Spring 2021 6.871: PSet 5

Collaboration Policy

- Your solution document must be created completely on your own.
- You are free to collaborate via discussions with other students. Please acknowledge all collaborators on your final write-up.

Submission: Gradescope & Dates

Please submit your report via Gradescope. The due date of Pset 5 is May 14, 11:59 pm ET.

Q1) Causality: Translating to Math

Written by Matthew McDermott and Monica Agrawal.

Overview

In this problem, we will present you with several free text scenarios. For each scenario, you must answer first whether or not causal inference is required in this scenario, and, if so, you must identify the relevant covariates (X), treatments (T), outcomes (Y), and any hidden confounders (H) that pose particular concern in this setting.

Learning Goals: We hope that this problem will help you develop the ability to recognize causal problems in the wild, to be able to translate them into the mathematical framework discussed in lectures, and to think critically about any missing confounders.

Problem

Example: You note that ice cream sales are strongly correlated to drowning rates in the United States, and decide to test if this is a causal phenomenon based on a large observational dataset at a monthly, county-specific level consisting of quantities of flavors of ice cream ordered, the quantity of hot-dogs eaten, the average socioeconomic status of residents in that county, and number of drownings.

Solution: Yes, this does require causal inference, as you wish to understand the causal link between ice cream sales and drowning rates. T would be whether ice cream of various flavors were eaten. Y would be the number of drownings observed. X would be the socioeconomic data about the county and the # of hot-dogs eaten. A critical hidden confounder that might damage this analysis is the average temperature that month, which likely both inspires ice cream consumption and swimming.

[4 points] **1.1.** You have collected physiological time series data (in particular, EEG traces, heart rate, SpO2, and blood pressure) from patients during a set of surgical procedures which are controlled by the application of anesthetic drug D at various dose levels continuously throughout the procedure. While most patients survive the procedure, a small minority do not. You want to design a model to automatically control the drug dosage level over time to maximize chance of patient survival.

[4 points] **1.2.** You work for a hospital and have just been assigned to help a new payer (e.g., insurance agent) satisfy an odd request. They like to receive advanced estimates of billing codes they will eventually be charged for their patients, and have requested that you provide them with an estimate of each patient's final billing code after only their first 24 hours. Contractually, this payer is prohibited from using these estimates to affect whether they accept or dispute claims, but instead uses them to more proactively assess their revenue streams. To accommodate this request, you plan to build a model based on retrospective data to take the first 24 hours of a patient's stay and predict the final ICD10 codes they received at discharge time.

[4 points] **1.3**. You work for a hospital which, like many hospitals, receives penalties from major payers if patients are readmitted within 30 days of discharge. To help prevent this, you decide to train a model based on past patient data, ingesting all EHR data up to discharge and predicting whether or not the patient will be readmitted within 30 days. Ultimately, you hope to use this model to audit and inform discharge decisions.

Q2) Computing ATE and CATE

Adapted from Uri Shalit and Rom Gutman.

Overview

In this problem, you will work with a simulated experiment and dataset to formulate an explicit causal graph and work through calculating CATE and ATE by hand.

Learning Goals: In this problem, we hope to give you a hands-on familiarity with ATE and CATE with real data in a simulated context.

Problem

The data in Table 1 comes from an experiment run from 2004-2008 at Sunnydale Hospital. In the data, you have binary indicators for: prior education (Z), whether the surgeon had ≥ 100 successful surgeries in 2004 (X), whether the surgeon enrolled in the fellowship from 2005-2006

(T), whether the surgeon had ≥ 150 successful surgeries in 2007 (Y), and whether the surgeon cumulatively had ≥ 500 successful surgeries in their lifetime by 2008 (W).

Z	X	T	Y	W
0	1	1	1	0
0	0	1	1	0
0	0	1	1	0
1	1	1	1	1
1	0	1	1	1
1	0	1	1	1
1	1	1	0	0
1	0	1	0	0
0	0	1	0	0
1	1	0	1	1
1	0	0	1	0
1	0	0	1	0
1	1	0	0	1
1	1	0	0	1
0	1	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Table 1: Sample data for problem 2.

We know the following:

(i) The number of successful surgeries in 2004 depends solely on the prior education.
(ii) Whether a doctor is selected to the surgery fellowship program depends on their prior education and number of successful surgeries in 2004.
(iii) The number of successful surgeries in 2007 depends on the fellowship, prior education, and number of successful surgeries in 2004.
(iv) The cumulative number of successful surgeries by 2008 is directly based on how many successful surgeries a doctor performs in 2004 and 2007.

Your task is as follows:

[6 points] **2.1.** Draw the causal graph that describes the above experiment.

[6 points] **2.2.** Calculate the Average Treatment Effect (ATE) of the fellowship (T) on 2007 success (Y). Use covariate adjustment and empirically estimate the probabilities/expectations from the observed data. Recall that

$$\text{ATE} = \mathbb{E}\left[Y_1 - Y_0\right]$$

where, for this causal graph,

$$\mathbb{E}\left[Y_t\right] = \sum_{z \in \{0,1\}} \sum_{x \in \{0,1\}} P(Z = z) P(X = x | Z = z) \mathbb{E}\left[Y | X = x, Z = z, T = t\right]$$

[6 points] **2.3.** Calculate the Conditional Average Treatment Effect (CATE) of the fellowship (T) on 2007 success (Y) for patients without prior education (Z = 0). Recall that

$$\text{CATE} = \mathbb{E}\left[Y_1 - Y_0 | Z = 0\right] = \mathbb{E}\left[Y_1 | Z = 0\right] - \mathbb{E}\left[Y_0 | Z = 0\right]$$

where, for this causal graph,

$$\mathbb{E}\left[Y_t | Z = 0\right] = \sum_{x \in \{0,1\}} P(X = x | Z = 0) \mathbb{E}\left[Y | X = x, Z = 0, T = t\right]$$