Causality Projects Description

ABSTRACT

For the project, we expect you to perform analyses on the data following what presented in the tutorials. In particular, we will be grading the output, which will be a 15 minutes group presentation (so around 15 slides in total) that should have all of the following points covered.

1 INTRODUCTION AND MOTIVATION (10%)

In this part, you will introduce the datasets, the assumptions and the causal questions you are investigating. This will be 10% of the grade of the report. In particular:

- Describe your dataset (e.g. what are the observational data and how they were collected, in case there are interventional data, also what are they and how they were collected).
- Describe the causal questions you wish to answer (e.g. "we investigate the effect of X on Y").
- Describe the assumptions of your dataset (causal sufficiency, no cycles in the causal graph, positivity, etc).

2 EXPLORATORY DATA ANALYSIS (15%)

In this part you will do exploratory data analysis as shown in Tutorial 2. This will be 15% of the grade of the presentation. In particular you can present:

- Testing correlation / dependence for the variables in the dataset and show how they are dependent;
- Discuss the true causal graph of the dataset, if it's known, and otherwise discuss a reasonable guess.

3 IDENTIFYING ESTIMANDS (20%)

Here you should identify possible adjustment sets by hand (showing that you understood the theory in the class) by using:

- Backdoor criterion (10%)
- Frontdoor criterion (5%)
- Instrumental variables (5%)

as per Tutorials 3 and 4. Keep in mind that you should report what happens for these methods even if they don't apply and explain why. You can also show the results you get for each of these estimands from doWhy and compare with the ones you found by hand. This will be 20% of the grade of the presentation.

4 ESTIMATING CAUSAL EFFECTS (15%)

Apply and explain different causal estimate methods (linear, inverse propensity weighting, two-stage linear regression, etc.) to your previously identified estimands, as shown in Tutorial 4. This will be 15% of the grade of the presentation.

5 CAUSAL DISCOVERY (20%)

In this part you will try out the two types of algorithms for learning causal graphs (constraint-based 10 % and score-based 10%) that will be shown in Tutorials 5 and 6. You are also expected to explain why each methods works or doesn't and what is identifiable in terms of the causal graph. This will be 20% of the grade of the presentation.

- Run a constraint-based algorithm (e.g. PC) and a score-based algorithm (e.g. GES) on your data, and report back any identifiable causal relations.
- *Optional:* If you cannot find any identifiable causal relation or just want to test the algorithms further, simulate some data that resemble your real data (but maybe with less edges).

6 VALIDATION AND SENSITIVITY ANALYSIS (10%)

In this part you will try out different ways to validate your results and do sensitivity analysis of the methods. This will be 10% of the grade of the presentation.

- Report using some of the results of the refutation strategies implemented in DoWhy and interpret what they mean.
- *Optional:* If your dataset includes interventional data, check that the estimated causal effects from the observational data are reflected in the interventional data.
- *Optional:* Try experimenting with graphs in which some of the edges are dropped, and see how the results in Section 3 and 4 change.
- *Optional:* Try relaxing some of the assumptions you discussed in the Introduction, e.g. try to see the effect on not observing a certain variable.

7 DISCUSSION AND CONCLUSION (10%)

In this part you will discuss the results of the previous sections and explain if they do answer the causal questions you described in the Introduction. You can also elaborate on the results you observed in the validation and discuss if the assumptions you had made initially were realistic.

REFERENCES