Strategic classification as an adversarial attack setting

Background: Strategic classification is an online classification problem in which the data is generated by strategic agents who manipulate their features aiming to change the classification outcome. In rounds, the learner deploys a classifier, then an adversarially chosen agent manipulates some data samples to optimally respond to the learner's choice of classifier. Such settings arise when machine learning models are used to make important decisions about the welfare (employment, education, health) of strategic individuals. Knowing information about the classifier, such individuals may manipulate their attributes to obtain a better classification outcome. Relevant papers: “Strategic classification from revealed preferences”, “Strategic classification”.
Project Description: In this project, we consider strategic classification as an adversarial attack setting and aim to utilize adversarial defense methods to improve the performance of strategic classification learners.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)