Homework 4

Introduction to Causal Inference -
a Machine Learning Perspective 0365-4094-01
Winter 2019/2020

Submission date: January 12 , 2020

In this homework assignment you will experience working with actual (partially simulated) data, as preparation for the final project. You are more than encouraged to work in pairs. We generated the data in a manner that guarantees there are no unmeasured confounders.

The Data

Together with this document are two data files (*"data1.csv"* and *"data2.csv"*). Each contains $n = 4802$ samples, with the following properties:

1. $X_1 \dots X_{58}$: Covariates of the model (same in both datasets).
2. T: The treatment assignment of the model (binary).
3. Y: The outcome of the model.

Requirements

Your task is to estimate the Average Treatment effect on the Treated (ATT)[1], in both of the datasets. For each dataset you should first calculate the propensity score, and then estimate the ATT using the following approaches:

(i) Inverse Propensity Score Weighting (IPW)[2]

(ii) S-learner

(iii) T-learner

(iv) Matching

[1] $ATT \equiv \mathbb{E}[Y_1 - Y_0 | T = 1]$

[2] you may use Abdia et al. (2017) or Austin (2011)

The grade of this assignment will have two parts: 85% of the grade will be based on your ability to calculate estimates for each of the models (i)-(iv) above. Then, for the remaining 15% of the grade, we ask you to give your best estimate of the true ATT, using any of the methods above or any other method you wish to use. The 15% of the grade which are part of the competition will be calculated as a function of the distance from the true result and your classmates performance.

You are free to use any programming language (we recommend on using Python or R) and any model for estimating the propensities, the outcomes, and matching. Please think carefully which model you wish to use, as this is what will make a difference between the more accurate and less accurate estimates.

Submission Instructions

You are required to submit a single `zip` file containing:

1. the ATT results in a single `csv` file named `ATT_results.csv` with the format:

$$
\begin{array}{c}
Type, data1, data2 \\
1, ATT_1^1, ATT_1^2 \\
2, ATT_2^1, ATT_2^2 \\
3, ATT_3^1, ATT_3^2 \\
4, ATT_4^1, ATT_4^2 \\
5, ATT_5^1, ATT_5^2
\end{array}
$$

where ATT_i^j is the ATT of approach i on dataset j, and ATT_5^j is your final estimate (for the competition).

2. The propensity scores in a single `csv` file named `models_propensity.csv` with the format:

$$
\begin{array}{l}
data1, e_1^1, \ \ldots \ , e_n^1 \\
data2, e_1^2, \ \ldots \ , e_n^2
\end{array}
$$

where e_i^j is the propensity score of sample i in dataset j.

3. All of your code should be under a directory named `code`.

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., and Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59(5):967–985.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.