Controlling for Text in Causal Inference with Double Machine Learning

Emaad Manzoor
University of Wisconsin - Madison
emanzoor@wisc.edu, emaadmanzoor.com

Abstract

Text plays an increasingly important role in the study of causal relationships. In this tutorial, we consider the specific case of using text as a control to eliminate bias from confounders operating through the text. We formalize the problem of controlling for text using causal graphs and the potential outcomes framework, describe a principled estimation and inference procedure to realize this goal using double/debiased machine learning, and compare this procedure (hands-on) against several alternatives such as controlling for low-dimensional representations of the text obtained via topic modeling, principal component analysis, or other techniques. We conclude with a case study on using text as a control to quantify the causal impact of status on persuasion online.

Duration

We propose delivering this tutorial over **2 hours** in total, with the antipated time allocation below:

1. *30 minutes*: Introduction to the two tasks of causal inference: causal identification and statistical estimation; observational studies and associations; confounding bias, randomized experiments, and quasi-experimental studies.
2. *30 minutes*: Unstructured text as a control: examples of applications, formalization using causal graphs and potential outcomes, estimation challenges, issues with naive dimensionality-reduction approaches.
3. *30 minutes*: Double machine learning for causal estimation and inference: theory, advantages, underlying assumptions, comparison with other approaches.
4. *30 minutes*: Using double machine learning to control for text: the recipe, hands-on example with simulated data, case study on conversations from ChangeMyView.

This time allocation is based on a tutorial previously delivered by the authors at the University of Michigan[1].

Interaction style for hybrid format

We plan to have a natively virtual tutorial.

Tutorial schedule and activities

The proposed tutorial will span 4 logical sections (detailed in the *duration* section), and include both lecture and hands-on components interleaved with each other. The hands-on components can be executed and modified by attendees to better understand the concepts discussed in the lecture components. However, the modifying and re-running the hands-on components will be optional. The hands-on components will be hosted on Google Colab to maximize accessibility.

Target audience, prerequisites and outcomes

The proposed tutorial is targeted at both methodological and substantive researchers with an interest in the intersection of causal inference and NLP. These include researchers in natural language processing, statistics and machine learning, and (computational) social science. The audience is expected to know graduate-level statistics. The required natural language processing and causal inference background will be covered in the tutorial. At the end of this tutorial, attendees will be able to (i) formulate real-world causal inference tasks using causal graphs, (ii) use causal graphs to evaluate the suitability of text as a control, (iii) use double machine learning to perform estimation and inference of causal effects while controlling for text.

Materials

Slides, Google Colab notebook, and links to optional background readings will be shared prior to the tutorial.

Past Precedent

The proposed tutorial is based on a tutorial previously delivered by the authors at the University of Michigan, which included both lecture[1] and hands-on[2] components. The proposed tutorial is an hour longer than the previously delivered tutorial to incorporate comparisons with a larger number of baseline approaches, and provide a more comprehensive description of the underlying statistical theory. The hands-on component of the proposed tutorial will be self-contained in a Google Colab notebook and will not require any software licenses or installation by attendees.

[1] Video: https://www.youtube.com/watch?v=DwUqA1ydJI0

[2] Google Colab Notebook: https://colab.research.google.com/drive/15Jz9QehJsT2um1cH5GEbbBOeJDcX0e2n?usp=sharing