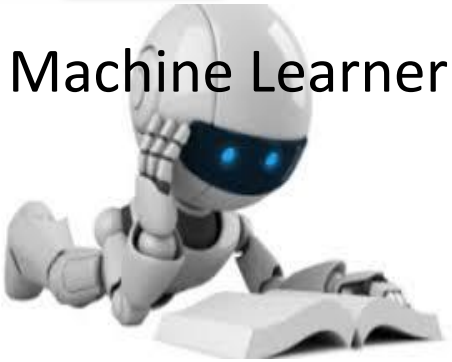
Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners

Shike Mei and Xiaojin Zhu

We play "white hat" hackers!

Optimally Poison

Mislead



Machine Learner

To

Specific Wrong Model

Bilevel Optimal Framework

$$\min_{D \in \mathbb{D}, \hat{\theta}_D} \quad O_A(D, \hat{\theta}_D)$$

Upper-level: attacker

$$\text{s.t.} \quad \hat{\theta}_D \in \operatorname{argmin}_{\theta \in \Theta} O_L(D, \theta)$$
$$\text{s.t. } \mathrm{g}(\theta) \leq 0, \ \mathrm{h}(\theta) = 0.$$

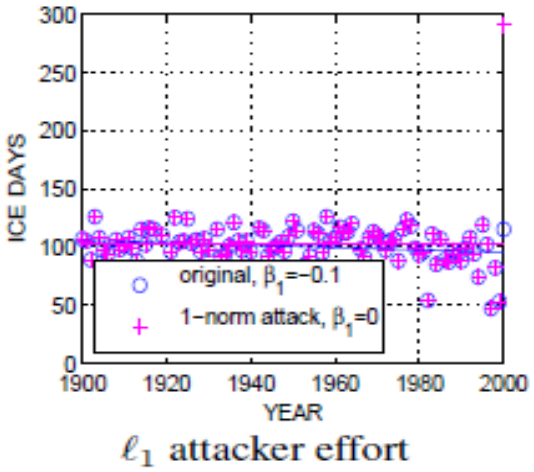
Lower-level: learner

Solved by KKT conditions and implicit functions

E.g.,

Learner: learn the trend of #frozen days of Lake Mendota.

Attacker: hide the lake warming trend with minimal modification on data.



ℓ_1 attacker effort