Machine Learning for Healthcare, 2019

Lab 2

Section 1:

For this lab, you will use the low birth weight (lbw) data. This is a publicly available dataset of linked birth and infant death data collected and maintained by the Center for Disease Control. Each line represents an infant, and has data about the infant's date of birth and death if he/she died within a year. It also has information about the mother, her education, race, smoking habits, alcohol consumption and other as well as some information about the father.

Download the data here:

http://people.csail.mit.edu/dsontag/courses/mlhc_summer19/day2/lab2/singletons.csv
http://people.csail.mit.edu/dsontag/courses/mlhc_summer19/day2/lab2/twins.csv

The following table is intended to clarify the differences between the following two questions:

Question	Treatment	Outcome	Dataset
Q1	Birth weight under 2700g? (binary)	1-Year Mortality? (binary)	twins.csv
Q2	Smoked during pregnancy? (binary)	Low birth weight? (real)	singletons.csv

Q1. Low Birth Weight Causes Infant Mortality?

You are doing so well at causal inference that Prof Sontag decides you are ready to meet some of his clinical collaborators. He introduces you to Dr. Buffy Summers at Sunnydale Hospital. She wants to know whether low birth weight (i.e. < 2700g) causes infant mortality, but she really can't figure out all of the jargon like "propensity" or "covariate." She hopes that you will be able to explain things in a more intuitive way.

Fortunately for you, you notice that the data has a section for twin babies (see twins.csv). Sometimes for a given pair of twins (where the environmental factors are the same), one baby is born below 2700g while the other is born above. "This will be great for explaining it to Dr. Summers without any fancy math!", you excitedly say to yourself.

a) Why would this dataset be so well-suited for studying the counterfactual effect of low birth weight on infant mortality? Explain in 1-3 sentences.

b) Filter the cohort so that you only have pairs where exactly one of them is below 2700g. Use these "counterfactual" pairs to estimate the ATE of low birth weight on the one year mortality rate for twins. Recall that

$$CATE(x) = y_1 - y_0$$

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} CATE(x_i)$$

Where t=0 indicates birth weight >= 2700g and t=1 indicates birthweight < 2700g.

c) Dr. Summers likes your explanation. She then asks whether this ATE generalizes to the whole population, including singletons? In other words, can we assume that the ATE of lbw in the singletons population is roughly the same? Justify why or why not. Hint: compute the mortality rates among the "counterfactual" twin pairs and among the singleton population.

Q1. Smoking During Pregnancy Causes Low Birth Weight?

Dr. Summers appreciates the simplified description, but she wants to focus on methods that can be applied to the full population of babies. She suggests that you shift focus and instead look at whether smoking during pregnancy causes low birth weight. She provides a large dataset with many factors about the mother and father (see singletons.csv).

Covariates: ['dmage', 'dmar', 'dlivord', 'anemia', 'cardiac', 'lung', 'diabetes', 'herpes',
'hydra', 'hemo', 'chyper', 'phyper', 'eclamp', 'incervix', 'pre4000', 'preterm',
'renal', 'rh', 'uterine', 'othermr', 'alcohol', 'm_race_black', 'm_race_other',
'm_race_white', 'm_edu_college', 'm_edu_elementary', 'm_edu_highschool',
'm_edu_morethancollege', 'm_edu_noedu']
Treatment: 'tobacco'
Outcome: 'dbirwt'

a) Dr. Summers asks you to start simple. Split the dataset into those who were treated (i.e. babies whose mothers smoked during pregnancy) and those who weren't. Compute the average birth weight of each cohort, and report the difference between these two average birth weights. Justify why, in general, this naive approach won't allow you to reason about the causal effects of the treatment.

b) Do a covariate adjustment to estimate the effect of mother's smoking habits on the baby's birth weight. Use sklearn's LinearRegression (with default hyperparameters) to fit a linear model, f, to predict the potential outcome given the covariates and treatment:

$$f(x,t) \approx \mathbb{E}[Y_t | T = t, x]$$

Report the estimated ATE using this method. In 1 sentence, describe the relationship between the ATE (as calculated using this method) and the coefficient for the treatment variable that is learned by the linear model. Recall that

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$$

c) An alternative to covariate adjustment is propensity score weighting. Use sklearn's LogisticRegression (with default hyperparameters) to fit a linear model to predict the treatment given the covariates. The probability of getting treatment t is known as the propensity score:

$\hat{p}(T = t|x)$

Report the estimated ATE using this method. State in 1 sentence what this ATE means to someone considering smoking during pregnancy. Recall that

$$A\hat{T}E = \frac{1}{n}\sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n}\sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Dr. Summers thanks you for your hard work.

Section 2:

Choose one of these two papers for your group to read and discuss:

1) Personalized diabetes management using electronic medical records
Dimitris Bertsimas, Nathan Kallus, Alexander M. Weinstein, and Ying Daisy Zhuo
Diabetes Care, 2016
http://care.diabetesjournals.org/content/early/2016/12/01/dc16-0826.full-text.pdf

2) Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study
Gabriel Brat et al., BMJ 2017
https://www.bmj.com/content/360/bmj.j5790

How does the causal inference problem that needs to be solved in the paper relate to what you learned in the lecture? Does it attempt to estimate the average treatment effect (ATE) or the conditional average treatment effect (CATE)? How does it use machine learning for this (e.g., through covariate adjustment, propensity score, or a variant of these)? What are potential weaknesses in the authors' analyses? How do the results inform health care policy?