithm

*Inverse probability of treatment weighted estimator*

How to calculate ATE with propensity score
for sample \((x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\)

1. Randomized trial \(p = 0.5\)

2. \[
\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i = 1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i = 0} \frac{y_i}{0.5} = \\
\frac{2}{n} \sum_{i \text{ s.t. } t_i = 1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i = 0} y_i
\]
Propensity scores – algorithm

*Inverse probability of treatment weighted estimator*

How to calculate ATE with propensity score for sample \((x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\)

1. Randomized trial \(p = 0.5\)

2. \[ \hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} y_i \frac{1}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} y_i \frac{1}{0.5} = \]

\[ \frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i \]

Sum over \(\sim \frac{n}{2}\) terms
Propensity scores - derivation

• Recall average treatment effect:
  \[ \mathbb{E}_{x \sim p(x)} \left[ \mathbb{E} [ Y_1 | x, T = 1 ] - \mathbb{E} [ Y_0 | x, T = 0 ] \right] \]

• We only have samples for:
  \[ \mathbb{E}_{x \sim p(x | T=1)} \left[ \mathbb{E} [ Y_1 | x, T = 1 ] \right] \]
  \[ \mathbb{E}_{x \sim p(x | T=0)} \left[ \mathbb{E} [ Y_0 | x, T = 0 ] \right] \]
Propensity scores - derivation

• We only have samples for:

\[
\mathbb{E}_{x \sim p(x|T=1)} \left[ \mathbb{E} [Y_1|x, T = 1] \right] \\
\mathbb{E}_{x \sim p(x|T=0)} \left[ \mathbb{E} [Y_0|x, T = 0] \right]
\]
Propensity scores - derivation

• We only have samples for:

\[
\begin{align*}
\mathbb{E}_{x \sim p(x|T=1)} & \left[ \mathbb{E} [Y_1|x, T = 1] \right] \\
\mathbb{E}_{x \sim p(x|T=0)} & \left[ \mathbb{E} [Y_0|x, T = 0] \right]
\end{align*}
\]

• We need to turn \( p(x|T = 1) \) into \( p(x) \):

\[
p(x|T = 1) \cdot \ ? \ = p(x)
\]
Propensity scores - derivation

• We only have samples for:
  \[ \mathbb{E}_{x \sim p(x | T=1)} [ \mathbb{E} [Y_1 | x, T = 1] ] \]
  \[ \mathbb{E}_{x \sim p(x | T=0)} [ \mathbb{E} [Y_0 | x, T = 0] ] \]

• We need to turn \( p(x | T = 1) \) into \( p(x) \):
  \[
p(x | T = 1) \cdot \frac{p(T = 1)}{p(T = 1 | x)} = p(x)
  \]
  \[ \text{Propensity score} \]
Propensity scores - derivation

• We only have samples for:

\[
\mathbb{E}_{x \sim p(x|T=1)} \left[ \mathbb{E} [Y_1 | x, T = 1] \right] \\
\mathbb{E}_{x \sim p(x|T=0)} \left[ \mathbb{E} [Y_0 | x, T = 0] \right]
\]

• We need to turn \( p(x|T = 0) \) into \( p(x) \):

\[
p(x|T = 0) \cdot \frac{p(T = 0)}{p(T = 0|x)} = p(x)
\]

\textit{Propensity score}
• We want: \( \mathbb{E}_{x \sim p(x)}[Y_1(x)] \)

• We know that:

\[
p(x|T=1) \cdot \frac{p(T=1)}{p(T=1|x)} = p(x)
\]

• Thus:

\[
\mathbb{E}_{x \sim p(x|T=1)} \left[ \frac{p(T=1)}{p(T=1|x)} Y_1(x) \right] = \mathbb{E}_{x \sim p(x)}[Y_1(x)]
\]

• We can approximate this empirically as:

\[
\frac{1}{n_1} \sum_{i \text{ s.t. } t_i=1} \left[ \frac{n_1/n}{\hat{p}(t_i=1 | x_i)} y_i \right] = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i=1 | x_i)}
\]

(similarly for \( t_i=0 \))
Problems with inverse propensity weighting (IPW)

• Need to estimate propensity score (problem in all propensity score methods)

• If there’s not much overlap, propensity scores become non-informative and easily mis-calibrated

• Weighting by inverse can create large variance and large errors for small propensity scores
  – Exacerbated when more than two treatments
Many more ideas and methods

• Natural experiments & regression discontinuity
• Instrumental variables
Many more ideas and methods – Natural experiments

- Does stress during pregnancy affect later child development?
- Confounding: genetic, mother personality, economic factors...
- Natural experiment: the Cuban missile crisis of October 1962. Many people were afraid a nuclear war is about to break out.
- Compare children who were in utero during the crisis with children from immediately before and after
Many more ideas and methods – Instrumental variables

• Informally: a variable which affects treatment assignment but not the outcome
• Example: are private schools better than public schools?
• Confounding: different student population, different teacher population
• Can’t force people which school to go to
Many more ideas and methods – Instrumental variables

• Informally: a variable which affects treatment assignment but not the outcome
• Example: are private schools better than public schools?
• Can’t force people which school to go to
• Can randomly give out vouchers to some children, giving them an opportunity to attend private schools
• The voucher assignment is the instrumental variable
Summary

• Two approaches to use machine learning for causal inference:
  1. Predict outcome given features and treatment, then use resulting model to impute counterfactuals \((\text{covariate adjustment})\)
  2. Predict treatment using features \((\text{propensity score})\), then use to reweight outcome or stratify the data

• Results only valid if there’s no unobserved confounding – should do a sensitivity analysis