Carnegie Mellon University

Bayesian Persuasion for Algorithmic Recourse

Keegan Harris keeganh@cmu.edu
Valerie Chen valeriechen@cmu.edu
Joon Kim joonkim@cmu.edu
Ameet Talwalkar talwalkar@cmu.edu
Hoda Heidari hheidari@cmu.edu
Zhiwei Steven Wu zstevenwu@cmu.edu

ML MACHINE LEARNING DEPARTMENT

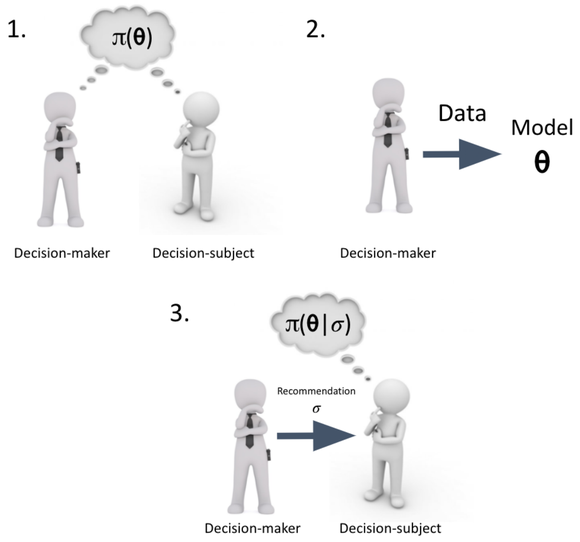
TL;DR

We use tools from Bayesian persuasion to design mechanisms for **incentivizing actionable interventions** which are **desirable to the decision-maker** in strategic-ML situations.

Strategic Responses to ML Models

When faced with automated decision-making tools, decision-subjects can take **actions** which modify their **observable features** and may therefore lead to different **outcomes**.

Bayesian Persuasion for Algorithmic Recourse



Goal: Design action recommendation policy to **maximize** chances decision-subject takes **desirable actions**.

This action recommendation policy must be **incentive-compatible**.

Feasibility of Computing Optimal Policy

Problem: Computing the optimal recommendation policy requires reasoning about all possible decision rules.

Key Observation: Space of all possible decision rules can be partitioned into a finite number of **equivalence regions**, based on the actions available to each decision-subject.

Decision-maker can then recommend actions based on these regions **w.l.o.g.**

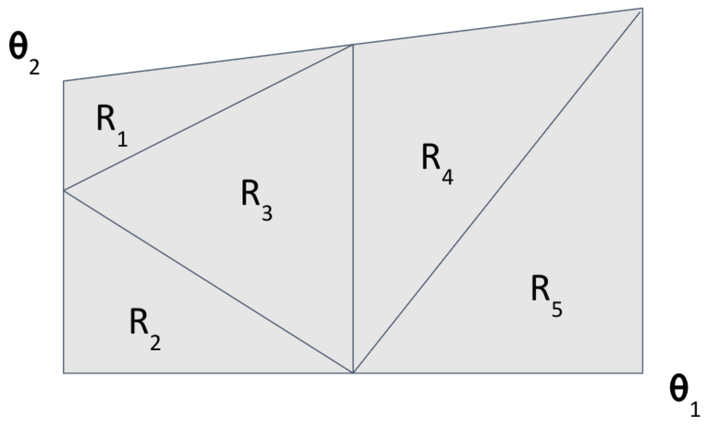


Figure 1. Example partitioning of the space of all possible decision rules. While there may possibly be an uncountably infinite number of possible decision rules, the number of regions will always be finite.

# of equivalence regions is possibly exponential, but can apply standard techniques to get a polynomial-time approximation w.h.p.

Optimal Recommendation Policy

pick action recommendation prob. in each region to
maximize decision-maker utility
s.t. recommendation policy is *incentive compatible*

Performance Guarantees

Optimal signaling policy

- is **never worse** than (i.) revealing model to decision subject or (ii.) providing no information about model to decision subject
- **can be arbitrarily better** than (i.) and (ii.)

Experiments

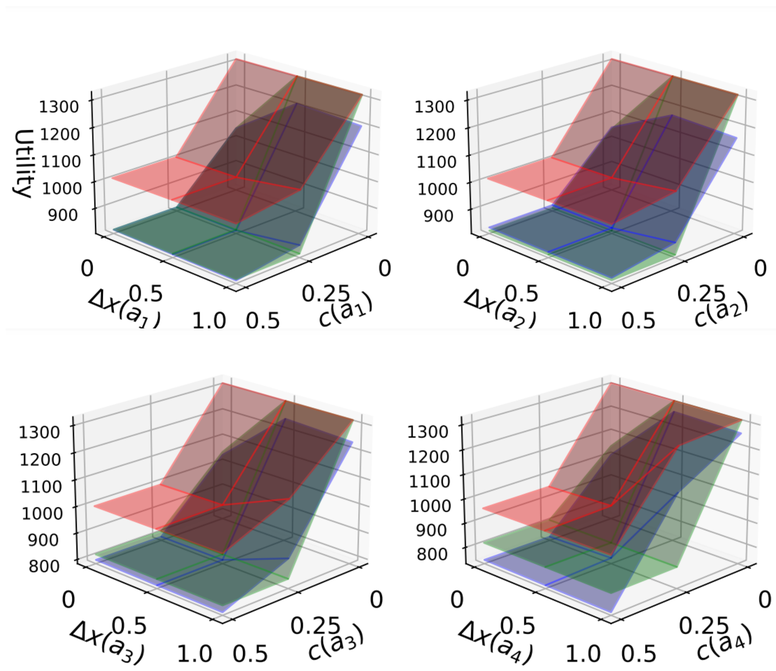


Figure 2. Expected utility across different $c(a)$ and $\Delta(a)$ configurations for $\sigma^2 = 0.4$. Optimal signaling policy (red) effectively upper-bounds the two baselines, revealing everything (blue) and revealing nothing (green) in all settings.

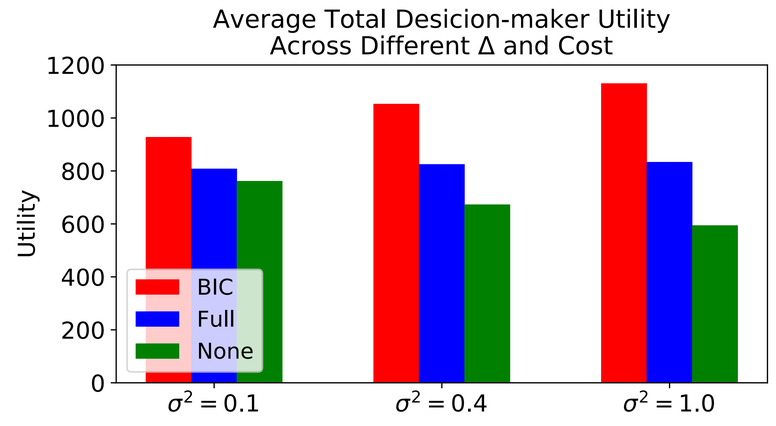


Figure 3. Total decision maker utility averaged across all cost and $\Delta(a)$ configurations for three different prior variances. The optimal signaling policy (red) consistently yields higher utility compared to the two baselines: revealing full information (blue) and no information (green). This gap increases when the decision subject is less certain about the model being used (higher σ^2).