Machine Teaching

Jerry Zhu

Department of Computer Sciences
University of Wisconsin-Madison

ICML 2015 Workshop on Machine Learning for Education
Example 1: Teaching a support vector machine

- Here is an SVM:
  \[
  \min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{n} \max(1 - y_i x_i^\top \theta, 0) + \frac{1}{2} \| \theta \|^2
  \]
Example 1: Teaching a support vector machine

- Here is an SVM: \( \min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{n} \max(1 - y_i x_i^\top \theta, 0) + \frac{1}{2} \| \theta \|^2 \)

- What (batch) training set can teach the model: \( \theta^* = (\frac{1}{2}, \frac{\sqrt{3}}{2})^\top \)?
Example 1: Teaching a support vector machine

- Here is an SVM: \( \min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{n} \max(1 - y_i x_i^\top \theta, 0) + \frac{1}{2} \|\theta\|^2 \)
- What (batch) training set can teach the model: \( \theta^* = \left( \frac{1}{2}, \frac{\sqrt{3}}{2} \right)^\top \)?
- Teach the exact \( \theta^* \), not just the decision boundary \( x^\top \theta^* = 0 \)
One training item is necessary and sufficient!

\[ x_1 = \left( \frac{1}{2}, \frac{\sqrt{3}}{2} \right)^\top, \quad y_1 = 1 \]

You don’t even need negative training items.
What’s the catch?

[Liu & Zhu unpublished]

**Theorem (Teaching Dimension of homogeneous SVM)**

To teach any target model $\theta^* \neq 0$ to a homogeneous SVM

$$\min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{n} \max(1 - y_i x_i^\top \theta, 0) + \frac{\lambda}{2} \|\theta\|^2$$

one needs $n = \lceil \lambda \|\theta^*\|^2 \rceil$ training items.
What’s the catch?

[Liu & Zhu unpublished]

**Theorem (Teaching Dimension of homogeneous SVM)**

To teach any target model $\theta^* \neq 0$ to a homogeneous SVM

$$\min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{n} \max(1 - y_i x_i^\top \theta, 0) + \frac{\lambda}{2} \|\theta\|^2$$

one needs $n = \lceil \lambda \|\theta^*\|^2 \rceil$ training items.

**Proposition**

One such training set is $n = \lceil \lambda \|\theta^*\|^2 \rceil$ identical copies of the following item:

$$x = \frac{\lambda \theta^*}{n}, \quad y = 1.$$
Example 2: Teaching a Gaussian density estimator

- Given \( x_1 \ldots x_n \in \mathbb{R}^d \), the student computes sample mean and sample covariance:

\[
\hat{\mu} = \frac{1}{n} \sum x_i, \quad \hat{\Sigma} = \frac{1}{n-1} \sum (x_i - \hat{\mu})(x_i - \hat{\mu})^\top
\]
Example 2: Teaching a Gaussian density estimator

- Given \( x_1 \ldots x_n \in \mathbb{R}^d \), the student computes sample mean and sample covariance:
  \[
  \hat{\mu} = \frac{1}{n} \sum x_i, \quad \hat{\Sigma} = \frac{1}{n-1} \sum (x_i - \hat{\mu})(x_i - \hat{\mu})^\top
  \]

- The teacher wants to teach the model \( \mathcal{N}(\mu^*, \Sigma^*) \)
$d + 1$ training items necessary and sufficient

Vertices of $d$-dim tetrahedron
Two sides of a coin

- Machine learning: given data $D$, find model $\theta$
Two sides of a coin

- Machine learning: given data $D$, find model $\theta$
- Machine teaching: given model $\theta$, find (the smallest) data $D$
Two sides of a coin

- **Machine learning**: given data $D$, find model $\theta$
- **Machine teaching**: given model $\theta$, find (the smallest) data $D$
  - for any given learner

> $D$ will usually not be i.i.d.

Optimal teaching studied as optimal teaching (Goldman & Kearns 1995, many others)

Traditional emphasis: version-space learners

Our emphasis: modern optimization-based learners
Two sides of a coin

- Machine learning: given data $D$, find model $\theta$
- **Machine teaching**: given model $\theta$, find (the smallest) data $D$
  - for any given learner
  - $D$ will usually not be i.i.d.
Two sides of a coin

- Machine learning: given data $D$, find model $\theta$
- Machine teaching: given model $\theta$, find (the smallest) data $D$
  - for any given learner
  - $D$ will usually not be i.i.d.
  - studied as optimal teaching [Goldman & Kearns 1995, many others]
Two sides of a coin

- Machine learning: given data $D$, find model $\theta$
- **Machine teaching**: given model $\theta$, find (the smallest) data $D$
  - for any given learner
  - $D$ will usually not be $i.i.d.$
  - studied as optimal teaching [Goldman & Kearns 1995, many others]
  - traditional emphasis: version-space learners
Two sides of a coin

- **Machine learning**: given data $D$, find model $\theta$
- **Machine teaching**: given model $\theta$, find (the smallest) data $D$
  - for any given learner
  - $D$ will usually not be i.i.d.
  - studied as optimal teaching [Goldman & Kearns 1995, many others]
  - traditional emphasis: version-space learners
  - our emphasis: modern optimization-based learners
(Optimization-based) machine learning

Given $D = \{z_1 \ldots z_n\}$, we consider any learner with regularized empirical risk minimization:

$$
\hat{\theta} \leftarrow \arg\min_{\theta \in \Theta} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta)
$$

- $z_i = (x_i, y_i)$ for supervised learning, $z_i = x_i$ for unsupervised learning
(Optimization-based) machine learning

Given $D = \{z_1 \ldots z_n\}$, we consider any learner with regularized empirical risk minimization:

$$\hat{\theta} \leftarrow \arg\min_{\theta \in \Theta} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta)$$

- $z_i = (x_i, y_i)$ for supervised learning, $z_i = x_i$ for unsupervised learning
- $|D| = n$
(Optimization-based) machine learning

Given $D = \{z_1 \ldots z_n\}$, we consider any learner with regularized empirical risk minimization:

$$
\hat{\theta} \leftarrow \arg \min_{\theta \in \Theta} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta)
$$

- $z_i = (x_i, y_i)$ for supervised learning, $z_i = x_i$ for unsupervised learning
- $|D| = n$
- $\ell(\cdot, \cdot)$ a loss function
(Optimization-based) machine learning

Given $D = \{z_1 \ldots z_n\}$, we consider any learner with regularized empirical risk minimization:

$$\hat{\theta} \leftarrow \arg\min_{\theta \in \Theta} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta)$$

- $z_i = (x_i, y_i)$ for supervised learning, $z_i = x_i$ for unsupervised learning
- $|D| = n$
- $\ell(\cdot, \cdot)$ a loss function
- $\Omega(\cdot)$ a regularizer
(Optimization-based) machine learning

Given $D = \{z_1 \ldots z_n\}$, we consider any learner with regularized empirical risk minimization:

$$\hat{\theta} \leftarrow \operatorname{argmin}_{\theta \in \Theta} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta)$$

- $z_i = (x_i, y_i)$ for supervised learning, $z_i = x_i$ for unsupervised learning
- $|D| = n$
- $\ell(\cdot, \cdot)$ a loss function
- $\Omega(\cdot)$ a regularizer
- \operatorname{argmin} is the learning algorithm $A$
The math of machine teaching

\[ \mathbb{A}(\theta)^{-1} \mathbb{A}(D) \theta^* \]
Machine teaching

Given $\theta^*$ and $A$, find the smallest training set:

$$D^* \leftarrow \arg \min_{D \in \mathcal{D}} \quad |D| \quad \text{Teacher’s problem}$$

subject to

$$\theta^* = \arg \min_{\theta \in \Theta} \frac{1}{|D|} \sum_{z_i \in D} \ell(z_i, \theta) + \Omega(\theta) \quad \text{learner’s algorithm } A$$

Bilevel optimization
Solution idea

Convert lower level problem to nonlinear constraints:

\[
D^* \leftarrow \arg\min_{D \in \mathbb{D}} |D|
\]

s.t. Karush-Kuhn-Tucker conditions of \( A \) at \( \theta^* \)
Machine teaching is stronger than active learning

Sample complexity to achieve $\epsilon$ error
- passive learning $1/\epsilon$
Machine teaching is stronger than active learning

- **Passive learning** "waits" with $O(1/n)$ sample complexity.
- **Active learning** "explores" with $O(1/2^n)$ sample complexity.
- **Teaching** "guides" with $O(1/2^n)$ sample complexity.

Sample complexity to achieve $\epsilon$ error:
- Passive learning: $1/\epsilon$
- Active learning: $\log(1/\epsilon)$
Machine teaching is stronger than active learning

Sample complexity to achieve $\epsilon$ error

- passive learning $1/\epsilon$
- active learning $\log(1/\epsilon)$
- machine teaching 2: the teacher knows $\theta$
Education potentials

Education

- **Assumption 1**: The educational goal can be reasonably approximated by an objective function on $\theta^*$
Assumption 1: The educational goal can be reasonably approximated by an objective function on $\theta^*$
- e.g. $\|\hat{\theta} - \theta^*\|^2$
Assumption 1: The educational goal can be reasonably approximated by an objective function on $\theta^*$
- e.g. $\|\hat{\theta} - \theta^*\|^2$
- e.g. test set accuracy
Assumption 1: The educational goal can be reasonably approximated by an objective function on $\theta^*$
- e.g. $\|\hat{\theta} - \theta^*\|^2$
- e.g. test set accuracy

Assumption 2: The student can be reasonably approximated by a machine learning algorithm $A: \mathbb{D} \mapsto \Theta$
Education

- **Assumption 1:** The educational goal can be reasonably approximated by an objective function on $\theta^*$
  - e.g. $\|\hat{\theta} - \theta^*\|^2$
  - e.g. test set accuracy

- **Assumption 2:** The student can be reasonably approximated by a machine learning algorithm $A : \mathbb{D} \mapsto \Theta$
  - e.g. regression, SVM, neural network
Machine teaching for education

1. \( D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \)
Machine teaching for education

1. $D^* \leftarrow \text{MachineTeaching}(\theta^*, A)$
2. Train human on $D^*$
Machine teaching for education

1. \( D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \)
2. Train human on \( D^* \)
3. Test human on a test set
Machine teaching for education

1. \( D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \)
2. Train human on \( D^* \)
3. Test human on a test set
Machine teaching for education

1. \(D^* \leftarrow \text{MachineTeaching}(\theta^*, A)\)
2. Train human on \(D^*\)
3. Test human on a test set

\(D^*\) should be better than any other lesson \(D\) by definition!
Machine teaching for education

1. $D^* \leftarrow \text{MachineTeaching}(\theta^*, A)$
2. Train human on $D^*$
3. Test human on a test set

$D^*$ should be better than any other lesson $D$ by definition!
What if it isn’t?
Machine teaching for education

1. \[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]
2. Train human on \( D^* \)
3. Test human on a test set

\( D^* \) should be better than any other lesson \( D \) by definition!

What if it isn’t?

... blame yourself (choice of \( A \))
Machine teaching for education

1. \( D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \)
2. Train human on \( D^* \)
3. Test human on a test set

\( D^* \) should be better than any other lesson \( D \) by definition!

What if it isn’t?

... blame yourself (choice of \( A \))
... blame us (bilevel optimization)
Evidence from cognitive psychology

Human categorization [Patil Z Kopeć Love 2014]
Evidence from cognitive psychology

- Human categorization [Patil Z Kopeć Love 2014]
- $A$ is a limited capacity retrieval cognitive model $\approx$ kernel density classifier
Evidence from cognitive psychology

- Human categorization [Patil Z Kopeć Love 2014]
- $A$ is a limited capacity retrieval cognitive model $\approx$ kernel density classifier
Evidence from cognitive psychology

- **Human categorization** [Patil Z Kopec Love 2014]
- A is a limited capacity retrieval cognitive model \( \approx \) kernel density classifier

<table>
<thead>
<tr>
<th>human trained on</th>
<th>human test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>optimal lesson ( D^* )</td>
<td>72.5%</td>
</tr>
<tr>
<td>iid</td>
<td>69.8%</td>
</tr>
</tbody>
</table>

(statistically significant)
Open research questions in machine teaching

- approximate teaching: \[ \| \hat{\theta} - \theta^* \| \leq \epsilon \]
Open research questions in machine teaching

- approximate teaching: $\|\hat{\theta} - \theta^*\| \leq \epsilon$
- teaching under budget: $n \leq B$
Open research questions in machine teaching

- approximate teaching: \[ \|\hat{\theta} - \theta^*\| \leq \epsilon \]
- teaching under budget: \[ n \leq B \]
- uncertainty in learner \[ A \]: unknown \[ \lambda \]
Open research questions in machine teaching

- approximate teaching: \( \| \hat{\theta} - \theta^* \| \leq \epsilon \)
- teaching under budget: \( n \leq B \)
- uncertainty in learner \( A \): unknown \( \lambda \)
- sequential learner
Open research questions in machine teaching

- approximate teaching: $||\hat{\theta} - \theta^*|| \leq \epsilon$
- teaching under budget: $n \leq B$
- uncertainty in learner $A$: unknown $\lambda$
- sequential learner
- reinforcement learner
Open research questions in machine teaching

- approximate teaching: \[ \| \hat{\theta} - \theta^* \| \leq \epsilon \]
- teaching under budget: \( n \leq B \)
- uncertainty in learner \( A \): unknown \( \lambda \)
- sequential learner
- reinforcement learner
- real world applications
Open research questions in machine teaching

- approximate teaching: $\|\hat{\theta} - \theta^*\| \leq \epsilon$
- teaching under budget: $n \leq B$
- uncertainty in learner $A$: unknown $\lambda$
- sequential learner
- reinforcement learner
- real world applications
- ...
Summary

\[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]

- not machine learning
Summary

\[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]

- not machine learning
- reverse engineering
Summary

\[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]

- not machine learning
- reverse engineering
- potential new paradigm for education
Summary

\[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]

- not machine learning
- reverse engineering
- potential new paradigm for education
Summary

\[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]
- not machine learning
- reverse engineering
- potential new paradigm for education

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

\[ D^* \leftarrow \text{MachineTeaching}(\theta^*, A) \]

- not machine learning
- reverse engineering
- potential new paradigm for education

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

Collaborators: Scott Alfeld, Martha Alibali, Michael Ferris, Ji Liu, Bradley Love, Percival Matthews, Shike Mei, Bilge Mutlu, Gorune Ohannessian, Martina Rau, Tim Rogers, Ayon Sen, Steve Wright.