# Machine Teaching
for
Personalized Education, Security, Interactive Machine Learning

Jerry Zhu

NIPS 2015 Workshop on Machine Learning from and for Adaptive User Technologies

# Supervised Learning Review

- $D$: training set $(x_1, y_1) \dots (x_n, y_n)$
- Learning algorithm $A : A(D) = \{\theta\}$
  - Example: regularized empirical risk minimization

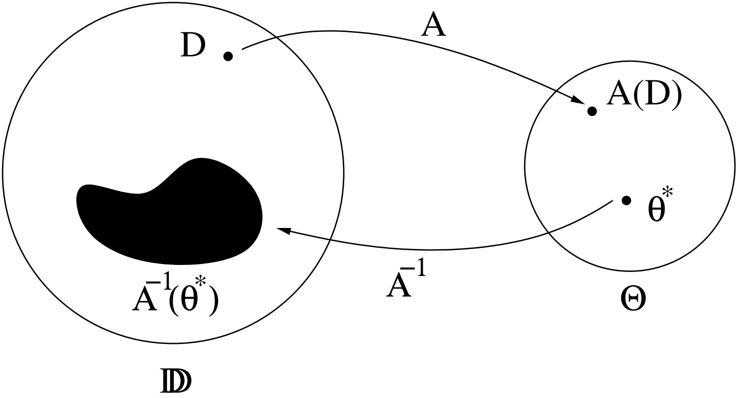
$$A(D) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \ell(\theta, x_i, y_i) + \lambda\Omega(\theta)$$

  - Example: version space learner

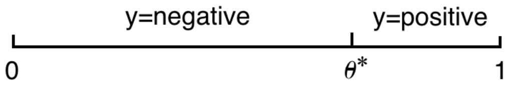
$$A(D) = \{\theta \text{ consistent with } D\}$$

# Machine Teaching = Inverse Machine Learning

- Given: learning algorithm $A$, target model $\theta^*$
- Find: the smallest $D$ such that $A(D) = \{\theta^*\}$

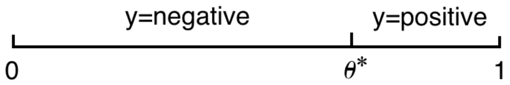


# Machine Teaching Example

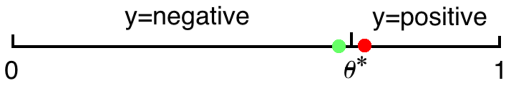



- Teach a 1D threshold classifier
- Given: $A =$ SVM, $\theta^* = 0.6$
- What is the smallest $D$?

# Machine Teaching Example




- Teach a 1D threshold classifier
- Given: $A =$ SVM, $\theta^* = 0.6$
- What is the smallest $D$?



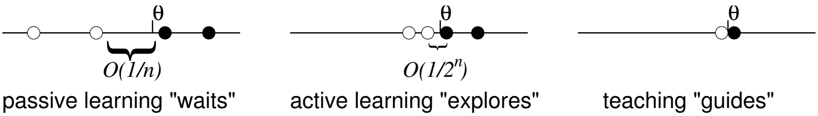
# Machine Teaching as Communication

- sender $=$ teacher
- message $= \theta^*$
- decoder $= A$
- codeword $= D$

# Machine Teaching "Stronger" Than Active Learning

Sample complexity to achieve $\epsilon$ error:

- passive learning $n = O(1/\epsilon))$
- active learning $n = O(\log(1/\epsilon))$: needs binary search
- machine teaching $n = 2$: teaching dimension [Goldman + Kearns 1995], the teacher knows $\theta^*$



# Why Machine Teaching?

Why bother if we already know $\theta^*$?

1. education
2. computer security
3. interactive machine learning

# Education

Teacher answers three questions:

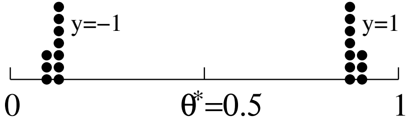
1. My student's cognitive model $A$ is a ___
   (a) SVM (b) logistic regression (c) neural net ...
2. Student success is defined by ___
   (a) learning a target model $\theta^*$ (b) excel on a test set
3. My teaching effort is defined by ___
   (a) training set size (b) training item cost ...

Machine teaching finds the optimal, personalized lesson $D$ for the student.

# Education Example

[Patil et al. 2014]

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




- Teaching humans with $D$:

| human trained on | human test accuracy |
| --- | --- |
| optimal $D$ | 72.5% |
| random items | 69.8% |

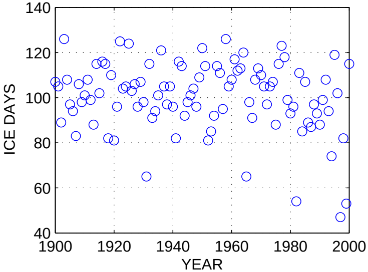
(statistically significant)

# Security: Data Poisoning Attack

[Alfeld et al. 2016], [Mei + Zhu 2015]

- Given: learner $A$, attack target $\theta^*$, clean training data $D_0$
- Find: the minimum "poison" $\delta$ such that $A(D_0 + \delta) = \{\theta^*\}$

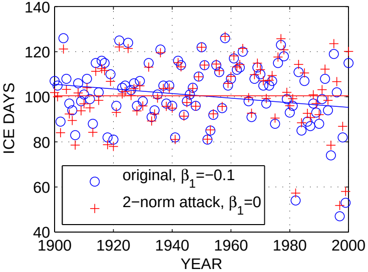
# Security: Lake Mendota Data



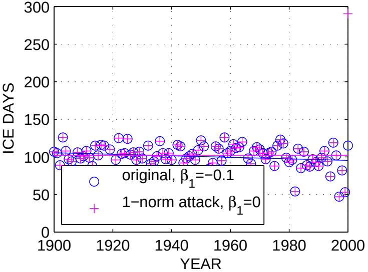

Lake Mendota, Wisconsin ice days

# Security: Optimal Attacks on Ordinary Least Squares

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




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




minimize $\|\delta\|_1$

# Security: Optimal Attack on Latent Dirichlet Allocation

[Mei + Zhu 2015]

# Optimization for Machine Teaching

Bilevel optimization

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

$$\text{s.t.} \quad \theta^* = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{|D|} \sum_{(x_i, y_i) \in D} \ell(x_i, y_i, \theta) + \Omega(\theta)$$

In general

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

- Sometimes closed-form solution [Alfeld et al. 2016] [Zhu 2013]
- Often nonconvex optimization [Bo Zhang, this workshop]

# Theoretical Question: Teaching Dimension

- Given: $A, \theta^*$
- Find: the smallest $D$ such that $A(D) = \{\theta^*\}$

Teaching dimension $TD = |D|$

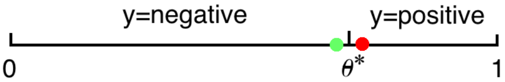
# Theoretical Question: Teaching Dimension

- Given: $A, \theta^*$
- Find: the smallest $D$ such that $A(D) = \{\theta^*\}$

Teaching dimension $TD = |D|$

- $A =$ version space learner: $TD = \infty$? [Goldman + Kearns 1995]
- $A = SVM$: $TD = 2$

Learner-specific teaching dimension

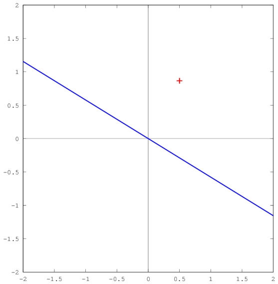


# Teaching Dimension Example: Teaching an SVM

- Learner
  $A(D) = \arg\min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{n} \max(1 - y_i \mathbf{x}_i^\top \theta, 0) + \frac{1}{2}\|\theta\|^2$
- Target model: $\theta^* = (\frac{1}{2}, \frac{\sqrt{3}}{2})^\top$
- One training item is necessary and sufficient:

$$x_1 = (\frac{1}{2}, \frac{\sqrt{3}}{2})^\top, \; y_1 = 1$$

- $TD(\theta^*, A) = 1$



# The Teaching Dimension of Linear Learners

[Liu & Zhu 2015]

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

| goal ↓ | homogeneous | | | inhomogeneous $\theta^* = [w^*; b^*]$ | | |
|---|---|---|---|---|---|---|
| | ridge | SVM | logistic | ridge | SVM | logistic |
| parameter | 1 | $\lceil \lambda \|\theta^*\|^2 \rceil$ | $\left\lceil \frac{\lambda \|\theta^*\|^2}{\tau_{\max}} \right\rceil$ | 2 | $2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2} \right\rceil^\dagger$ | $2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2\tau_{\max}} \right\rceil^\dagger$ |
| boundary | - | 1 | 1 | - | 2 | 2 |

# The Teaching Dimension of Linear Learners

[Liu & Zhu 2015]

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

| goal ↓ | homogeneous | | | inhomogeneous $\theta^* = [w^*; b^*]$ | | |
|---|---|---|---|---|---|---|
| | ridge | SVM | logistic | ridge | SVM | logistic |
| parameter | 1 | $\lceil \lambda \|\theta^*\|^2 \rceil$ | $\left\lceil \frac{\lambda \|\theta^*\|^2}{\tau_{\max}} \right\rceil$ | 2 | $2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2} \right\rceil^\dagger$ | $2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2\tau_{\max}} \right\rceil^\dagger$ |
| boundary | - | 1 | 1 | - | 2 | 2 |

Teaching Dimension (independent of $d$) distinct from VC-dimension $(d+1)$

# Interactive Machine Learning

- You train a classifier
- Only use item, label pairs $(x_1, y_1) \ldots (x_n, y_n)$
- Cost $= n$ to reach small $\epsilon$ error

# “Speed of Light”

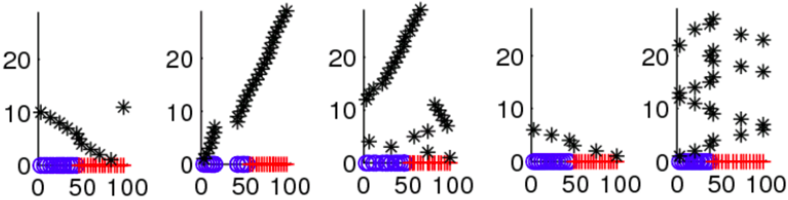
[Goldman+Kearns 1995], [Angluin 2004], [Cakmak+Thomaz 2011], [Zhu *et al.* in preparation]

Fundamental Cost of Interactive Machine Learning
$n \geq$ Teaching Dimension

- Optimal teacher achieves $n =$Teaching Dimension
- Can be much faster than active learning (recall 2 vs. $\log \frac{1}{\epsilon}$)
- Must allow teacher-initiated items (unlike active learning)
- But some humans are bad teachers ...

# Some Humans are Suboptimal Teachers

[Khan, Zhu, Mutlu 2011]




Challenge for the computer: make human behave more like the optimal teacher

# Summary

Machine teaching

- Theory: teaching dimension
- Algorithm: bilevel optimization
- Applications: education, security, interactive machine learning

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