Machine Teaching

- Kevin, Justin, Zilong, Kaikai
Contents

• Introduction
• Problem Formulation
• Application on Computer Security
• Application on Education
• Open Questions
<table>
<thead>
<tr>
<th></th>
<th>Active Teacher</th>
<th>Passive Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Active Learner</strong></td>
<td>Interactive Machine Teaching</td>
<td>Active Learning</td>
</tr>
<tr>
<td><strong>Passive Learner</strong></td>
<td><strong>Machine Teaching</strong></td>
<td>Passive Learning</td>
</tr>
</tbody>
</table>
Consider a “student” who is a machine learning algorithm, for example, a SVM.

Consider a “teacher” who wants the student to learn a target model $\theta^*$, for example, a specific hyper-plane in SVM.

The teacher knows $\theta^*$ and the student’s learning algorithm $A$, and teaches by giving the student training examples.

Machine teaching aims to design the optimal training set $D$. 
Steve (the student) runs a linear SVM
Given a training set with \( n \) items \( x_i \in \mathbb{R}^d \), \( y_i \in \{-1, 1\} \), Steve learns \( w \in \mathbb{R}^d \)

Tina (the teacher) wants Steve to learn a target \( w^* \)

What is the smallest training set Tina can give Steve?
Example One

- Tina’s $non - iid$ training set with $n = 2$ items
  One positive sample + One negative sample
Example Two

- Steve wants to estimate a Gaussian density
- Given \( x_1 \cdots x_n \in \mathbb{R}^d \)
- Steve learns

\[
\hat{\mu} = \frac{1}{n} \sum x_i \\
\hat{\Sigma} = \frac{1}{n-1} \sum (x_i - \hat{\mu})(x_i - \hat{\mu})^T
\]

- Tina wants Steve to learn a target Gaussian with \((\mu^*, \Sigma^*)\)
Example Two

- Tina’s minimal training set of \( n = d + 1 \) tetrahedron vertices
Introduction

Comparison between 3 algorithms

• Teach a 1D threshold classifier
  • For example, given: $A = \text{SVM}$, $\theta^* = 0.6$

• What is the smallest training set $D$?
  • Passive Learning?
  • Active Learning?
  • Machine Teaching?
Sample complexity to achieve $\epsilon$ error:

- **Passive learning** $n = O\left(\frac{1}{\epsilon}\right)$
- **Active learning** $n = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$: needs binary search
- **Machine teaching** $n = O(1)$: teaching dimension [Goldman + Kearns 1995], the teacher knows $\theta^*$, only need two samples

$$n_1 = \left(\theta^* - \frac{\epsilon}{2}, -1\right), \quad n_2 = \left(\theta^* + \frac{\epsilon}{2}, +1\right)$$
Machine Teaching, an inverse problem of Machine Learning

Problem Formulation
Machine Teaching, an inverse problem of Machine Learning

- Teacher wants Student to learn a target model $\theta^*$
  - not machine learning: Teacher already knows $\theta^*$
  - Teacher knows Student’s learning algorithm $A$
- Teacher seeks **the best training set** within $A^{-1}(\theta^*)$ for Student
How to define “Best”?

- Formulate it as an optimization problem
  \[
  \min_{D \in \mathbb{D}} \epsilon(D) \\
  s.t. \quad A(D) = \theta^*
  \]

- \( \epsilon(D) \): “Teaching effort function” which we must define to capture the notion of training set optimality
- \( \mathbb{D} \): Search space of training sets
- \( D \): Selected training set
- \( A \): Learning algorithm
- \( \theta^* \): Target model
Problem Formulation

How to define “Best”?

• $\min_{D \in \mathbb{D}} \epsilon(D)$
  
  s.t. $A(D) = \theta^*$

• **Question 1**: How to define the teaching effort function $\epsilon(D)$?

• **Question 2**: Can we get a closed-form solution for $A(D) = \theta^*$?
How to define the teaching effort function $\epsilon(D)$?

- Normally, we prefer small training sets, so we can define 
  $$\epsilon(D) = |D|$$

- If we require the optimal training set to contain exactly $n$ items (imagine the limited capacity of human brains), we may define 
  $$\epsilon(D) = \mathbb{I}_{|D|=n}$$

- If we teach a classification task, we may prefer that any two training items from different classes be clearly distinguishable. Here, $D$ is of the form $D = (x_1, y_1), \ldots, (x_n, y_n)$, we may define 
  $$\epsilon(D) = \sum_{i,j: y_i \neq y_j} \|x_i - x_j\|^{-1}$$
Problem Formulation

Can we get a closed-form solution for $A(D) = \theta^*$?

• For some learners, we can.
  • One example is ordinary least squares regression where
    \[ A(D) = (X^T X)^{-1} X^T y, \text{ with } D = (X, y) \]

• For most learners, there is no closed-form $A(D)$ and we can only approach the problem with an optimization-based method
Problem Formulation

General Machine Teaching Framework

\[
\min_{D \in \mathcal{D}, \hat{\theta} \in \Theta} R_T(\hat{\theta}) + \lambda E_T(D)
\]

s.t. \[
\hat{\theta} = A(D)
\]

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

Machine Teaching on Education and Computer Security

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

- What if we can contaminate the training set $D$?
  - $A(D + \delta) = \{\theta'\}$, Adversarial Machine Teaching --> Wrong model
  - Application on Computer Security

- What if the learning algorithm is unknown?
  - Human teaching, with limited brain capacity
  - Application on Education
Adversarial Machine Teaching
“Microsoft’s website notes that Tay has been built using ‘relevant public data’ that has been ‘modeled, cleaned, and filtered,’ but it seems that after the chatbot went live filtering went out the window.”
Application - Teaching the Wrong Model

Why do we want to teach the wrong model?

How can we poison the training data?

What is the goal?
Contaminating Training Data

Hacking

Optimally Poison → Training Data → Mislead → Machine Learner

Specific
Wrong Model

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

Yellow line - correct model
Green line - target model

Computer Virus
Contaminating Training Data

Poisoned Data Points
Contaminating Training Data

Training Set Attack Algorithm

\[
\min_{D \in \mathcal{D}, \hat{\theta}_D} \quad O_A(D, \hat{\theta}_D) \\
\text{s.t.} \quad \hat{\theta}_D \in \arg \min_{\theta \in \Theta} O_L(D, \theta) \\
\text{s.t.} \quad g(\theta) \leq 0, \ h(\theta) = 0
\]

Overall attacker objective function

Learner’s objective

Bilevel optimization problem
Bilevel optimization problems are NP-hard in general.

Assume attack space is differentiable.

Can reduce problem to single-level constrained optimization problem by replacing lower-level problem with its Karush-Kuhn-Tucker (KKT) conditions (the constraints are stationarity, complementary slackness, primal and dual feasibility)

\[
\begin{align*}
\min_{D \in D, \theta, \lambda, \mu} & \quad O_A(D, \theta) \\
\text{s.t.} & \quad \partial_\theta (O_L(D, \theta) + \lambda^T g(\theta) + \mu^T h(\theta)) = 0 \\
& \quad \lambda_i g_i(\theta) = 0, i = 1 \ldots m \\
& \quad g(\theta) \leq 0, h(\theta) = 0, \lambda \geq 0
\end{align*}
\]
Contaminating Training Data

Experiments - SVM

SVM rating wine as good or bad
Goal is to teach model that only the feature “alcohol” correlates with wine quality
Improvement from $E_A = 515$ to $E_A = 370$

Figure 1: Training-set attack on SVM. The “alcohol” feature is marked by a red star in (b,c).
Contaminating Training Data

Experiments - Logistic Regression

Logistic Regression calculating spam likelihood
Goal is to teach model that the word “credit” does not affect spam likelihood
Improvement from $E_A = 390$ to $E_A = 232$

Figure 2: Training-set attack on logistic regression. The 20th feature on “frequency of word credit” is marked
Contaminating Training Data

Experiments - Linear Regression

Linear Regression learning a warming trend based on number of frozen days for Lake Mendota
Goal is to hide the warming trend
Different norms for attacker effort

Figure 3: Training-set attack on OLS
Machine Teaching on Education
Machine Teaching on Education

Education System Overview

1. Teaching Image
2. Student's Estimate
3. Class Label
Two fundamental questions

- **Teaching strategy**
  - How to teach one to achieve the expectation given a budget
  - Evaluate students’ performance

- **Human learning model**
  - How to know human’s learning algorithm and feature representation :
  - Limited and imperfect memory for recognition :
  - Generalization power: generalize to unknown examples and perform domain adaptation given only few instances :(
Machine Teaching in image classification training

- Motivation: image labeling which needs expertise like Chinese characters

- The goal is to choose teaching images that will maximize the student’s classification ability in the minimum amount of teaching time
Teaching Strategy

- Random sampling: randomly choose the examples to teach
  - redundantly present teaching examples of concepts that have already been learned
  - not reinforce concepts that the student is uncertain about
Teaching Strategy

- “worst predicted”: optimally seeks to show the student the image that they are currently most uncertain about
Teaching Strategy

- “worst predicted”: optimally seeks to show the student the image that they are currently most uncertain about
Teaching Strategy

- “worst predicted”: optimally seeks to show the student the image that they are currently most uncertain about
Machine Teaching on Education

Teaching Strategy

- Expected error reduction teaching
  - it concentrates on regions of high density in the feature space
Teaching Strategy

- Expected error reduction teaching
  - it concentrates on regions of high density in the feature space
Teaching Strategy

- Expected error reduction teaching
  - it concentrates on regions of high density in the feature space
Teaching Strategy

- Expected error reduction teaching
  - it concentrates on regions of high density in the feature space
Teaching Strategy

- Expected error reduction teaching
  - it concentrates on regions of high density in the feature space
Machine Teaching on Education

Teaching Strategy

- Expected error reduction teaching
  - it concentrates on regions of high density in the feature space
Machine Teaching on Education

Performance comparison

Chinese

Butterflies

Seabed
Machine Teaching on Education

Teaching process

Teaching images

Random

Centroids

Worst Pred

Batch

EER

Correct Incorrect

Grass Mound Stem
Human Teaching
How do Humans teach?

Research by Faisal Khan, Xiaojin Zhu, Bilge Mutlu

Basic research question: can we use Machine Teaching to model and analyze how humans teach?

Secondary Question: Does how human teach show us anything about how humans learn?
Human Teaching

Classic Teaching Dimension Model

Classification of feature over a singular axis

Optimal teaching strategy is the boundary strategy, where you present the two examples closest the the boundary

Alternative model is the extreme strategy

Alternate easy to hard

\[(x_1, 0), (x_n, 1), (x_2, 0), (x_{n-1}, 1), \ldots, (x_j, 0), (x_{j+1}, 1)\]
31 Volunteers were given the task of teaching whether an object in a picture is graspable or not.

Target was a robot that simply followed motion in the room and did not learn anything.

Each participant had their own labeling of graspable or not graspable.
Three Major Human Teaching Strategies

1. **Extreme strategy** - starts with objects at extremes and moves towards decision boundary (14/31)
2. **Linear Strategy** - moves from one side to the other (14/31)
3. **Positive-only strategy** - only gave examples of objects that were graspable. (3/31)
4. None used the boundary strategy, and people typically started at the extremes.

![Extreme](Extreme.png) ![Linear](Linear.png) ![Positive-only](Positive-only.png)
Theoretical Account of the “Extreme” Teaching

Data shows that extreme teaching is a popular strategy, but boundary teaching is never used.

New proposed model: humans represent everyday objects in a highly dimensional feature space $X = [0,1]^d$. Assume binary label $y = 1 | x_1 > 1/2$.

Assumption: Learners selects a hypothesis $h$ and follows it until it no longer works, then picks working hypothesis

The boundary strategy isn’t good for this model!
Consider optimizing learning with two examples, one positive and one negative. We choose $x_1 = (a, x_{12}, x_{13}, ...)$ and $x_2 = (b, x_{22}, x_{23}, ...)$ as two examples.

Risk is achieved when $a = 1$, $b = 0$ when $d \to \infty$.
Human Teaching

Teaching Sequence should Gradually Approach Boundary

Assumption: Teacher only cares about the 1st dimension.

Collorary: All other dimensions can be treated as random variables

Suppose $V_k(t)$ are the hypotheses in the $k$th dimension that are viable. $V$ is non-empty when the points revealed are separable in dimension $k$

Choosing extreme points makes sure that the other $V_k$ for $k \neq 1$ are weeded out before as it is bound to become inseparable quickly as they are chosen randomly

Choosing extreme points ensure that the majority of hypotheses left are good
Human Teaching

Potential Takeaways on Human Learning

Humans utilize a multidimensional representation of objects.

Humans use the extreme strategy to minimize the per-iteration expected error, rather than worst-case error.

A theoretical simulation of extreme teaching shows that it approaches optimal in minimizing per-iteration expected error.

This may be due to the teacher being limited to objects in the pool of objects, whereas the goal is for generalization of other objects.
Criticisms and proposed extensions to the study

1. Students are assumed to be unable to provide live feedback to the teacher while they can in real life.
2. A centroid based learning model would likely explain the extreme strategy better.
3. The study used everyday people to show how they teach. It would be interesting to see how educators or people trained in education would teach differently.
4. The paper only explains half of the strategies used. What is the justification for linear or positive-only teaching?
Extensions To Machine Teaching for Humans

Increase human learning rate through rapid teaching strategies

Curriculum design for multi-concept machine teaching

Modeling memory loss and long term memory tradeoff

Modeling relational concepts such as in Physics

Interactive Machine Learning

Improve human accuracy through better teaching techniques
Open Questions

- Optimization
  - Solving for optimal training data set D
- Theory
  - Theoretical study of teaching dimension (maybe information theory)
- Psychology
  - Adjudicate existing cognitive models for human categorization
- Education
- Novel Applications
Reference

1. The Teaching Dimension of Linear Learners, Liu, Ji; Zhu, Xiaojin

2. How Do Humans Teach: On Curriculum Learning and Teaching Dimension, Faisal Khan, Xiaojin Zhu, and Bilge Mutlu


4. Near-Optimally Teaching the Crowd to Classify, Adish Singla, Ilija Bogunovic, Gabor Bartok, Amin Karbasi, and Andreas Krause

Thank you!

Q&A