Learning from Data Through Models

Alberto Bisin, Guillaume Frechette and Jimena Galindo

July 11, 2023

Abstract

TBW

Introduction

TBW

Literature Review

TBW

Framework and Predictions

A finite number of observable random variables, $X = (x_1, x_2, ..., x_N)$, determine a binary state, y. The observable variables are independently distributed and each follows a Normal distribution with corresponding mean μ_i and standard deviation σ_i with $i \in \{1, 2, ..., N\}$.

The state is *Red* (**R**) or *Blue* (**B**) and is determined in the following way:

$$y = \begin{cases} \mathbf{B} & \text{if } a_1 x_1 + ... + a_N x_N + K \geq 0 \\ \mathbf{R} & \text{otherwise} \end{cases}$$

An observable variable x_i is *relevant* if the associated coefficient, a_i, is non-zero. Likewise, variable i is *irrelevant* if $a_i = 0$. We call the set of relevant variables $\mathcal{R}$ and the set of irrelevant variables $\mathcal{I}$.

In each period, the decision-maker (DM) observes the realization of all observable variables and predicts the state. Their flow payoff in period t is 1 if they predict the state correctly, and it is zero otherwise. Denote the prediction by $\hat{y}$.

In order to make a prediction, the DM uses a map $f : M_t \to \{\mathbf{B}, \mathbf{R}\}$
where $M_t \subseteq \{x_1, x_2, ..., x_N\}$ is the set of variables that they choose to consider in period t. Our main object of study is the set of variables that the DM decides to consider and not the properties of the mapping from the variables to the state. We refer to the set of variables that are taken into consideration, M_t, as the DM's *model of the world*, and we abstract away from how the variables are mapped into a prediction of the state.

A model of the world is said to be *simple* if it does not include all the relevant variables, *i.e.* if $M \subset \mathcal{R}$. And it is complex if it includes all relevant variables, $M = \mathcal{R}$.

Optimal Rules

The most accurate prediction rule for this setting will be one that predicts the state to be **B** whenever, given the observables, **B** is more likely. And it predicts **R** whenever **R** is more likely given the observables. Because we abstract away from the procedure through which the rule is updated, we will make predictions under the assumption that the DM has learned the parameters perfectly. That is, they have seen enough information that, under an updating procedure that converges to the truth—for example, correctly specified regression model or Bayes rule—they would have converged to the true parameters already.

We refer to a prediction rule as an *Optimal Rule* if it is the rule that a fully rational Decision maker would converge to with perfect Bayesian learning. In this context, Bayesian learning corresponds to using Bayes rule to learn the coefficients $a = (a_1, a_2, ..., a_N)$ as well as the constant K. This means that in an N dimensional space, the Decision Maker must learn $N + 1$ parameters which corresponds to learning the halfspace in which the state is **B**. The optimal rule for the complex model coincides with the true data-generating process when learning is perfect.

Conditional on using a simple model, the optimal rule for that model is the rule that, when restricted to the variables being considered, does the best at predicting the state. In this case, the Decision Maker is learning only the parameters that pertain the variables in the model. Hence, for a model that considers K variables, the DM will have to learn $K + 1$ parameters. The following proposition illustrates exactly what the optimal prediction rules look like when the DM is able to learn.

Proposition. *Fix any model that considers $k \leq N$ variables and relabel the variable so that $x_1, ..., x_K$ are the variables that the model considers and $x_{k+1}, ..., x_N$ are the ones that the model does not consider. Relabel the coefficients $a = (a_1, ..., a_N)$ accordingly as well.*

proof. *Let $M := a_1x_1 + ... + a_kx_k$ and $M^c = a_{k+1}x_{k+1} + ... + a_Nx_N + k$ and define the latent variable $y^L := M + M^c$. The optimal prediction rule for a model that considers the variables in M the state to be **B** whenever $P[y^L|M] \geq \frac{1}{2}$ and predicts **R** otherwise.*

Using the fact that for a random variable $z \sim N(\mu_z, \sigma_z)$ we have that $P[z \geq \mu_z] = \frac{1}{2}$ and noticing that M^c and $y^L|M$ are Normally distributed with means $\mathbb{E}[M^c]$ and $M + \mathbb{E}[M^c]$

respectively, it is easy to see that $P[y^L \geq 0|M] \geq P[y^L \geq M + \mathbb{E}[M^c|M]] = \frac{1}{2}$ *whenever* $M+\mathbb{E}[M^c] \geq 0$. *Similarly, whenever* $M+\mathbb{E}[M^c] < 0$ *we will have that* $P[y^L \geq 0|M] < P[y^L \geq M + \mathbb{E}[M^c|M]] = \frac{1}{2}$. *Therefore, the optimal rule is to predict* **B** *whenever* $a_1x_1 + ... + a_kx_k \geq -\mathbb{E}[M^c]$ *and* ***R*** *otherwise.*

With limited data, it is not guaranteed that the DM is able to perfectly learn the parameters required for the optimal rule. However, the notion of optimal rules is useful for determining the scenarios in which we can expect polarization to arise. The next section uses the notion of optimal rules to formalize the concept of polarization and to establish the main prediction of the theory.

Polarization

To study polarization, we take the definition of Haghtalab et al. [2021]: two DMs are said to be polarized if, when observing the same realization of observables, they predict the state to be different. That is, they both observe the same draw of $x = (x_1, ..., x_N)$ but one DM predicts the state to be **B** and the other predicts the state to be **R**.

Within out framework—as in Haghtalab2021—polarization may persist even when DMs have access to infinite data because they might be using different models of the world in order to make their predictions. Nevertheless using different models does not necessarily imply that beliefs will always be polarized; it must also be the case that the realization of observables falls in a *disagreement zone*. By disagreement zone, we simply mean that if the two models that are being used for prediction are M_1 and M_2, the realization of the observables $x = (x_1, ..., x_N)$ must be such that M_1 makes a different prediction from M_2. The following 2-dimensional example illustrates why the use of different models is necessary but not sufficient for polarization to arise.

Example *The optimal rule for the complex model coincides with the truth, it predicts* ***B*** *if* $x_2 + 1 \geq x_1$. *In addition, there are two simple models: one that considers only* x_1*—call it* M_1*—and one that considers only* x_2*—call it* M_2. *The optimal rules for* M_1 *and* M_2 *are threshold rules that predict the state is* ***B*** *if the value of* x_i *is below (above) the threshold* $t_i := \mu_{-i} + k$ *and the predicted state is* ***R*** *otherwise. These prediction rules are illustrated in . As the figure makes clear, both rules agree on the predicted state for some values of* (x_1, x_2) *and disagree for other values. Theregions in which the rules make different predictions are the disagreement zones.*

For any pair of agents who observe the same realization of teh observables and use different models, it is possible to determine whether they will be polarized or not by looking at the predictions made by their models.

Experimental Design

Subjects for the experimet were recruited from the CESS lab subject pool at NYU. We recruited 30 undergraduate students who participated in this experiment in person at the CESS lab. The experiment was coded using o-tree CITATION and it consisted of two parts: Part 1 was meant to expose the subjects to the data generating process and allow them to learn how to predict; part 2 was meant to elicit the models of the world that subjects might have developed in part 1.

Upon arrival subjects were randomly placed in pairs, however, they were not informed of this fact. Part 1 consited of 20 rounds in which subjects observed the realizations of all of the observables and were asked to predict the state. We had 5 observable variables called variable 1, 2, 3, 4 and 5. The variables were independent and Normally distributed with means and variances as described in TABLE.

The state was determined by the following rule: $11x_1 + 6x_2 + 4.5x_3 + 2.4 * x_4 + k \geq 0$. The values of the coefficients were chosen so that all variables had a similar but unique level of informativeness when considered on their own. In this case, because all the variables are normally distributed, the information that they convey about the state is related to the variance of the modified normal $a_i x_i$. The constant k was chosen so that the state was **B** in 50% of the cases. The last column of TABLE shows the probability of the optimal prediction rule making a correct prediction when the model used considers only each variable by itself. This is the measure that we use to determine the informativeness of each variable and is reported in the last column of TABLE.

Variable	Mean	Variance	Modified Variable	Modified Variance	Informativeness
x_1	0	1	$11x_1$	11	.65
x_2	-10	2	$6x_2$	12	.67
x_3	5	3	$4.5x_3$	13.5	.69
x_4	-5	4	$2.4 * x_4$	12	.63
x_5	100	5	$0x_5$	0	0

In each round, subjects were asked to predict the state. They were given feedback on whether their prediction was correct or not. Both subjects within a pair were shown exactly the same realization of the variables and had to predict the same state. This allows us to determine whether the pair was polarized or not for that particular observation.

Part 2 started immediatly after part 1. In this part each subject was randomly allowed to use only a certain number of variables in order to make their prediction. Instead of showing them the realizations of all 5 variables, they were told that they can disclose up to K variables, where K was drawn at random from the set $\{1, 2, ..., 5\}$ at every round and it was drawn independently across subjects. We take the variables that they chose to reveal as the model that each subject using for predicting the state and use that model to detrmine whether the theory would predict the pair to be polarized or not. In addition, we use the random assignment of K to investigate whether the subjects are using siumple or complex models. If

allowing subjects to disclose an additional variable improves their prediction accuracy, then it must be that they are in fact using the additional information for prediciton. If, on the other hand, they do not perform better when they have access to more information, it must be that although they are disclosing an additional variable, they do not correctly incorporate the infomation into their prediciton rule. Thus, if we observe that the performance of our subjects does not improve when they are allowed to disclose $K+1$ variables, relative to their performance when they were allowed to disclose only K variables, their model of the world must be of size K or lower.

Data

Results

Conclusion

References

Appendix

Inline code example: 30

References

Nika Haghtalab, Matthew O. Jackson, and Ariel D. Procaccia. Belief polarization in a complex world: A learning theory perspective. *PNAS*, 118, 2021.