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

$$y = \begin{cases} \mathbf{B} & \text{if } a_1x_1 + ... + a_Nx_N + K \geq 0 \\ \mathbf{R} & \text{otherwise} \end{cases}$$

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, an agent 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 agent uses a map $M_t \rightarrow \{\mathbf{B}, \mathbf{R}\}$ 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 agent decides to consider in a given period, and the procedure by which the maping or the set are updated. We refer to the set of variables that are taken into consideration, M_t, as the agent's *mental model*, and assume that the mapping is linear.

A mental model is *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, $M = \mathcal{R}$. In addition, mental models can be ranked in terms of their simplicity by the number of relevant variables that they consider.

Whenever agents use simple models, they ignore some of the relevant information that is available to them. In contrast, whenever agents use complex models, they use all the relevant information available. This means that when agents with complex mental models are restricted in terms of how many variables they can consider, their ability to predict the state will be impaired. On the contrary, restricting the number of variables that an agent with a simple model can use, does not necessarily affect their ability to predict the state. In particular, if they are considering fewer variables than what the restriciton allows, their predictions should be as accurate as without the restriciton. These two effects are captured in predictions 1 and 1B respectively.

Prediction 1: If subjects use complex models, then allowing them to use an additional variable will always increase the frequency with which they make correct predictions.

Prediction 1B: If subjects use a simple mental model that considers K variables, allowing them to use $K + 1$ variables will not increase the frequency with which they make correct predictions.

Optimal Rules

Conditional of using a particular mental model, the most accurate prediction rule that the agent can use is one that predicts the state to be **B** whenever, given the observables under consideration, **B** is more likely. And it predicts **R** whenever **R** is more likely. Because we abstract away from the procedure through which the rule is updated, we will make predictions under the assumption that the agent 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. This provides an upper bound for how well any model can perform.

We refer to a prediction rule as an *Optimal Rule* if it is the rule that a fully rational agent 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 agent 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 and the parameters are identified.

An optimal rule conditional on model M is the rule that, when restricted to the variables being considered, does the best at predicting the state. In this case, the agent is learning only the parameters that pertain the variables in M plus a constant term. The following proposition illustrates exactly what the optimal prediction rules look like when the agent is able to learn perfectly.

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

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

With limited data, it is not guaranteed that the agent is able to perfectly learn the parameters required for the optimal rule. However, the notion of optimal rules is useful for determining a benchmark for how well each model can perform, as well as 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

We take the definition of polarization from Haghtalab et al. [2021]: two agents 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 agent predicts the state to be $\mathbf{B}$ and the other predicts the state to be $\mathbf{R}$.

Within out framework—as in Haghtalab et al. [2021]—polarization may persist even when agents have access to infinite data because they might be using different models of the world. 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* $\boldsymbol{B}$ *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* $\boldsymbol{B}$ *if the value of* x_i *is below (above) the threshold* $t_i := \mu_{-i} + k$ *and the predicted state is* $\boldsymbol{R}$ *otherwise. These prediction rules are illustrated in . As Figure 1 makes clear, both rules agree on the predicted state for some values of* (x_1, x_2) *and disagree for other values. The regions in which the rules make different predictions are the disagreement zones.*

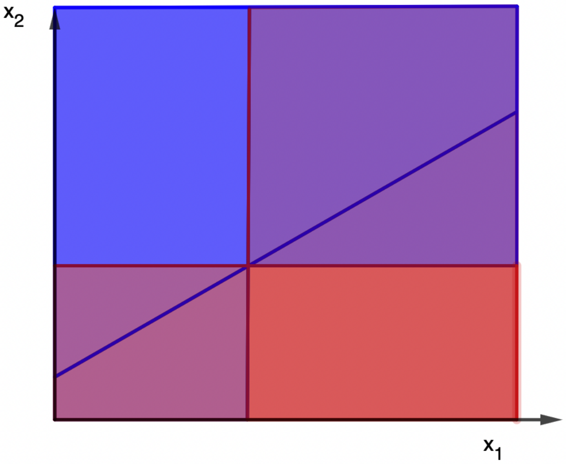

Figure 1: Polarization regions for two single variable models

For any pair of agents who observe the same realization of the observables and use different models, it is possible to determine whether they will be polarized or not by looking at the predictions made by the optimal rule corresponding to their models. Notice that the polarizing observations—those that fall into a polarization zone—are different for each pair of models. Also notice that our definition of the polarization zone assumes that the agents are using optimal rules. This is a very strong assumption and will most likely not be the case that subjects in the experiment learn perfectly. This is a theoretical construct that allows us to formalize Prediction 2.

Prediction 2 : Polarization will arise more often when the theory predicts so. That is, when two agents who have different models of the world, are facing a realization of observables that falls into the polarization zone corresponding to 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 durinf the summer and fall of 2023. The experiment was coded using o-tree Chen et al. [2016] 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.

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 and has 70 rounds. 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 $\{1, 2, ..., 5\}$. K was redrawn independently every round and across subjects. 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. This feature design allows us to test prediciton 1 and 1B and determine the number of variables that subjects are using.

In order to explore the polarization prediciton, we use the fact that subjects are in fixed pairs throughout the experiment. Both subjects in a pair observe the same realization of all the variables in each round. Therefore, we can determine whether they are polarized or not by looking at whether they make the same prediction or not. We can then determine whether the theory predicts that they should be polarized or not by looking at the polarization zone corresponding to the models that they are using. If the draw of the variables falls into the polarization zone, then the theory predicts that they should be polarized. The models that we use for this prediction are the ones that include the variables that they chose to reveal in that round. It could be that although they are revealing certain variables, they are not using them for prediction, however, we do not have a good way to determine this. Instead, we use the variables that they reveal as a proxy for the variables that they are using for prediction and determine the polarization zones for those models.

Results

In this section we present the results from the experiment. We start by exploring the general learning patterns of the subjects as well as the model choices. We then look at the effect of allowing subjects to disclose an additional variable and lastly we look at the polarization results.

Learning

In the first part of the experiment, subjects were asked to predict the state given the realization of all the variables. Having access to all the information they could predict the state perfectly if they managed to learn the parameters of the hyperplane that determines the state. Figure 2 shows the share of correct predictions across rounds in part 1. We see that subjects are able to learn how to predict the state at a rate that is higher than random. They initially predict the state corectly about 50% of the time, which is consistent with random guessing. The share of correct guesses increases to 59% by the end of part 1. And this last number is significantly higher than 50% (p-value $= 3 \times 10^{-4}$). They continue to learn even after the first 20 rounds when we look at the rounds in which subjects had access to all the information. By the end of the experiment the rate of correct predictions is 59% which is still significantly higher than 50% (p-value $= 0$).

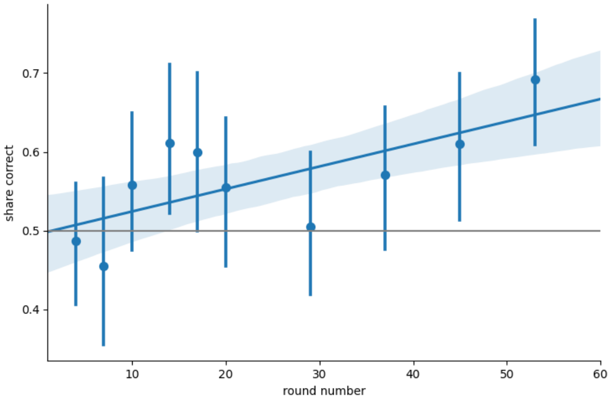

Figure 2: Share of correct predictions across rounds

Conclusion

Appendix

Daniel L. Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 3 2016. ISSN 2214-6350. doi: 10.1016/J.JBEF.2015.12.001.

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