A Survey on Policy Learning

MATH 818.01 Midterm Survey

Wonjun Choi

October 19, 2020

Abstract

This paper surveys growing literature on policy learning in the interdiscipline area of economis, statistics, and computer science. Policy learning incorporates statistical decision making, AI/ML algorithms, and potential oucome model in economics literatue to find optimal policy assignment rule. Policy learning predicts expected outcome of a policy given practical restrictions such as budget constraint, fairness, or political reasons. Considerations rasied in the economics literature, applications of AI/ML algorithms, and mathematical foundation of the results are briefly introduced in turn. A suggestion for the final project is also presented in the final section.

1 Introduction

Suppose you are a dean of the department and you are to allocate each graduate student to put an elephant into a refrigerator.

Online vs Offline setup

Bayesian vs Frequentist

2 Economic Modeling of the Allocation Problem

It is worth to stop and introduce typical approaches in economics for policy evaluation. Introduced notations would ease our conversation throught this paper.

Policy evaluation has been one of the most important topic in the field of economics. As the government implements various economic policies, the outcome of the interventions need to be analyzied and has been studied in the framework of *potential outcome*, which was frequently refered as *Rubin's causal model*.

Potential Outcome Framework

Consider we are giving aspirins to patients and observing their body temperatures. After treatments are assigned to each patient, we can only observe only *one side* of the outcome; the temperature with

or without taking aspirin. If one takes her aspirin, we cannot know what the temperature would have been without taking it, and vice versa. This is the fundamental problem of *treatment effect* analysis.

What would be the effect of an aspirin on body temperature? It would be the difference of boty temperature between taking aspirin and not. With D equals 1 if the treatment is applied and 0 otherwise, let the potential outcome $Y(1)$ be the outcome with the treatment and $Y(0)$ be the outcome without the treatment. Now we can denote the treatment effect τ as

$$\tau = Y(1) - Y(0).$$

τ cannot be obtained without further assumptions since one of the outcome is not observed. In economics/statistics literature, the unobserved outcome is called *counterfactual*. I refer [Imbens and Rubin, 2015], among many others, for detailed explanations and issues in treatment effect analysis.

Regret Function and Optimal Policy

One practical concern in designing policy could be 'How should we assign (limited) treatments for the best outcome?'. [Manski, 2004] introduced a framework for this analysis to economists based on *statistical decision rule* of statistics literature([Wald, 1950]).

The concern of statistical treatement rule is that how to assign a treatment D to the population based on their covariates(features) $X \in \mathcal{X}$ to maximize utilitarian welfare U[1]. Let's call this assignment rule(function) as *policy* $\pi : \mathcal{X} \to \{0, 1\}$ and the collection of possible policies as $\Pi = \{\pi : \text{some restirctions on } \pi\}$.

It is not difficult to find this kind of problem in reality. For example, consider a Youtube's recommendation algorithm. For a given video A, the algorithm decides whether to recommend this video to you or not based on your characteristics[2]. Now let's say Youtube has decided to recommend this video to total 100 people among its users. Then Youtube would want to find an algorithm(policy) to maximize its total viewership(utilitarian welfare) among possible policies.

How can we measure the succesfulness of a given policy? In what sense, the optimal policy can be regarded as the best? As already mentioned, we are maximizing the utilitarian welfare so the best policy could be thought as

$$\pi_{opt} = \arg\max_{\pi \in \Pi} \; U(\pi)$$

where $U(\pi)$ is a total welfare of population when the policy π is implemented. By defining a *regert function* $R(\pi)$ as a difference between the welfare with the policy π and the best possible outcome,

$$R(\pi) = U(\pi_{opt}) - U(\pi),$$

we can measure the successfulness of the policy π by comparing $R(\pi)$. Now our objective becomes clear: finding a policy funtion π that minimizes the functional $R(\pi)$.

[1] If U is stochastic, consider the mean $E(U)$.

[2] Of course, if we set a policy as $\pi : X \to \{A, B, C, D, ...\}$, we can consider a more complexed decision making.

Theoretical efforts has been made to bound a reget function. With a properly decided policy, we can find a uniform bound of the regret function. Notice that if a regret function degenerate fastly enough, we might be willing to adopt that that policy as a solution to our decision making problem. For some results regarding a regret bound and the minimax decision criteria, refer [Manski, 2004] [Stoye, 2009], and [Hirano and Porter, 2009].

3 Policy Learning Embedding AI/ML Algorithms

The similar context can be found in computer science literature; *multi-armed bandit* and their close cousins. [about MAB]

[Dudik et al., 2011] proposes a *doubly robust* estimator for policy evaluation and optimization(learning). As mentioned in the previous section, the treatment effect of a policy cannot be estimated without further assumptions for the missing data problem(counterfactual). There are two typical approaches which the authors refer as *direct method*(DM) and *inverse propensity score*(IPS). For detailed treatment of these methods, refer their paper or [Imbens and Rubin, 2015]. I briefly introduce their experimental setups which are more relevant to our remaining discussions. Notations and explanations are mostly theirs([Dudik et al., 2011]).

Consider i.i.d. data drawn from a distribution D: $(x, c) \sim D$, where $x \in \mathcal{X}$ is the feature vector and $c \in C = \{1, 2, .., , k\}$ is the class label. An action a is chosen by the policy $p(a|x, h)$, where h is the history of previous observations. A reward r_a is revealed while other potential rewards $r_{a'}$ remain unknown. Defining the *value* of a policy π as

$$V^{\pi} = E_{x,\vec{r}}[r_{\pi(x)}],$$

our policy learing objective is to find a policy that maximizes the value function.

[Dudik et al., 2011] provides simulation results by transforming a classical classification problem[3] into a policy learning framework. They consider (x, c) as a observed sample and let (potential) loss as $(x, l_1, l_2, \cdots, l_k)$ with $l_a = 1[a \neq c]$ where $1[\cdot]$ is an indicator function. Then the two problems become identical. The opposite way of transforming is also interesting: we can use an optimization tool for classification to solve our policy optimization.

[Kitagawa and Tetenov, 2018] uses [Dudik et al., 2011]'s IPS estimator and proves that their *empirical welfare maximize* methods meets semiparametric efficient minimax regret bound under some assumptions that are commonly used in economics literature. For readers who are interested in the assumptions that economists use may refer their paper.

As the flexible and powerful feature of various AI/ML algorithms remain attractive to economists' eyes, there has been trials to embrace those methods for policy evaluation. Among many others, I introduce [Athey and Wager, 2017] whose method can be equipped with AI/ML techniques. They suggests an algorithm to find such optimal policies:

[3] In a classification problem, we are searching for a classifier $\pi : \mathcal{X} \to C$ that minimizes the classification error.

1. Estimate the potential oucome equation $\hat{m}$ and the some function $\hat{g}$ using any methods whose rate of convergence is known.

2. Construct a score function for the value function $\hat{\Gamma}$ using nuisance components estimated in 1.

3. Find $\hat{\pi} = \arg\max\{\sum(2\pi(X_i) - 1)\hat{\Gamma}\}$ where $\pi(\cdot)$ is a trained weighted classifier.

For detailed instruction for $\hat{m}$, $\hat{g}$, $\hat{\Gamma}$, refer their paper. Here, part 1 and 3 of the algorithm can be obtained with AI/ML methods.

For the part 1, we can use any methods whose rate of convergence in mean square error(MSE) is known. As many statistical(machine) learning techniques use a MSE criteria for their loss functions, various ML estimators can be used for the part 1. Also their convergence rate has been widely investigated nowadays. While optimization in the part 3 is not a convex optimization, the computation of the part 3 could be troublesome. As mentioned before, the problem can be translated into a weighted classification problem so we can use techniques developed for those classifications.

4 Mathematical/Statistical Foundations

Theoretical properties of aforementioned methods rely on the calculation of *Vapnik-Chervonenkis dimesion*[Cite] which I abbreviate hereafter as *VC-dimension*. VC-dimension is a kind of measurement of the complexity of a collection. In our problem, we are searching for the optimal policy π within the collection of feasible policies Π. Thus, our result depends on the complexity of the collenction Π.

Definition 1 (VC-dimension)**.** *dd*

As we compare policies in a policy class Π, the things we have in our consideration depend on the complexity of Π. Statistical literature offers some valuable tools to control the complexity of a class. VC-dimensions and entropy integrals.

Also related to *learnable*.

5 Discussion/Conclusion

In the survey, I introduced a problem of treatment assign rule and one path of the recent development.

While the literature is fastly growing, there are still many unsolved questions remaining. First, in the real application of the problem, using the introduced methods requires several selection from user. For example, to estimate the part 1 of [Athey and Wager, 2017]'s algorithm, one have to decide which algorithm to use(or ensemble). Moreover, if the rate of convergence of that method is not known, one might have to derive it. The part 3 of their algorithms is also computationally tricky.

Developed models so far are still parsimonious for many applications. For example, as in MAB, a treatment assignment could be repeated during the process. Also, the data collection can be augmented

during the process(online setup). Using a longitudinal data might need more modifications especially when using ML methods.

For the final project, I would like to study the case of 'Disaster support' for COVID-19 in Korea. There was an debate on the way to distribute this aid: whether to provide it for everyone or low-incomed. I would like to investigate on this debate in terms of the consumption stimuli effect as many countries around the world considered similar policy as a fiscal stimuli for economy. Specifically I would like to adopt [Kitagawa and Tetenov, 2018]'s method or [Athey and Wager, 2017]'s method which involves an optimization using AI/ML algorithms.

References

[Athey and Wager, 2017] Athey, S. and Wager, S. (2017). Policy Learning with Observational Data. *arXiv*.

[Dudik et al., 2011] Dudik, Langford, and Li (2011). Doubly Robust Policy Evaluation and Learning.

[Hirano and Porter, 2009] Hirano, K. and Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, 77(5):1683–1701.

[Imbens and Rubin, 2015] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[Kitagawa and Tetenov, 2018] Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616.

[Manski, 2004] Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246.

[Stoye, 2009] Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.

[Wald, 1950] Wald, A. (1950). Statistical decision functions.