Machine Teaching

An Inverse Problem to Machine Learning and an Approach Toward Optimal Education

Jerry Zhu
Department of Computer Sciences
University of Wisconsin-Madison

AAAI 2015

Example One

- Steve the student runs a linear SVM:

$$\text{Given a training set with } n \text{ items } \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

$$\text{Steve learns } \mathbf{w} \in \mathbb{R}^d$$

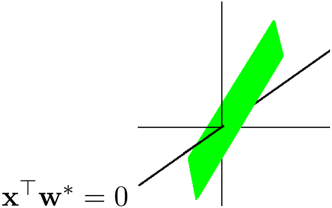
Example One

- Steve the student runs a linear SVM:

$$\text{Given a training set with } n \text{ items } \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

$$\text{Steve learns } \mathbf{w} \in \mathbb{R}^d$$

- Tina the teacher wants Steve to learn a target $\mathbf{w}^*$



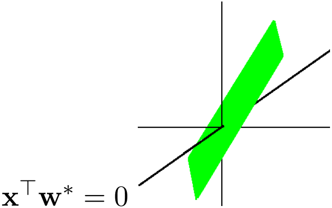
Example One

- Steve the student runs a linear SVM:

  Given a training set with n items $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$

  Steve learns $\mathbf{w} \in \mathbb{R}^d$

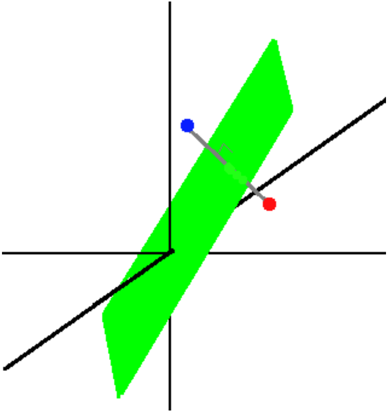
- Tina the teacher wants Steve to learn a target $\mathbf{w}^*$



- What is the smallest training set Tina can give Steve?

Example One

Tina's non-iid training set with $n = 2$ items



Example Two

- Steve estimates a Gaussian density:

$$\text{Given } \mathbf{x}_1 \ldots \mathbf{x}_n \in \mathbb{R}^d$$

$$\text{Steve learns } \hat{\mu} = \frac{1}{n} \sum \mathbf{x}_i, \quad \hat{\Sigma} = \frac{1}{n-1} \sum (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^\top$$

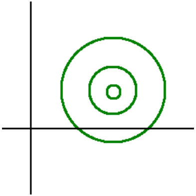
Example Two

- Steve estimates a Gaussian density:

$$\text{Given } \mathbf{x}_1 \ldots \mathbf{x}_n \in \mathbb{R}^d$$

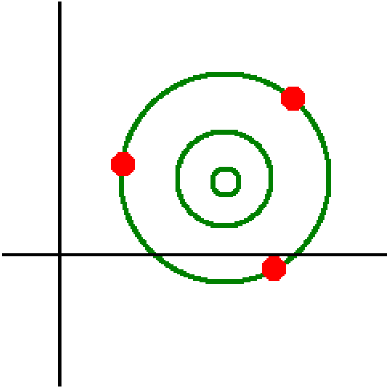
$$\text{Steve learns } \hat{\mu} = \frac{1}{n} \sum \mathbf{x}_i, \quad \hat{\Sigma} = \frac{1}{n-1} \sum (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^\top$$

- Tina wants Steve to learn a target Gaussian with (μ^*, Σ^*)

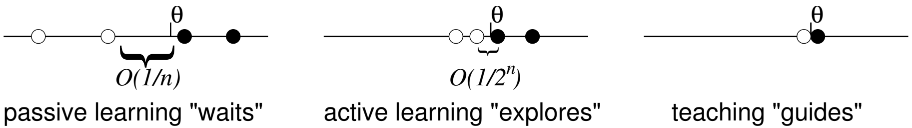


Example Two

Tina's minimal training set of $n = d + 1$ tetrahedron vertices



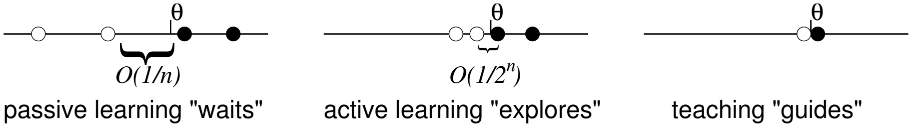
Machine Teaching More Powerful Than Active Learning



Sample complexity to achieve ϵ error

- passive learning $1/\epsilon$

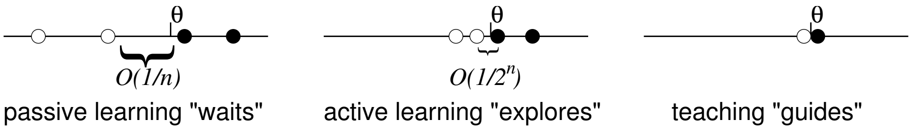
Machine Teaching More Powerful Than Active Learning



Sample complexity to achieve ϵ error

- passive learning $1/\epsilon$
- active learning $\log(1/\epsilon)$

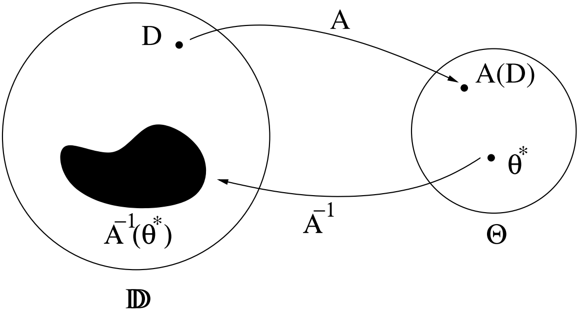
Machine Teaching More Powerful Than Active Learning



Sample complexity to achieve ϵ error

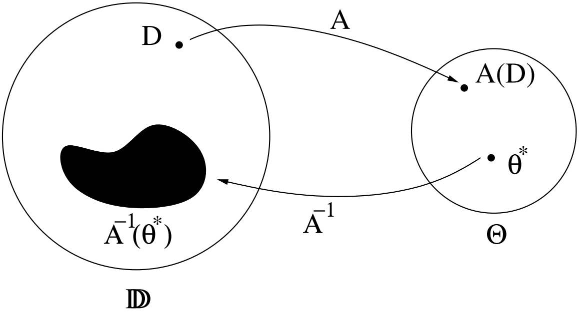
- passive learning $1/\epsilon$
- active learning $\log(1/\epsilon)$
- machine teaching 2: Tina knows θ

Machine Teaching



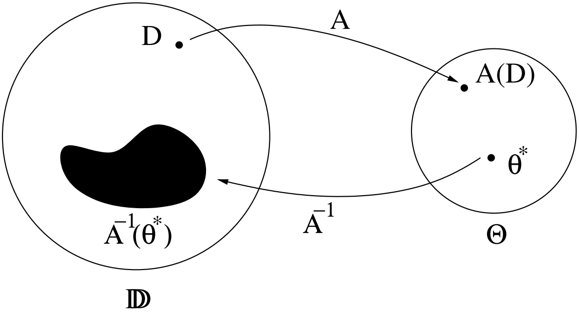
- Tina wants Steve to learn a target model θ^*

Machine Teaching



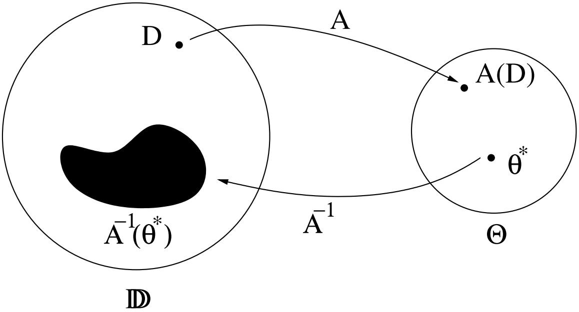
- Tina wants Steve to learn a target model θ^*
  - not machine learning: Tina already knows θ^*

Machine Teaching



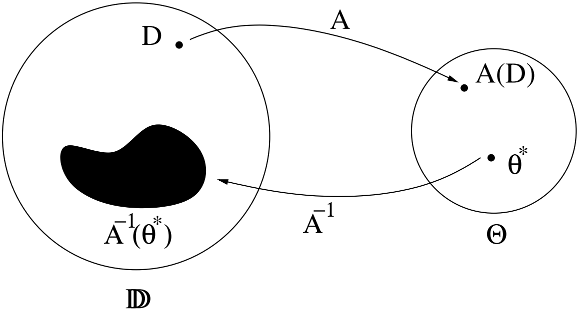
- Tina wants Steve to learn a target model θ^*
  - not machine learning: Tina already knows θ^*
- Tina knows Steve's learning algorithm A

Machine Teaching



- Tina wants Steve to learn a target model θ^*
  - not machine learning: Tina already knows θ^*
- Tina knows Steve's learning algorithm A
- Tina seeks the best training set within $A^{-1}(\theta^*)$ for Steve

Machine Teaching



- Tina wants Steve to learn a target model θ^*
  - not machine learning: Tina already knows θ^*
- Tina knows Steve's learning algorithm A
- Tina seeks the best training set within $A^{-1}(\theta^*)$ for Steve
  - best = the smallest (Teaching Dimension [Goldman Kearns 1995]), or other criteria

Machine Teaching Harder Than Machine Learning

Special case: teaching the exact parameters, minimizing training set size

$$\min_{D \in \mathbb{D}} \quad |D| \qquad \text{Tina's problem}$$

$$\text{s.t.} \quad \theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta) \qquad \text{Steve's algorithm } A$$

Bilevel optimization, NP-hard in general

General Machine Teaching Framework

[poster 819 Wednesday]

$$\min_{D \in \mathbb{D}, \hat{\theta} \in \Theta} \quad R_T(\hat{\theta}) + \lambda E_T(D)$$
$$\text{s.t.} \quad \hat{\theta} = A(D)$$

- $R_T()$: teaching risk function e.g. $\|\hat{\theta} - \theta^*\|_2^2$

General Machine Teaching Framework

[poster 819 Wednesday]

$$\min_{D \in \mathbb{D}, \hat{\theta} \in \Theta} \quad R_T(\hat{\theta}) + \lambda E_T(D)$$
$$\text{s.t.} \quad \hat{\theta} = A(D)$$

- $R_T()$: teaching risk function e.g. $\|\hat{\theta} - \theta^*\|_2^2$
- $E_T()$: teaching effort function e.g. different item costs

General Machine Teaching Framework

[poster 819 Wednesday]

$$\min_{D \in \mathbb{D}, \hat{\theta} \in \Theta} \quad R_T(\hat{\theta}) + \lambda E_T(D)$$
$$\text{s.t.} \quad \hat{\theta} = A(D)$$

- $R_T()$: teaching risk function e.g. $\|\hat{\theta} - \theta^*\|_2^2$
- $E_T()$: teaching effort function e.g. different item costs
- Tina's search space $\mathbb{D}$: constructive or pool-based, batch or sequential

General Machine Teaching Framework

[poster 819 Wednesday]

$$\min_{D \in \mathbb{D}, \hat{\theta} \in \Theta} \quad R_T(\hat{\theta}) + \lambda E_T(D)$$
$$\text{s.t.} \quad \hat{\theta} = A(D)$$

- $R_T()$: teaching risk function e.g. $\|\hat{\theta} - \theta^*\|_2^2$
- $E_T()$: teaching effort function e.g. different item costs
- Tina's search space $\mathbb{D}$: constructive or pool-based, batch or sequential
- Tractable solutions when Steve runs linear regression, logistic regression, SVM, LDA, etc. [Mei Z 2015a, Mei Z 2015b]

Education

- Steve may actually be a human!

Education

- Steve may actually be a human!
- Tina computes the best lesson D for Steve

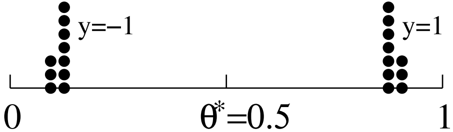
Education

- Steve may actually be a human!
- Tina computes the best lesson D for Steve
- Needs Steve's cognitive model A

Education

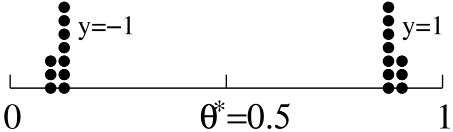
- Steve may actually be a human!
- Tina computes the best lesson D for Steve
- Needs Steve's cognitive model A
  - a "good enough" A is fine

A Case Study



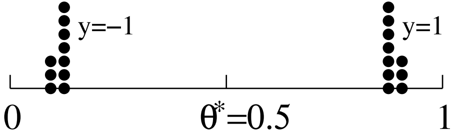
- Human categorization [Patil Z Kopeć Love 2014]

A Case Study



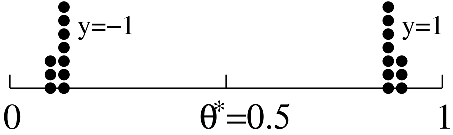
- Human categorization [Patil Z Kopeć Love 2014]
- A: a limited capacity retrieval cognitive model

A Case Study



- Human categorization [Patil Z Kopeć Love 2014]
- A: a limited capacity retrieval cognitive model

A Case Study



- Human categorization [Patil Z Kopeć Love 2014]
- A: a limited capacity retrieval cognitive model

human training set	human test accuracy
D	72.5%
iid	69.8%

(statistically significant)

Open Problems for the AI Community

- **Optimization**: solve Tina's problem for any Steve

Open Problems for the AI Community

- **Optimization**: solve Tina's problem for any Steve
- **Theory**: generalize Teaching Dimension [Goldman Kearns 1995]

Open Problems for the AI Community

- **Optimization**: solve Tina's problem for any Steve
- **Theory**: generalize Teaching Dimension [Goldman Kearns 1995]
- **Cognitive Science and Education**: Steve's A, real education tasks

Open Problems for the AI Community

- **Optimization**: solve Tina's problem for any Steve
- **Theory**: generalize Teaching Dimension [Goldman Kearns 1995]
- **Cognitive Science and Education**: Steve's A, real education tasks
- **Computer Security**: data poisoning attacks

Open Problems for the AI Community

- **Optimization**: solve Tina's problem for any Steve
- **Theory**: generalize Teaching Dimension [Goldman Kearns 1995]
- **Cognitive Science and Education**: Steve's A, real education tasks
- **Computer Security**: data poisoning attacks

Open Problems for the AI Community

- **Optimization**: solve Tina's problem for any Steve
- **Theory**: generalize Teaching Dimension [Goldman Kearns 1995]
- **Cognitive Science and Education**: Steve's A, real education tasks
- **Computer Security**: data poisoning attacks

References:

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