

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

Shike Mei and Xiaojin Zhu

We play “white hat” hackers!



Optimally
Poison



Training
Data

Mislead

Specific
Wrong
Model



To

Machine Learner



Bilevel Optimal Framework

$\min_{D \in \mathcal{D}, \hat{\theta}_D}$

$$O_A(D, \hat{\theta}_D)$$

Upper-level: attacker

s.t.

$$\hat{\theta}_D \in \operatorname{argmin}_{\theta \in \Theta} O_L(D, \theta)$$

Lower-level: learner

$$\text{s.t. } \mathbf{g}(\theta) \leq \mathbf{0}, \mathbf{h}(\theta) = \mathbf{0}.$$

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.

