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 \( \Pr(\text{lung cancer} \mid \text{smoker}) \) vs \( \Pr(\text{lung cancer} \mid \text{nonsmoker}) \)?
• No! Answering such questions from observational data is difficult because of *confounding*
To properly answer, need to formulate as *causal* questions:

Patient, $X$ (including all confounding factors) → Intervention, $T$ (e.g. medication, procedure) → Outcome, $Y$

*High dimensional* observational data
Outline for lecture

• How to recognize a causal inference problem

• Potential outcomes framework
  – Average treatment effect (ATE)
  – Individual treatment effect (ITE)

• Algorithms for estimating ATE and ITE

• When a feature’s weight has a correct causal interpretation
Example 1
Precision medicine: Individualized Treatment Effect (ITE)
Which treatment is best for me?

- Which anti-hypertensive treatment?
  - Calcium channel blocker (A)
  - ACE inhibitor (B)

- Current situation:
  - Clinical trials
  - Doctor’s knowledge & intuition

- Use datasets of patients and their histories

- Blood pressure = 150/95
- WBC count = 6*10⁹/L
- Temperature = 98°F
- HbA1c = 6.6%
- Thickness of heart artery plaque = 3mm
- Weight = 65kg
Which treatment is best for me?

• Which anti-hypertensive treatment?
  – Calcium channel blocker (A)
  – ACE inhibitor (B)

• Future blood pressure: treatment A vs. B

• Individualized Treatment Effect (ITE) also called Conditional Average Treatment Effect (CATE)
Which treatment is best for me?

• Which anti-hypertensive treatment?
  – Calcium channel blocker (A)
  – ACE inhibitor (B)

• Potential confounder: maybe rich patients got medication A more often, and poor patients got medication B more often
**Example 2**

Screening for prostate cancer: Average Treatment Effect (ATE)

---

**Annals of Internal Medicine**

**SCREENING FOR PROSTATE CANCER**

**CLINICAL SUMMARY OF U.S. PREVENTIVE SERVICES TASK FORCE RECOMMENDATION**

<table>
<thead>
<tr>
<th>Population</th>
<th>Adult Males</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendation</td>
<td>Do not use prostate-specific antigen (PSA)–based screening for prostate cancer.</td>
</tr>
<tr>
<td></td>
<td>Grade: D</td>
</tr>
</tbody>
</table>

**Balance of Harms and Benefits**

- The reduction in prostate cancer mortality 10 to 14 years after PSA-based screening is, at most, very small, even for men in the optimal age range of 55 to 69 years.
- The harms of screening include pain, fever, bleeding, infection, and transient urinary difficulties associated with prostate biopsy, psychological harm of false-positive test results, and overdiagnosis.
- Harms of treatment include erectile dysfunction, urinary incontinence, bowel dysfunction, and a small risk for premature death. Because of the current inability to reliably distinguish tumors that will remain indolent from those destined to be lethal, many men are being subjected to the harms of treatment for prostate cancer that will never become symptomatic. The benefits of PSA-based screening for prostate cancer do not outweigh the harms.

**Other Relevant USPSTF Recommendations**

- Recommendations on screening for other types of cancer can be found at www.uspreventiveservicestaskforce.org.
Should men over the age of 50 have a prostate-specific antigen (PSA) test?

• Screening does not reliably distinguish lethal tumors from those that will remain indolent
• Harms of screening include pain, fever, overdiagnosis
• Should doctors continue to screen?
• **Average Treatment Effect (ATE)**
• Potential *confounder*: Maybe those who are screened are more likely to have prostate cancer?
Observational studies

A major challenge in causal inference from observational studies is how to control or adjust for the confounding factors.
Potential Outcomes Framework
(Rubin-Neyman Causal Model)

• Each unit $x_i$ 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”

• Individual Treatment Effect for unit $i$:
  
  $ITE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1 | x_i)}[Y_1 | x_i] - \mathbb{E}_{Y_0 \sim p(Y_0 | x_i)}[Y_0 | x_i]$

• Average Treatment Effect:

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

• Each unit $x_i$ 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:
  $$y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$
Example – Blood pressure and age

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

\[ Y_1(x) \]
\[ Y_0(x) \]
\[ y = \text{blood_pres}. \]

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

\[ \begin{align*}
  Y_1(x) \\
  Y_0(x)
\end{align*} \]

\[ x = \text{age} \]

\[ \text{ATE} \]
Blood pressure and age

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

\[ Y_1(x) \]

\[ Y_0(x) \]

- Treated
- Control

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

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

\[ x = \text{age} \]
“The fundamental problem of causal inference”

We only ever observe one of the two outcomes
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 – no unmeasured confounders

$Y_0, Y_1$: potential outcomes for control and treated

$x$: unit covariates (features)

$T$: treatment assignment

We assume:

$$(Y_0, Y_1) \perp T \mid x$$

Ignorability
Ignorability

Potential outcomes

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

\[ \text{age, gender, weight, diet, heart rate at rest,...} \]

\[ x \]

\[ T \]

\[ (Y_0, Y_1) \perp T \mid x \]
No Ignorability

age, gender, weight, diet, heart rate at rest, ...

\( x \)

\( T \)

\( h \)

\( (Y_0, Y_1) \perp T \mid x \)
Typical assumption – common support

$Y_0, Y_1$: potential outcomes for control and treated

$x$: unit covariates (features)

$T$: treatment assignment

We assume:

$$p(T = t | X = x) > 0 \ \forall t, x$$
Framing the question

• Now that we have the formalism, let’s return to the questions from the beginning of this lecture:
  1. Where could we go to for data to answer these questions?
  2. What should $X$, $T$, and $Y$ be to satisfy ignorability?
  3. What is the specific causal inference question that we are interested in?
  4. Are you worried about common support?
Outline for lecture

• How to recognize a causal inference problem
• Potential outcomes framework
  – Average treatment effect (ATE)
  – Individual treatment effect (ITE)
• Algorithms for estimating ATE and ITE
• When a feature’s weight has a correct causal interpretation
Tools of the trade

Matching
Covariate adjustment
Propensity score
Matching

• Find each unit’s long-lost counterfactual identical twin, check up on his outcome
Matching

- Find each unit’s long-lost counterfactual identical twin, check up on his outcome

*Obama, had he gone to law school*  *Obama, had he gone to business school*
Matching

• Find each unit’s long-lost counterfactual identical twin, check up on his outcome
• Used for estimating both ATE and ITE
Match to nearest neighbor from opposite group

- Treated
- Control

Charleson comorbidity index

Age
Match to nearest neighbor from opposite group

Charleson comorbidity index

Treated

Control

Age
1-NN Matching

- Let $d(\cdot, \cdot)$ be a metric between $x$'s
- For each $i$, define $j(i) = \arg\min_{j \text{ s.t. } t_j \neq t_i} d(x_j, x_i)$
  - $j(i)$ is the nearest counterfactual neighbor of $i$
- $t_i = 1$, unit $i$ is treated:
  $$\hat{ITE}(x_i) = y_i - y_{j(i)}$$
- $t_i = 0$, unit $i$ is control:
  $$\hat{ITE}(x_i) = y_{j(i)} - y_i$$
1-NN Matching

• Let $d(\cdot, \cdot)$ be a metric between $x$'s

• For each $i$, define $j(i) = \arg\min_{j \text{ s.t. } t_j \neq t_i} d(x_j, x_i)$
  
  $j(i)$ is the nearest counterfactual neighbor of $i$

• $\hat{ITE}(x_i) = (2t_i - 1)(y_i - y_{j(i)})$

• $\overline{ATE} = \frac{1}{n} \sum_{i=1}^{n} \hat{ITE}(x_i)$
Matching

• Interpretable, especially in small-sample regime
• Nonparametric
• Heavily reliant on the underlying metric
• Could be misled by features which don’t affect the outcome
Tools of the trade

Matching
Covariate adjustment
Propensity score
Covariate adjustment

- Explicitly model the relationship between treatment, confounders, and outcome
- Also called “Response Surface Modeling”
- Used for both ITE and ATE
- A regression problem
Regression model

Outcome

Covariates (Features)

\( x_1 \)

\( x_2 \)

\( \vdots \)

\( x_d \)

\( T \)

\( f(x,T) \)
Nuisance Parameters

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

Regression model

\[ f(x, T) \]

Outcome

\[ y \]

Parameter of interest

\[ T \]
Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of $T$ on $Y$:
  \[ \mathbb{E}_{x \sim p(x)} \left[ \mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right] \]
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

\[
\overline{ATE} = \frac{1}{n} \sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)
\]
Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of $T$ on $Y$:
  \[
  \mathbb{E}_{x \sim p(x)} \left[ \mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right]
  \]
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

\[
\widehat{ITE}(x_i) = f(x_i, 1) - f(x_i, 0)
\]
\[ y = \text{blood}_{\text{pres}}. \]

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

\[ Y_1(x) \]

\[ Y_0(x) \]

Treated

Control

\[ x = \text{age} \]
$y = \text{blood\_pres.}$

$\begin{align*}
\text{Tre} & \text{ated} \\
\text{Control} \\
\text{Counterfactual treated} \\
\text{Counterfactual control}
\end{align*}$
Example of how covariate adjustment fails when there is no overlap.

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

\[ Y_1(x) \]

\[ Y_0(x) \]

Treated

Control

\[ x = \text{age} \]
Tools of the trade

Matching
Covariate adjustment
Propensity score
Propensity score

• Tool for estimating ATE
• Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples, similar to importance sampling
Inverse propensity score re-weighting

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

\( x_2 = \) Charlson comorbidity index

\( x_1 = \) age

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.
How to obtain ATE with propensity score
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. 

\[
\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i = 1} \frac{y_i}{\hat{p}(t_i = 1 | x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i = 0} \frac{y_i}{\hat{p}(t_i = 0 | x_i)}
\]
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. \[
\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1}^n \frac{y_i}{\hat{p}(t_i = 1 | x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0}^n \frac{y_i}{\hat{p}(t_i = 0 | x_i)}
\]
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. $\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} y_i \frac{1}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} y_i \frac{1}{0.5}$ = 
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. \[ \hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} = \]

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

Sum over \(\sim \frac{n}{2}\) terms
Outline for lecture

• How to recognize a causal inference problem
• Potential outcomes framework
  – Average treatment effect (ATE)
  – Individual treatment effect (ITE)
• Algorithms for estimating ATE and ITE
• When a feature’s weight has a correct causal interpretation
Linear model

• Assume that:

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

• Then:

\[ ITE(x) := \mathbb{E}[Y_1(x) - Y_0(x)] = \]
Linear model

- Assume that:

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

- Then:

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

\[ ATE := \gamma \]
Linear model

• Assume that:

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

\[ ATE = \mathbb{E}[Y_1(x) - Y_0(x)] = \gamma \]

• We care about \( \gamma \), not about \( Y_t(x) \)

Identification, not prediction
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:
  \[
  \hat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t
  \]

\[
\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}
\]
Better approach is to learn a non-linear model

• Random forests and Bayesian trees

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

• Neural nets
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)
Summary

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

• Causal graphs important for thinking through whether problem is setup appropriately and whether assumptions hold