DS 574 Algorithmic Mechanism Design **Lecture #21**
Fall 2023 Prof. Kira Goldner

Machine Learning and Incentives

ML algorithms for decision-making are almost everywhere nowadays. Examples from headlines and responses to them:

- Whether you qualify for a loan or not.
  - Increase your number of credit cards.
  - Increase you number of bank accounts.
  - Improve your credit history.
- Whether you qualify for probation.
- Whether you get invited for an onsite interview after a video screening call.
  - Dress a certain way.
  - Hide piercings/tattoos.
  - Change the way you talk.
- Whether students get admitted to college.
  - Improve their GPA.
  - Retake the GRE or pay for prep classes.
  - Change to a different school where they can be a higher rank.

Are these responses honest effort exertion or gaming?

Problem: If ML algorithms ignore this strategic behavior, they risk making policy decisions that are incompatible with the original policy's goal.

The goal of policy makers' and mechanism designers' in using ML for decision-making is to learn from human data to create better decisions.

However, those who have a stake in the outcome can also learn and manipulate their data.

Example: Strategic Classification

Consider a university trying to classify student applicants as qualified (positive) or unqualified (negative) for admission. A datapoint represents a student's features: SAT score, GPA, class ranking, etc.

Suppose the university solves this problem. Now, given this classifier (trained on training data), what will happen?

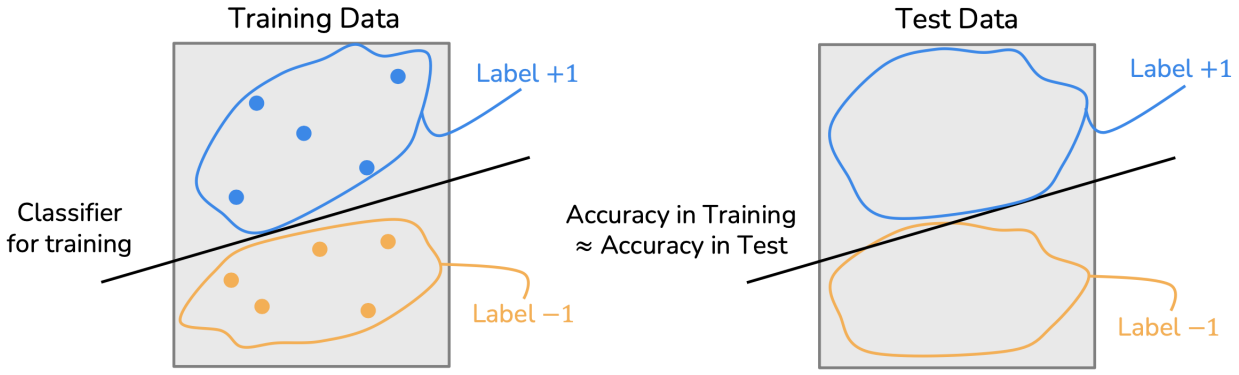

Figure 1: Example of Strategic Classification with students.

What will happen: Students who are near the border but below in the test set will be aware of the classifier and will manipulate their data points to be above the classifier (e.g., retake the SAT). If the classifier's goal is accuracy of the original data points, it is not succeeding.

Root of the problem: Data corresponds to individuals who have **agency** and want to affect the decisions made on them by the ML algorithms.

The Offline Model

Let's formalize the above example with a university and student applicants into a *game*, as was first done by Hardt et al. [2016]. This will be a Stackelberg game, where the main player we worry about goes first and acts in a way that is defensive against all possible best-responses.

The players:

- University: Their objective is to admit the most qualified candidates (accuracy). Their action is to produce a linear classifier.
- Individual students: Their objective is to be admitted. Their actions are to strategically change features.

The game:

1. Nature draws each agent's features (e.g., SAT score, class ranking, ...) $x \in \mathcal{X}$ from distribution $\mathcal{D}$.
2. The learner commits to classifier $\alpha \in \mathcal{A} : \mathcal{X} \to \{-1, +1\}$.
3. An agent observes the classifier α and the x.
4. An agent reports to learner feature vector $\Delta(x)$ (see below—$\neq x$).
5. The learner observes label $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. The learner gets utility: $\Pr_{x\sim\mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

Let $\Delta(x) = \arg\max_{y\in\mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$ where $c(x, y)$ is the manipulation cost and we make a crucial assumption that it is *separable*, e.g., $c(x, y) = \max\{0, c_2(y) - c_1(x)\}$. Let $\alpha(y)$ be the value for passing the classifier.

The learner's goal is to select the best classifier given strategic responses, that is, to compute a Stackelberg Equilibrium:

$$\alpha^* = \arg\max_{f\in\mathcal{H}} \Pr_{x\sim\mathcal{D}}[h(x) = f(\Delta(x))]$$

The main result of Hardt et al. [2016] is that the algorithm learns α^* in poly time and sample complexity.

Follow up work by Zrnic et al. [2021] shows that the order of play is crucial: they study when it's determined by how fast the principal and the agent adapt to each other, and that the agent's equilibria may be favorable for both.

The Online Model

We'll now look at the model adapted for a *dynamic* or an *online* learning setting. The players are the same. The game is updated slightly.

For round $t \in [T]$:

1. Nature chooses an agent's features (e.g., SAT score, class ranking, ...) $x_t \in \mathcal{X} \subseteq [0, 1]^d$.
2. The learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1, 1]^{d+1}$.
3. The agent observes the classifier α_t and the datapoint (x_t, y_t) where $y_t \in \{-1, 1\}$.
4. The agent reports to the learner a feature vector $\hat{x}_t(\alpha_t)$ $(\neq x_t)$
5. The learner observes true label y_t.
6. The learner incurs classification loss $\ell(\alpha_t, \hat{x}_t(\alpha_t))$.

The learner's goal is to minimize Stackelberg Regret:

$$R(T) = \sum_{t=1}^{T} \ell(\alpha_t, \hat{x}_t(\alpha_t)) - \min_{\alpha^* \in \mathcal{A}} \sum_{t=1}^{T} \ell(\alpha^*, \hat{x}_t(\alpha^*))$$

Assumptions are critical:

- What does the cost function to manipulate the agent's data look like?
- What does the learner's loss function look like? Binary? [Chen et al., 2020, Ahmadi et al., 2021] Hinge? Logistic? [Dong et al., 2018]

Big critiques:

- This is only for gaming, not honest effort.
- Agents have exact knowledge of the classifier. In reality, most classifiers are not exactly known, and we usually only have sample or probe access.

Fairness with Heterogenous Populations

So far we've been imagining that all students come from the same population, but that's obviously not true in reality. Two concurrent papers examine this problem when there are two populations: an *advantaged* population A and a *disadvantaged* population B.

We'll assume that the two populations are drawn from two different unknown distributions, which might even be the same. That is, both populations have qualified individuals.

For the advantaged population A, they have more access to tools to manipulate their feature vector, i.e., SAT prep classes and retakes. That is, it is less *costly* for those in A to change from x to $\hat{x}$ than it for those in B: $c_A(x, y) \leq c_B(x, y)$.

Main question of Hu et al. [2019]: What if you subsidize the disadvantaged population by some β (fractionally *or* additively) so that it's less expensive for the population to manipulate? The learner pays for the subsidies, doesn't want false positives to be disproportionate between the two groups, but their primary objective is to maximize accuracy.

Result: There are examples where subsidies are not Pareto-improving for both groups A and B. That means that there exist individuals in both groups who are worse off after the subsidies.

Milli et al. [2019] shows a necessary trade-off between agent utility and learner utility (accuracy), and that this disproportionately impacts the disadvantaged population.

Acknowledgements

This lecture was developed by using materials from Chara Podimata, and in particular, her 2021 tutorial on Incentive-Aware Machine Learning at ACM STOC.

References

Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The Social Cost of Strategic Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34:15257–15269, 2021.