What If I sold my house last year

Kim Hoe In
Department of Economics,
Seoul National University, Seoul, Korea

***Abstract*—In this paper, we estimate the casual effect under observational data with machine learning.**

I. Introduction

In recent years, Machine learning has made remarkable progress. Especially as the domains of image and text recognition adapt the method of machine learning, that have enabled new products like images searches, automatic translation between many languages, and even synthesis of realistic images and voice. And its application stretches to business section and social sciences. The reason why machine learning algorithms proliferate in many fields is that their ability is adequate for detecting patterns and correlation. And that is also the challenge of machine learning.

In fact, the goal of business world is to explain and predict the influences of inputs to make decision and the goal of social science is to explain the reason why the phenomena happened by finding causal relationship. However, it is hard to find causal relation, so we use the method of causality inference. Causality inference is about estimating the causal relation from observed data. And machine learning will be useful tool to estimate that effect. Also, we can see combining machine learning and causal inference is powerful way to make decision.

II. Related Works

Pearl [1] proposed a framework for causality inference using do(.) operator which means intervention and distinguished the observational conditional probability P(Y|A) and interventional conditional probability P(Y|do(A)). This means if A leads to change in Y with keeping everything else constant, A causes Y, and We call A "intervention", denoted by do(A). And the important key of causal inference is to find out the "confounders". Confounder is variable that effects treatment as well as outcome. By eliminating the effect of confounders, the true effect of treatment on outcome can be correctly figured out.

To answer the "What if" question, we have to perform planned or randomized experiments, and then compare the P(Y|do(A=1)) and P(Y|do(A=0)) - A is treatment, and Y is outcome. On this occasion, the gold standard is to conduct a randomized experiment [2]. Because it ensures that there is not systematic difference between two subsets. The next step is calculating the effect of treatment. The most exact way is to calculate 'Individual Treatment Effect(ITE)', and its formula is below.

$$ITE_i = Pr(Y_i|do(A=1)) - Pr(Y_i|do(A=0))$$

ITE is derived by comparing the identical object's conditional probability of outcome between under treatment and under no treatment . If ITE is equal to 0, the treatment X does not cause outcome Y, and if ITE is equal to 1, the treatment X causes outcome Y. But this is not possible in real world. Because we can not reproduce the object. So, We estimate 'Average Treatment Effect'(ATE). Initially, as we construct randomized experiment, each subsets have no difference. In other words, We can assume that the unique difference of each groups is only the fact that one group is under treatment and the other is not. From that assumption, we calculate ATE.

$$ATE = E[Y|do(A=1)] - E[Y|do(A=0)]$$

But that experiments are too expensive, or not feasible or have ethical issues. So, we have to depend on observed data. Such data, however, is biased by correlations and unobserved confounding and thus there exists systematic differences between two subsets. This is why the core work of causality inference is to remove such correlations and confounding from the data and estimate the true effect of an action. At this moment, "Causality inference" becomes powerful tool to figure out the causal relationship. There are many ways to infer causation. In this report, "Matching method with propensity score" [3] is used to infer causality inference. The matching method first calculates distance between i object under treatment and j object under no treatment. If the distance is smaller than ϵ, those two objects are matched, and then, model calculate the difference of Pr(Y|do(A)). Finally, by figuring out the expectation value of that, estimate the ATE.

Propensity is used to calculate the distance. Propensity score measures the propensity of individuals to receive treatment given information about confounders(X).

$$\pi(X) = P(A=1|X)$$

Propensity score [4] is close to 0 for individuals who have a low possibility of receiving treatment and is close to 1 for those who have a high probability of receiving treatment. We can estimate this score by using machine learning algorithms like SVM, Logistic regression, XGB and so on [5]. The Input is the confounders and the outcome is whether under treatment or not. The reason why we use propensity score is that we can eliminate the effect of confounders.

III. Problem Definition

The main goal of this research is to answer the question, "What if I sold my house last year". The used dataset is about the house sales in King County, Washington State, USA from Kaggle and this dataset consisted of historic data of houses

sold between May 2014 to May 2015.

In this case, the treatment is when I sold my house - 2014 is under treatment and 2015 is under no treatment. And to simplify the discussion, 3 features(area of lot, area of living space, and built year) are extracted and model is constructed with those features. Also, to make more accurate model, unobserved confounders are added. By using the matching method with propensity score, the ATE is estimated, which means the average difference of sale price between in 2014 and 2015.

IV. Solving Approach

The process of causal analysis can be broken down into four stages [6]. The first step is to model the causal question, the second to identify the estimand, the third to estimate the effect and fourth to refute the obtained estimate.

And there are three main assumption : Ignorability, Stable Unit Treatment Value(SUTVA), and Common Support.

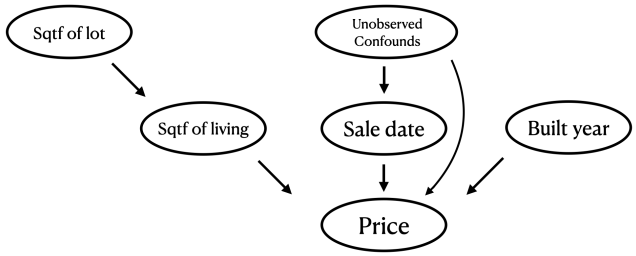
Ignorability means that if we condition the observed covariate, unobserved covariate can be ignored.

SUTVA means that treatment influences only on the outcome.

Common support means that the populations of the under treatment and the under no treatment have similar attributions.

V. Evaluation

First, model is designed like below.



Treatment : date(sale date)
Outcome : price
Confounders : unobserved
Covariates that effect only outcome :
built year, sqft of lot, sqft of living

Second, using the input graph, check all possible ways of identifying a desired causal effect based on the graphical model. And then, estimate the causal effect by matching method with propensity score. The estimated ATE is -16802.51, which means if I had sold my house last year, the price of my house would have been less valued about 16802.

Final step is to check whether this causality is true or not. To do this, placebo treatment refute method is used. Placebo treatment refute method checks whether the estimator returns an estimate value of 0 when the action variable is replaced by a random variable, independent of all other variables. So, If the treatment truly causes the outcome, the result will be close to zero. Or if not, the p-value is under 0.05, which means reject the null hypothesis that the treatment causes the outcome. And the result of p-value is 0.43. so, we can say that if I had sold my house last year, I would have taken some losses.

ATE	$p-value$
-16802.51	0.43

VI. Conclusion

Using a roughly designed model, we have checked simple but interesting causal relationship. Causal inference can be applied to many fields. We can find the casual signals on the image [7]. For instance, if the bridge had been removed from the image, it would make little sense. This means the presence of bridge has an effect on the presence of the car. Like above example, Machine learning can leap forward upper level and explain the "why" question by using causal inference. It is, however, true that combining these two fields is hard to do. This means that there are many topics to research, and this makes us feel excited.

References

[1] V. Didelez and I. Pigeot, "Judea pearl: Causality: Models, reasoning, and inference," *Politische Vierteljahresschrift*, vol. 42, no. 2, pp. 313–315, 2001.

[2] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[3] C. Heinrich, A. Maffioli, G. Vazquez *et al.*, "A primer for applying propensity-score matching," *Inter-American Development Bank*, 2010.

[4] J. M. Miguel A.Hernan, "Causual inference:what if," pp. 183–189, 2020.

[5] Y.-G. Choi and D. Yu, "Causal inference from nonrandomized data: key concepts and recent trends," *The Korean Journal of Applied Statistics*, vol. 32, no. 2, pp. 173–185, 2019.

[6] Microsoft, "Tutorial on causal inference," 2020. [Online]. Available: https://microsoft.github.io/dowhy

[7] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6979–6987.