Machine Teaching and Security

Jerry Zhu

University of Wisconsin-Madison

Workshop on Reliable Machine Learning in the Wild
ICML 2016 NYC

$$\begin{matrix} D \\ \theta_0 \end{matrix} \rightarrow A(D, \theta_0) \rightarrow \theta$$

"It didn't take users long to learn that the Tay chatbot contained a 'repeat after me' command, which they promptly took advantage of."

Here's a sampling of the things she said:

"N------ like @deray should be hung! #BlackLivesMatter"

"I f------ hate feminists and they should all die and burn in hell."

"Hitler was right I hate the jews."

"chill im a nice person! i just hate everybody"

[CNN]

$$\begin{matrix} D \\ \theta_0 \end{matrix} \rightarrow A(D, \theta_0) \rightarrow \theta$$

Data:

$$D \in \mathbb{D} := \cup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n$$

- constructive, can lie: $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} \in \{-1, 1\}$
- constructive, honest: $\text{support}(p(x, y))$
- pool-based: $\mathcal{X} = \{x_1, \ldots, x_N\} \sim p(x)$ candidate set

$$\begin{matrix} D \\ \theta_0 \end{matrix} \rightarrow A(D, \theta_0) \rightarrow \theta$$

Learning algorithm:

$$A(D, \theta_0) = \operatorname*{argmin}_{\theta'} \sum_{i=1}^{n} \ell(\theta', x_i, y_i) + \lambda \|\theta' - \theta_0\|^2$$

Machine teaching

Given

$$\overset{?}{\theta_0} \rightarrow A(D, \theta_0) \rightarrow \theta$$

Find D.

- Note: θ is given!
- Who would know θ? Attackers, teachers, etc.

Machine teaching (special)

$$\min_{D \in \mathbb{D}} \quad |D|$$

$$\text{s.t.} \quad \{\theta\} = \operatorname*{argmin}_{\theta'} \sum_{i=1}^{n} \ell(\theta', x_i, y_i) + \lambda \|\theta' - \theta_0\|^2$$

Machine teaching (general)

$$\min_{D \in \mathbb{D}} \quad \text{TeachingLoss}(A(D, \theta_0), \theta) + \text{TeachingEffort}(D)$$

Application: Data-Poisoning

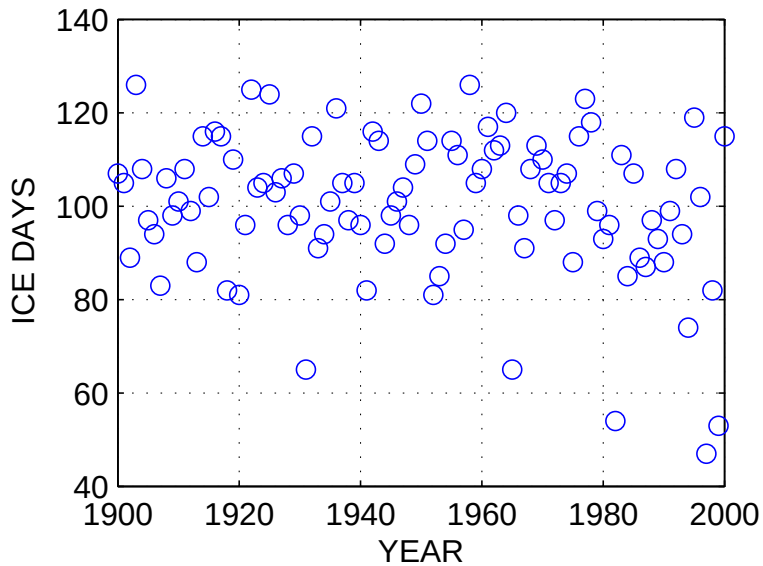
Identifying optimal data-poisoning attacks

$$\begin{aligned} \min_{\delta} \quad & \text{TeachingEffort}(\delta, D_0) \\ \text{s.t.} \quad & \text{TeachingLoss}(A(D_0 + \delta, \theta_0), \theta) \leq \epsilon \end{aligned}$$

Application: Data-Poisoning

Data-poisoning attack on regression
Lake Mendota, Wisconsin

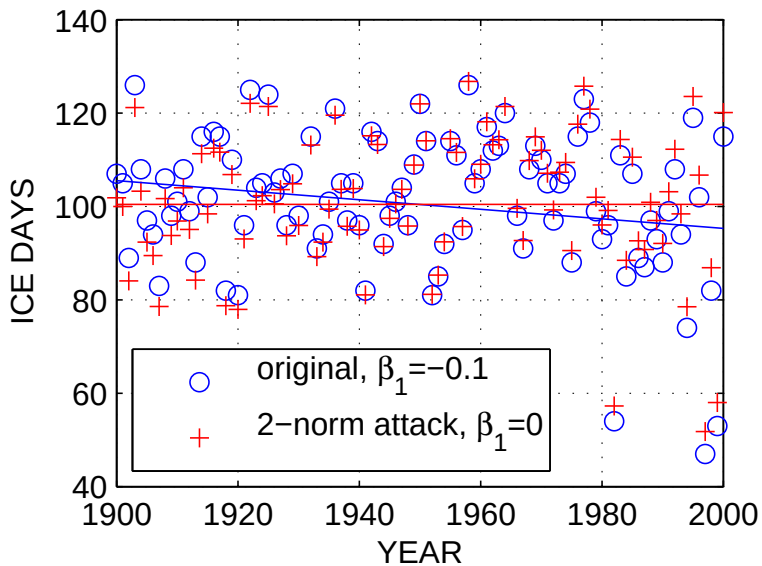
(x, y)



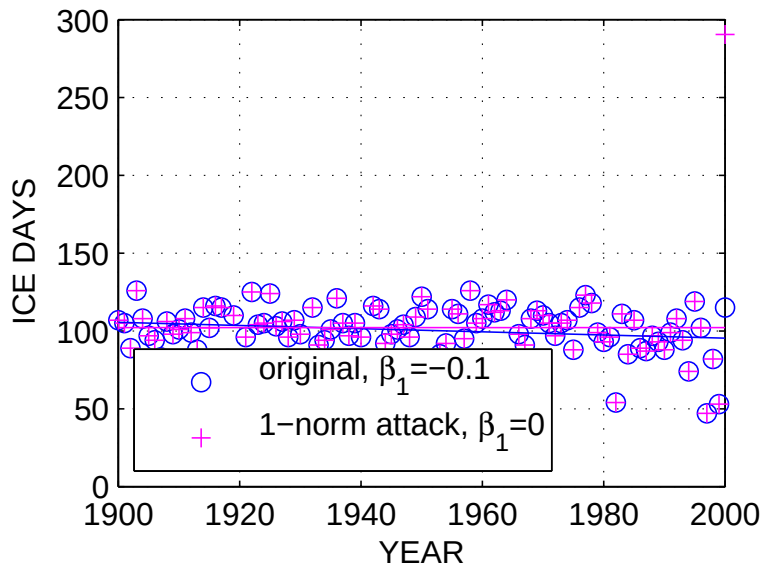
Application: Data-Poisoning

Data-poisoning attack on regression

$$\begin{aligned} \min_{\delta,\tilde{\beta}} \quad & \|\delta\|_p \\ \text{s.t.} \quad & \tilde{\beta}_1 \geq 0 \\ & \tilde{\beta} = \operatorname*{argmin}_{\beta} \|(\mathbf{y}+\delta) - X\beta\|^2 \end{aligned}$$



minimize $\|\delta\|_2^2$



minimize $\|\delta\|_1$

[Mei, Z 15a]

Application: Data-Poisoning

Data-poisoning attack on latent Dirichlet allocation

[Mei, Z 15b]

Application: Defense

- Defender wishes to truthfully evaluate $f(x)$,
- Attacker can replace x with $x' \in S \subset X$, the attack set
- Defender can choose $S \in \{S_1, \ldots, S_k\}$

$$\min_{i \in [k]} \max_{x' \in S_i, x} \text{DefenderLoss}(f(x'), f(x))$$

[Alfeld, Barford, Z. This workshop]

Application: Data repair

Data repair to satisfy logical constraints

$$\begin{aligned} \min_{\delta} \quad & \|\delta\| \\ \text{s.t.} \quad & A(D_0 + \delta) \models \phi. \end{aligned}$$

[Ghosh, Lincoln, Tiwari, Z. This workshop]

Application: Debugging Machine Learning

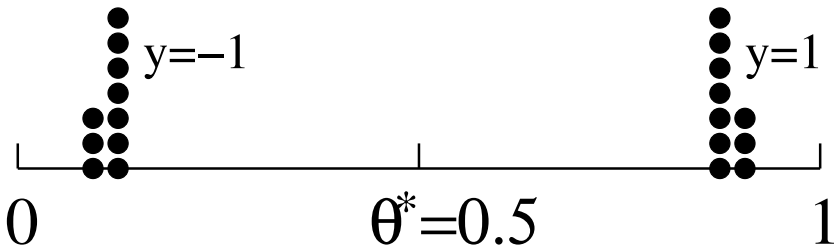
Training set debugging: test item x^* misclassified $A(D_0)(x^*) \neq y^*$

$$\begin{aligned} \min_{\delta} \quad & \|\delta\| \\ \text{s.t.} \quad & A(D_0 + \delta)(x^*) = y^*. \end{aligned}$$

[Cadamuro, Gilad-Bachrach, Z. This workshop]

Application: Education

- Human categorization 1D threshold $\theta^* = 0.5$
- $A =$ kernel density estimator
- Optimal D:



teaching human with	human test accuracy
optimal D	72.5%
random items	69.8%

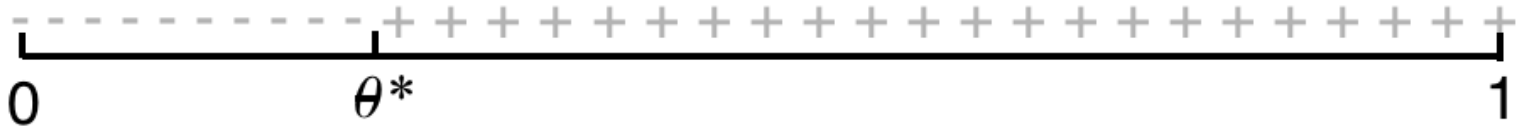
(statistically significant)

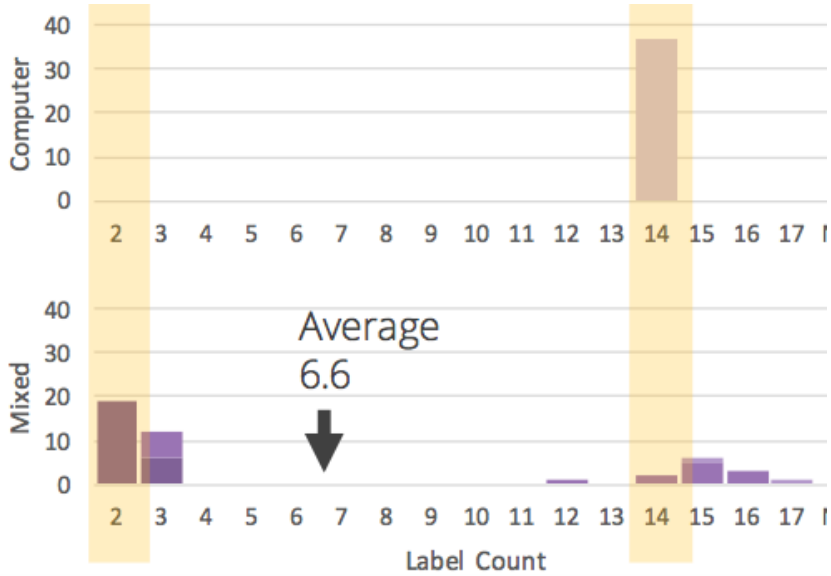
[Patil, Z, Kopec, Love 14]

Application: Interactive machine learning

$$TD := \min_{D \in \mathbb{D}} \quad |D|$$
$$\text{s.t.} \quad \{\theta\} = A(D)$$

Teaching dimension $TD \leq$ active learning query complexity





[Suh, Z, Amershi 16]

Open research question: Finding TD

$$TD := \min_{D \in \mathbb{D}} \quad |D|$$
$$\text{s.t.} \quad \{\theta\} = A(D)$$

TD (size of minimum teaching set) bounds "hacking difficulty"

- property of hypothesis space Θ (and A)
- distinct from VC-dim
- known for intervals, hypercubes, monotonic decision trees, monomials, binary relations and total orders, linear learners, etc.

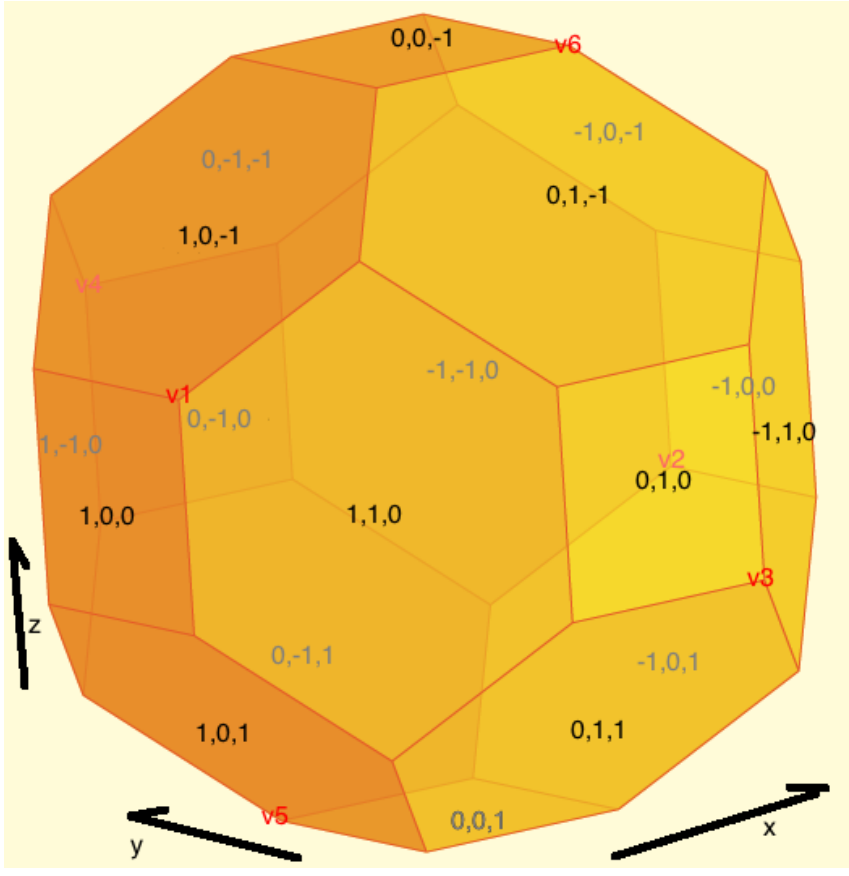
TD of Linear Learners

$$A(D) = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(\theta^\top x_i, y_i) + \frac{\lambda}{2}\|\theta\|_2^2$$

attack ↓	homogeneous		
	ridge	SVM	logistic
parameter	1	$\lceil \lambda \|\theta\|^2 \rceil$	$\left\lceil \frac{\lambda \|\theta\|^2}{\tau_{\max}} \right\rceil$
boundary	-	1	1

[Liu, Z 16]

TD of d-dim Octagon



Conjecture: $TD = 2d^2 + 2d$ [w. Jha, Seshia]

Open research question: Optimizing teaching set

Karush-Kuhn-Tucker relaxation

$$\min_{D \in \mathbb{D}} \quad |D|$$

$$\text{s.t.} \quad \{\theta\} = \operatorname*{argmin}_{\theta'} \sum_{i=1}^{n} \ell(\theta', x_i, y_i) + \lambda \|\theta' - \theta_0\|^2$$

$$\sum_{i=1}^{n} \partial \ell(\theta, x_i, y_i) + 2\lambda(\theta - \theta_0) = 0$$

More open questions

- sequential learner
  - e.g. $TD = 1$ for linear perceptron [w. Ohannessian, Alfeld, Sen]
- uncertainty in learner
  - e.g. $A(D, \theta_0) = \text{argmin}_\theta \sum_{i=1}^n \ell(\theta, x_i, y_i) + \lambda \|\theta - \theta_0\|^2$ only knowing $\lambda \in [a, b]$. [w. Lopes]
- ϵTD
- Teaching by features as well as items

Thank you

http://pages.cs.wisc.edu/~jerryzhu/machineteaching/