Causality and Propensity Score Methods
PyData Meetup Berlin, April 19th, 2017
Florian Wilhelm
About me

• Studies of mathematics and PhD at the Karlsruhe Institute of Technology about “Parallel Preconditioners for an Ocean Model in Climate Simulations”,

• Data scientist at Blue Yonder, leading provider of machine learning solutions for retail,

• Data scientist at inovex, an IT project house with focus on digital transformation. Currently working in the Data Team at mobile.de for more than one year.

@FlorianWilhelm
florianwilhelm.info
Outline

1. Motivation
2. Theory
3. Examples
4. Q&A
Correlation vs. causality

“Correlation trumps causation“

“...society will need to shed some of its obsession for causality in exchange for simple correlation: not knowing why but only what.”
Where does causality matter?

Commercial features at mobile.de

Advertisement campaigns with coupons
Let $y_{1i}$ be the target if the $i$-th sample had the do-feature set to 1, and let $y_{0i}$ be the target if the do-feature was 0.

The causal effect is the comparison of $y_{1i}$ with $y_{0i}$, e.g. $\frac{y_{1i}}{y_{0i}}$ or $y_{1i} - y_{0i}$, average causal effect is $E(y_1) - E(y_0)$.

$p(y|do(z))$ denotes the "causal effect" of $Z$ on $Y$, i.e. the distribution of $Y$ after setting variable $Z$ to a constant $Z = z$ by external intervention.
Fundamental problem

• A causal claim is a statement about what did not happen (counterfactuals, potential outcomes)
• Individual causal effects cannot be measured (fundamental problem)
• Correlation is not causation

Source: https://xkcd.com/552/
Clinical trials

**Randomized trial:**
- treatment assignment $Z$ is random thus independent of features $X$ and potential outcomes $Y$
- $Z$ is *controllable*
- gold standard

**Observational trial:**
- treatment assignment $Z$ depends on $X$
- $Z$ is not *controllable*
- sometimes necessary for e.g. ethical reasons
- causal danger zone
Let $X$ denote the covariates, $Y_0, Y_1$ the potential outcomes for treated $Z = 1$ and control $Z = 0$ units.

Treatment assignment $Z$ is strongly ignorable given $X$ if

$$(Y_0, Y_1) \perp Z \mid X \text{ and } 0 < p(Z = 1|x) < 1.$$ 

Using this assumption for causal inference is equivalent to $X$ being admissible, i.e.

$$p(y \mid \text{do}(z)) = \sum_x p(y \mid x, z)p(x).$$
Using machine learning to estimate individual causal effect:

1. train a model with covariates $X$ and $Z$ as feature and $Y$ as target,
2. predict for a given $x$ the response $\hat{y}_1$ with $Z = 1$ and $\hat{y}_0$ with $Z = 0$,
3. calculate the effect with $\hat{y}_1 - \hat{y}_0$ or $\frac{\hat{y}_1}{\hat{y}_0}$.
What if the do-feature is correlated with other features?

How do we isolate the causal effect?
Propensity score

Propensity of receiving a treatment:
- propensity score $e_i$ defined as $p(Z = 1|x_i)$,
- estimate with a classification method returning class probabilities.

Use $e_i$ to define propensity weights $w_i$ as

$$w_i := \frac{z_i}{e_i} + \frac{1 - z_i}{1 - e_i}.$$ 

Weight each sample $i$ by its weight $w_i$ in order to generate synthetic samples so that $Z$ is no longer correlated to $X$. This is called \textit{inverse probability of treatment weighting (IPTW)}. 
Causal effect in non-randomized trials

Using machine learning and propensity:

1. train a model with covariates $X$ in order to predict $Z$,
2. calculate the propensity scores $e_i$ by applying the trained model to all $x_i$,
3. train a second model with covariates $X$ and $Z$ as features and response $Y$ as target by using $w_i$ as sample weight for the $i$-th observation,
4. use this model to predict the causal effect like in the approach of the randomized trial.
Let the expected recovery time in days be Poisson distributed with

$$E(t_{recovery}) = \exp(2 + 0.5 \cdot I_{male} + 0.03 \cdot age + 2 \cdot severity - 1 \cdot I_{medication}),$$

where $I$ is an indicator function.

For the covariates we have:

- $sex \sim U\{0, 1\}$
- $age \sim \gamma(8, 4)$
- $severity \sim \beta(3, 1.5)$

Use this to generate 10,000 samples and note that $\exp(-1) \approx 0.37$. 
Distribution of age and severity
1. Randomized trial
   a) assign treatment randomly with \( p(Z = 1) = \frac{1}{2} \)
   b) use Poisson regression
   c) use Random forest

2. Non-randomized trail
   a) assign treatment based on \( X \)
   b) use Poisson regression
   c) use Random forest
Correlation matrix in randomized trial

No correlation between $X$ and $Z$
Poisson regression in randomized trial

Poisson regression correctly estimates the coefficients of the features

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>GLM</td>
<td>Df Residuals:</td>
<td>9995</td>
</tr>
<tr>
<td>Model Family:</td>
<td>Poisson</td>
<td>Df Model:</td>
<td>4</td>
</tr>
<tr>
<td>Link Function:</td>
<td>log</td>
<td>Scale:</td>
<td>1.0</td>
</tr>
<tr>
<td>Date:</td>
<td>Fri, 31 Mar 2017</td>
<td>Deviance:</td>
<td>10080.</td>
</tr>
<tr>
<td>Time:</td>
<td>14:30:21</td>
<td>Pearson chi2:</td>
<td>1.00e+04</td>
</tr>
<tr>
<td>No. Iterations:</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|       | coef   | std err | z      | P>|z|  | [0.025 | 0.975] |
|-------|--------|---------|--------|------|--------|--------|
| sex   | 0.4994 | 0.002   | 211.934| 0.000| 0.495  | 0.504  |
| age   | 0.0301 | 8.95e-05| 335.807| 0.000| 0.030  | 0.030  |
| severity | 2.0000 | 0.006   | 309.610| 0.000| 1.987  | 2.013  |
| medication | -1.0024 | 0.003   | -387.721| 0.000| -1.007 | -0.997 |
| const | 1.9990 | 0.006   | 326.234| 0.000| 1.987  | 2.011  |
Random forest in randomized trial

- Estimation of the average effect quite accurate
- Estimation of individual effects still decent
Assign treatment based on covariates, i.e.

\[
Z = \begin{cases} 
1, & \frac{1}{3} \cdot I_{male} + \frac{2}{3} \cdot severity + \epsilon > 0.8, \\
0, & \text{otherwise}
\end{cases}
\]

where \( \epsilon \sim \mathcal{N}(0, 0.15^2) \).
Feature *sex* and *severity* are highly correlated to *medication*.
Poisson regression in non-randomized trial

- Poisson regression still correctly estimates the coefficients
- Model dependence works in our favor

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>GLM</td>
<td>Df Residuals:</td>
<td>9995</td>
</tr>
<tr>
<td>Model Family:</td>
<td>Poisson</td>
<td>Df Model:</td>
<td>4</td>
</tr>
<tr>
<td>Link Function:</td>
<td>log</td>
<td>Scale:</td>
<td>1.0</td>
</tr>
<tr>
<td>Date:</td>
<td>Fri, 31 Mar 2017</td>
<td>Deviance:</td>
<td>10018.</td>
</tr>
<tr>
<td>Time:</td>
<td>14:30:23</td>
<td>Pearson chi2:</td>
<td>9.98e+03</td>
</tr>
<tr>
<td>No. Iterations:</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| coeff  | std err | z       | P>|z|  | [0.025 | 0.975 |
| ------ | ------- | ------- | -----|-------|-------|
| sex    | 0.5043  | 0.002   | 203.256 | 0.000 | 0.499 | 0.509 |
| age    | 0.0299  | 8.58e-05| 349.024 | 0.000 | 0.030 | 0.030 |
| severity | 1.9996 | 0.006   | 313.055 | 0.000 | 1.987 | 2.012 |
| medication | -1.0063 | 0.003 | -302.201 | 0.000 | -1.013 | -1.000 |
| const  | 2.0013  | 0.006   | 340.305 | 0.000 | 1.990 | 2.013 |
Random forest in non-randomized trial

- Average causal effect is quite off
- Quite many individual effects are estimated too high or the treatment effect too low resp.
Propensity scores in non-randomized trial

Calculation of the propensity scores in order to get the weights for the samples.
By using the propensity weights feature sex and severity are no longer as strongly correlated as before.
Propensity score in the randomized trial

- for comparison the propensity score in a randomized trial
- propensity score is $\frac{1}{2}$ as expected
Random forest with propensity

- estimation of the average causal effect improved
- estimation of individual causal effects improved in most cases but some outliers
Comparison

treatment effects for:
• randomized trial
• non-randomized trial
• non-randomized trial with IPTW
Conclusion

- IPTW improves estimating the causal effect in an observational trial compared to a naive approach.
- Postulating a model, e.g. Poisson regression, works for data from observational trials but is a bold assumption.
- Randomized trials remain gold standard, use it whenever possible.
- Further improvements can be accomplished by using the do feature only in a residual training.
References

- **E. Stuart;** The why, when, and how of propensity score methods for estimating causal effects; Johns Hopkins Bloomberg School of Public Health, 2011
- **Peter C. Austin;** “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies”; Multivariate Behav Res. 2011 May; 46(3): pp. 399–424

Blog post available at [florianwilhelm.info](http://florianwilhelm.info)
Questions?