Exploiting the Classification Algorithm for Robust Multidimensional Homogeneity Test in Causal Inference

Proposal

Aukkawut Ammartayakun

MA 590 Special Topics: Causal Inference

Spring 2023

1 Introduction

Homogeneity is one of the crucial assumptions in the causal inference field. Traditionally, this can be test with hypothesis testing like the Kolmogorov-Smirnov test. However, those test has a limitation on the dimension of the data. This work challenges that limitation by looking at this problem from a different perspective. Recall that hypothesis testing is essentially the classification problem of whether one has statistical evidence to reject the null hypothesis or not. Semantically, they both are different approaches but those problems are equivalent in the sense that we can have the classification functions $f : \mathcal{D} \to [0, 1]$ such that $f(x, y)$ for $(x, y) \in \mathcal{X} \supseteq \mathcal{D}$ reflecting the probability of rejecting the null hypothesis $H_0 : X = Y$ for paired data (x, y) in the data space $\mathcal{X}$ which now, the p-value can be estimated from f and the decision can be done with thresholding.

This work will explore the possibility of exploiting the classification algorithm for robust multidimensional homogeneity test with the application in causal inference. First, this work will explore the known method for homogeneity test in high dimensional space like exploiting f-divergence, maximum mean discrepancy, C2ST, etc. Then, the proposed method will be analyzed and tested in comparison to the former methods.

1.1 Hypothesis testing exploitation

As the process of hypothesis testing involves evaluating the test statistics and using that to find the critical region under the distribution. However, that can be simplified, hopefully, into the classification problem that can be exploited with the machine learning model. The transformation of the data into its representation form, i.e., embedding might reveal the underlying causal relation. The use of a machine learning model in substitution for hypothesis testing for causal inference will be explored, along with the analysis of the power of testing using a classification model.

1.2 Logistic Regression Analysis

Let say we have the data space $\mathcal{X}$ and the paired dataset $\mathcal{D} \subseteq \mathcal{X}$. The goal of this is to infers to one paired sample data $(x, y) \in \mathcal{D}$ whether classify the rejection of the null hypothesis $H_0 : X = Y$. For the convience, let assume that x and y are the same dimension, the probability of rejecting the null hypothesis (that is, X and Y are not the same distribution) for a given paired sample data vectors (x, y) can be determined.

For the naive case, let assume that X and Y are linearly separateable. That is, there exists a hyperplane H such that $x \in H$ if and only if $y \in H$. Then, the probability of rejecting the null hypothesis can be determined by the distance of the paired sample data vectors (x, y) to the hyperplane H.

$$P(H_0 \mid (x, y)) = \frac{1}{1 + \exp(-\alpha d(x, y))}, \tag{1}$$

where α is the hyperparameter that controls the steepness of the sigmoid function and $d(x, y)$ is the distance of the paired sample data vectors (x, y) to the hyperplane H.

2 Dataset

ASSISTments dataset[1] will be explored in various aspects.

3 Proposed Method (Roughly)

3.1 Hypothesis testing exploitation

1. Explore the hypothesis testing and analyze the classification problem as the statistical test to show the feasibility of the method and analyze the tradeoff of this method.
2. Comparing the traditional method of determining the causal relationship with the deep learning model mainly aims to classify whether to reject the null hypothesis (whether A has an effect on B) or not.

4 Goal

To apply the exploited homogenity test to test whether the result from the treatment group and outcome group are homogenious or not. If the result is not homogenious, then the treatment group has an effect on the outcome group.

5 Literature Review

5.1 Explotation on f-divergence

Definition 1. *Let $p(x)$ and $q(x)$ be two probability distributions such that $q(x) > 0$ for all x and let f be a function that is convex and satisfies $f(0) = 0$. Then, f-divergence is defined as*

$$D_f(p||q) = \int p(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{2}$$

One can see that f-divergence family would include the KL divergence, Jensen-Shannon divergence, etc. With this, one can exploit the f-divergence in order for semiparametric homogenity test.

5.2 Homogenity test

6 Approach

6.1 Prove that the transformation of embedding is sufficient for test statistic

6.2 Prove that the model itself have the same classification power as the hypothesis testing

In this path, one can easily show empirically that the type II error rate is the same as the power of the test. However, the problem arise when the prove is needed to be done analytically, which is the main goal of this project.

References

[1] Wang, Y., Heffernan, N. T., and Heffernan, C. Towards better affect detectors: Effect of missing skills, class features and common wrong answers. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (New York, NY, USA, 2015), LAK '15, Association for Computing Machinery, p. 31–35.