DS 574 **Lecture #21 Notes**
Fall 2023 Prof. Kira Goldner

Strat Class

- Loan:
  - Increase your number of credit cards.
  - Increase you number of bank accounts.
  - Improve your credit history.
- Interview:
  - Dress a certain way.
  - Hide piercings/tattoos.
  - Change the way you talk.
- Admission:
  - Improve their GPA.
  - Retake the GRE or pay for prep classes.
  - Change to a different school where they can be a higher rank.

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 point, it is not succeeding.

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:

a. Nature draws each agent's features (e.g., SAT score, class ranking, ...) $x \in \mathcal{X}$ from distribution $\mathcal{D}$.

b. The learner commits to classifier $\alpha \in \mathcal{A} : \mathcal{X} \to \{-1, +1\}$.

c. An agent observes the classifier α and the x.

d. An agent reports to learner feature vector $\Delta(x)$ (see below—$\neq x$).

e. The learner observes label $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.

f. 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.

Stackelberg Eq:

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

Stackelberg Regret:

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