Targeted Assistance

Bryce McLaughlin

Research Question: Will human decision-makers shift their decisions more in response to a prediction when the are given selective access to it? Can this be leveraged to increase the prediction accuracy of a human-controlled decision system?

Motivation

The practices of designing a prediction tool for the uses of automation and assistance are often confused as being identical. When a prediction is used for automation it acts as a decision rule: turning observable inputs into a final decision which is then put into practice. In contrast, an assistive prediction attempts to provide information to an autonomous decision-maker in hopes that they will prevent the decision-maker from committing would-be errors without eroding their properly taken decisions. In this way a decision rule wants to act as a substitute for the information a decision-maker stores while the assistive prediction wishes to acts a complement. This distinction suggests that the optimal predictions (in terms of minimizing system loss) might differ in the information they represent.

Even when the two have the exact same structure, a function $s : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is the set of observables and $\mathcal{Y}$ the possible prediction outcomes, the optimal times to utilize the prediction differ. In automation $\mathcal{Y}$ need always be implemented, but in assistance additional information can be encoded through selecting prediction sending. If the prediction is only sent at the times the user is *at high risk of error*, the decision-maker should have less faith in their prior and rely more on the prediction to make their final decision.

Model

A **product** arrives with observables $x \in \mathcal{X}$. The observables determine the distribution from which the product's value $v \in \mathbb{R}$ is drawn. The realization of the products value v remains random upon observation of x due to unknown factors such as condition or quality. From x an **evaluator** must form an estimate of the value v, $\hat{v} \in \mathbb{R}$, which earns them a utility of $U = 1 - L(v, \hat{v})$ for some loss function $L : \mathbb{R}^2 \to [0, 1]$.

When the evaluator is not interferred with, they take a decision according to their prior belief, Θ, of possible valuations for products to minimize the expected loss,

$$\hat{v}_0 = \arg\min_{y \in \mathbb{R}} E_{\Theta}[L(v, y)|x].$$

When the evaluator is interferred with by receiving a data-driven prediction $s : \mathcal{X} \to \mathcal{Y}$, they form a posterior belief $\Theta|s$ and take a decision,

$$\hat{v}_s = \arg\min_{y \in \mathbb{R}} E_{\Theta|s}[L(v, y)|x].$$

Under Bayes rule, the distance between $\hat{v}_0$ and $\hat{v}_s$ for any given x will depend on the ratio of the evaluator's certainty in $\hat{v}_0$ to their belief in the accuracy of s for products with observables x.

Utilizing the same decision rule we can increase the evaluators prediction-taking by only offering them the advice in the cases they are most likely to commit an error. If a prediction is selectively offered, more

information is offered, specifically that the evaluator is likely to error. This requires access to Z, a dataset of past observations where the decision-maker was not interferred. Using Z we can form an estimate of the evaluator's loss under v_0, $\phi : \mathcal{X} \to [0, 1]$.

If we only send the evaluator signals for x such that $\phi(x)$ is in the top ϵ percentile, the signal begins to carry additional information about the accuracy of v_0. As $\epsilon \to 0$, the evaluator's confidence in v_0 deteriorates on the observables x for which they receive a signal s as they need to rationalize their past mistakes on individuals who 'resemble' the current product (according to ϕ). Thus, a strategic limitation of interference should cause greater prediction-taking from the evaluator when a signal is sent. If the evaluator performs adequately enough under scenarios for which they do not receive s and implement v_0, the overall expected loss of the evaluator should reduce.

Hypotheses

A. Participants take predictions more as their access to them becomes limited to instances where they are more likely to make a mistake.

B. System accuracy can be improved through restricting prediction access strategically.

C. The evaluator will not be to able make proper use of ϕ when it is provided directly.

Experiment

1. Participants are asked to evaluate the monthly rent on 10 different appartments based on photos as well as a feature profile including location, beds, bath, and sq footage. This information is used to build a prediction function $\phi : \mathcal{X} \to \mathbb{R}_+$ on the expected error of the participant.

2. Participants fill out an attention check and some questions on home ownership unrelated to any specific numbers.

3. Participants are asked to evaluate monthly rent on 10 more apparentments while being assigned to one of the following testing conditions:

    - *Always* - Participants always receive a prediction to help them estimate the rent
    - *Rand-50* - Participants receive a prediction to help them estimate the rent randomly with probability 0.5
    - *Strat-50* - Participants receive a prediction to help them if the are likely to make a mistake according to ϕ (in expectation half of cases)
    - *Rand-20* - Participants receive a prediction to help them estimate the rent randomly with probability 0.2
    - *Strat-20* - Participants receive a prediction to help them if the are likely to make a mistake according to ϕ (in expectation one-fifth of cases)
    - *Always-Phi* - Participants always receive a prediction to help them estimate the rent alongside ϕ.

4. Participants are debriefed and paid based on the MSE of their predictions.

Tests

A1. Evaluators in *Always* will predict further from s on the 'worst 50%' than evaluators in *Strat-50* will. Similarly, evaluators in the *Always* will predict further from s on their 'worst 20%' than evaluators in *Strat-20* will.

A2. Evaluators in *Rand-50* will predict further from s on the 'worst 50%' than evaluators in *Strat-50* will. Similarly, evaluators in the *Rand-20* will predict further from s on their 'worst 20%' than evaluators in *Strat-20* will.

B1. Evaluators in *Always* will perform worse than worse on average than evaluators in either *Strat-50* or *Strat-20*.

B2. Evaluators in *Rand-50* will perform worse on average than evaluators in *Strat-50*. Similarly, evaluators in *Rand-20* will perform worse than worse on average than evaluators in *Strat-20*.

C1. Evaluators in the *Always-Phi* condition will predict further from s on the 'worst 50%' than evaluators in *Strat-50* will. Similarly, evaluators in the *Always-Phi* condition will predict further from s on their 'worst 20%' less than evaluators in *Strat-20* will.

C2. Evaluators in the *Always-Phi* condition will perform worse on average than evaluators in either *Strat-50* or *Strat-20*.