Business Analytics

Session 6a. Causal Inference and Potential Outcomes Model

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.

- Do the magnitudes/units of different independent variables matter?

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.

- Do the magnitudes/units of different independent variables matter?
  - They matter a lot, because we need to calculate the distances between points and clusters.

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.

- Do the magnitudes/units of different independent variables matter?
  - They matter a lot, because we need to calculate the distances between points and clusters.

- What are the advantages of hierarchical clustering?

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.
- Do the magnitudes/units of different independent variables matter?
  - They matter a lot, because we need to calculate the distances between points and clusters.
- What are the advantages of hierarchical clustering?
  - Intuitive; no need to specify the number of clusters.

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.
- Do the magnitudes/units of different independent variables matter?
  - They matter a lot, because we need to calculate the distances between points and clusters.
- What are the advantages of hierarchical clustering?
  - Intuitive; no need to specify the number of clusters.
- What are the advantages of $k-$means clustering?

Exercises

- (T/F) To perform clustering, we need to specify the outcome variable Y.
  - False. We just need to have the covariates X.

- Do the magnitudes/units of different independent variables matter?
  - They matter a lot, because we need to calculate the distances between points and clusters.

- What are the advantages of hierarchical clustering?
  - Intuitive; no need to specify the number of clusters.

- What are the advantages of $k-$means clustering?
  - Works well with a large data set; minimizes the squared errors.

Why Should We Care About Causal Inference?

Causality

- Recall our regression/classification model:

$$Y \approx f(X)$$

- Goal: To identify that changes in X cause changes in Y.
  - To understand the cause of a problem.
  - To understand the effect/consequence of a change.
  - To inform what changes should be made to improve the status quo.

- Questions of interest:
  - Are human activities causing climate changes?
  - Will more education lead to higher incomes?
  - Will inventory information change customers' purchasing decisions?
  - How much more time will you use Kuaishou per day if a more interactive UI is available?

Causation vs. Association

Does hospitalization lead to better health?

- Data from National Health Interview Survey:

Group	Sample Size	Mean Health Status	Std Error
Hospital	7774	3.21	0.014
No hospital	90049	3.93	0.003

* 1 refers to poor health; 5 refers to excellent health.

- Hospitalization is associated with poorer health.
  - Can we say that hospitalization leads to poorer health?
- Another example: WeChat promotion.

Group	Converted	Not Converted
Promotion	20	230
No promotion	23	227

- The two groups of customers seem to have very similar (i.e., statistically the same) conversion rates.
  - Can we say that WeChat promotion has little (or even negative) value?

Causation vs. Association

- Key issue: We are unable to observe what would have happened to each individual if the alternative action had been applied.
- People who are seriously ill are more likely to be admitted into a hospital in the first place, but we are unable to see:
  - What happens to a seriously sick person if not admitted into a hospital?
  - What happens to a slightly sick person if admitted into a hospital?
- Only those people who are unlikely to convert received the WeChat promotion. We are unable to see:
  - What happens to a likely-to-convert customer if receiving the promotion?
  - What happens to an unlikely-to-convert customer if not receiving the promotion?

Potential Outcomes Model

Counterfactuals and Potential Outcomes

- We call the unseen information about each individual the counterfactual.
  - Without reasoning about counterfactuals, we cannot draw causal inferences.
- Two possible actions applied to an individual:
  - 1 or "treatment"
  - 0 or "control"
- For each individual, two associated potential outcomes:
  - $Y(1)$: Outcome if treatment applied
  - $Y(0)$: Outcome if control applied

Causal Effect

- Causal effect: The difference between the outcome if they are assigned treatment or control.

$$\text{Causal effect} = Y(1) - Y(0)$$

- Fundamental problem in causal inference: For each individual, we either observe $Y(1)$ or $Y(0)$, but not both.
  - Causal inference is a problem of missing data.

- Question: How can we resolve this issue?

Assignment

- Assignment mechanism (W): $W = 1$ (resp. 0) if an individual is assigned to treatment (resp. control).
- In the hospital example, individuals are partially self-assigned, and partially assigned by doctors.
- In the promotion example, the firm assigns the customers to treatment or control, but there may be some biases in the assignment.
- Randomized assignment: Assigned to treatment or control at random.

WeChat Promotion Revisited

- $W = 1$ means promotion received; $Y = 1$ means individual converted.
- The stared entries $(*)$ are what we observe.

Individual	W_i	$Y_i(1)$	$Y_i(0)$	Causal Effect
1	0	1	0 $(*)$	1
2	0	1	0 $(*)$	1
3	0	1	0 $(*)$	1
4	0	1	1 $(*)$	0
5	0	1	1 $(*)$	0
6	0	1	1 $(*)$	0
7	1	1$(*)$	0	1
8	1	1$(*)$	0	1
9	1	1$(*)$	0	1
10	1	0$(*)$	0	0
11	1	0$(*)$	0	0
12	1	0$(*)$	0	0

- The average conversion rate is the same for treatment and control groups (50%). Can we say anything about the causal effect of WeChat promotions?

Estimating Causal Effects

Average Treatment Effect

- We cannot observe both potential outcomes for each individual.
  Possible approaches:
  - Observe the same individual at different points in time.
  - Observe two individuals who are nearly identical, and give one treatment and the other control.

- Average Treatment Effect:

$$\mathsf{ATE} = \mathbb{E}[\mathsf{Y}(1)] - \mathbb{E}[\mathsf{Y}(0)]$$

  - We lose individual information, but get an estimation of both terms in expectation.

- Estimating ATE:

$$\widehat{\mathsf{ATE}} = \frac{1}{n_1} \sum_{W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{W_i=0} Y_i(0) \approx \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0]$$

- Question: When is $\widehat{\mathsf{ATE}}$ a good estimate of ATE?

Selection Bias

$$\widehat{\mathsf{ATE}} \approx \underbrace{\mathbb{E}[\mathsf{Y}(1) - \mathsf{Y}(0)|\mathsf{W} = 1]}_{\text{Expected Causal Effect for Treatment}} + \underbrace{\mathbb{E}[\mathsf{Y}(0)|\mathsf{W} = 1] - \mathbb{E}[\mathsf{Y}(0)|\mathsf{W} = 0]}_{\text{Selection Bias}}$$

$$\widehat{\mathsf{ATE}} \approx \underbrace{\mathbb{E}[\mathsf{Y}(1) - \mathsf{Y}(0)|\mathsf{W} = 0]}_{\text{Expected Causal Effect for Control}} + \underbrace{\mathbb{E}[\mathsf{Y}(1)|\mathsf{W} = 1] - \mathbb{E}[\mathsf{Y}(1)|\mathsf{W} = 0]}_{\text{Selection Bias}}$$

- Theorem. $\widehat{\mathsf{ATE}} \approx \mathsf{ATE}$ if there is no selection bias:

$$\mathbb{E}[\mathsf{Y}(0)|\mathsf{W} = 1] = \mathbb{E}[\mathsf{Y}(0)|\mathsf{W} = 0], \ \mathbb{E}[\mathsf{Y}(1)|\mathsf{W} = 1] = \mathbb{E}[\mathsf{Y}(1)|\mathsf{W} = 0]$$

- No selection bias: Assignment to treatment uncorrelated with outcomes.
    - Not satisfied for the cases of hospitalization and WeChat promotion.

- Selection bias is the biggest challenge in causal inference.

Selection Bias: Example

- WeChat promotion result:

Individual	W_i	$Y_i(1)$	$Y_i(0)$	Causal Effect
1	0	1	0 $(*)$	1
2	0	1	0 $(*)$	1
3	0	1	0 $(*)$	1
4	0	1	1 $(*)$	0
5	0	1	1 $(*)$	0
6	0	1	1 $(*)$	0
7	1	1$(*)$	0	1
8	1	1$(*)$	0	1
9	1	1$(*)$	0	1
10	1	0$(*)$	0	0
11	1	0$(*)$	0	0
12	1	0$(*)$	0	0

- $\text{ATE} = \frac{6}{12} = 0.5$
- $\widehat{\text{ATE}} = \frac{3}{6} - \frac{3}{6} = 0$
- ATE for Treatment: $\text{ATT} = \mathbb{E}[Y(1) - Y(0)|W = 1] = \frac{3}{6} = 0.5$
- ATE for Control: $\text{ATC} = \mathbb{E}[Y(1) - Y(0)|W = 0] = \frac{3}{6} = 0.5$
- Selection biases:

$\mathbb{E}[Y(0)|W = 1] - \mathbb{E}[Y(0)|W = 0] = \frac{0}{6} - \frac{3}{6} = 0.5$ and $\mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(1)|W = 0] = \frac{3}{6} - \frac{6}{6} = -0.5$

Randomized Experiment

- Randomized experiment: Subjects are randomly assigned to treatment or control (W is completely random).
  - No selection bias: W and and the outcomes are independent.
  - Randomized experiments are the "gold standard" of causal inference.
  - Other names: Randomized controlled trial; A/B test.

- In a randomized experiment, $\widehat{\mathsf{ATE}}$ is a good estimator of ATE.
  - Estimated standard error $\widehat{\mathsf{SE}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}$.
  - Easy to compute the t-statistic and the p$-$value.
  - 95% confidence interval: $[\widehat{\mathsf{ATE}} - 1.96\widehat{\mathsf{SE}}, \widehat{\mathsf{ATE}} + 1.96\widehat{\mathsf{SE}}]$.

Regression Analysis

- An alternative approach: Linear regression.

$$\mathsf{Y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 \mathsf{W}_i, \text{ where } \mathsf{W}_i = 0, 1$$

- In a randomized experiment:
  - $\hat{\beta}_0$: Average outcome in the control group.
  - $\hat{\beta}_0 + \hat{\beta}_1$: Average outcome in the treatment group.
  - $\hat{\beta}_1 = \widehat{\mathsf{ATE}}$.

- Let's see an example. Read in the data "RE.csv".
  - Y =life of the patient (in months); W = adoption of a new drug.
  - We flip a coin to determine whether a patient will receive this drug.
  - What are the estimated values for $\widehat{\mathsf{ATE}}$, $\hat{\beta}_0$, and $\hat{\beta}_1$?
  - How do you interpret these estimates?