An Overview of Machine Teaching

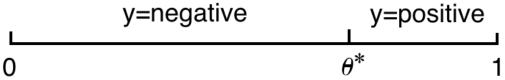
Adish Singla, Jerry Zhu

NIPS 2017 Workshop on
Teaching Machines, Robots, and Humans

A prototypical machine teaching task

Compare passive learning, active learning, teaching

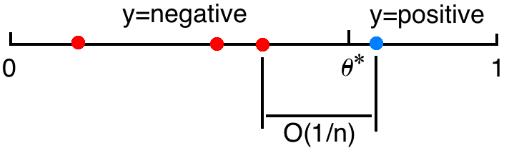
Example: learn a 1D threshold classifier



Passive learning

$$x_1, \ldots, x_n \sim U[0, 1]$$

$$y_i = \theta^*(x_i)$$



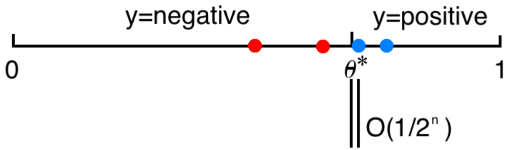
With large probability

$$|\hat{\theta} - \theta^*| = O(n^{-1}) \leq \epsilon$$

$$n \geq O(\epsilon^{-1})$$

Active learning

- learner picks query x, oracle answers $y = \theta^*(x)$
- binary search

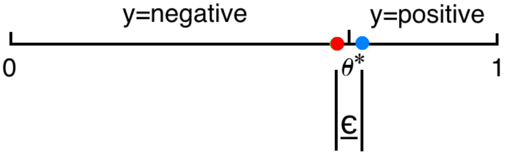


$$|\hat{\theta} - \theta^*| = O(2^{-n}) \leq \epsilon$$

$$n \geq O(\log(\epsilon^{-1}))$$

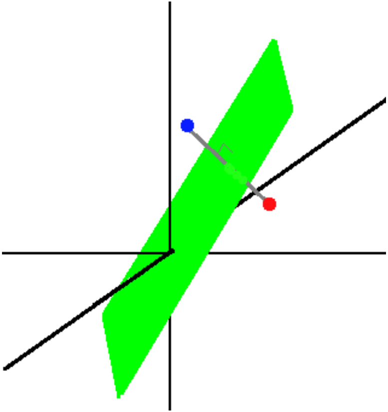
Machine teaching

- teacher can design an optimal training set of size 2



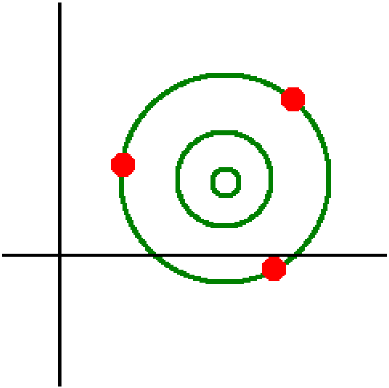
$$n = 2, \ \forall \epsilon > 0$$

Another example: teach hard margin SVM



Remark: Teaching Dimension $TD = 2$ but $VC = d + 1$

Yet another example: teach Gaussian density



$TD = d + 1$: tetrahedron vertices

Machine learning vs. machine teaching

Machine learning (D given, learn θ)

$$\min_{\theta} \sum_{(x,y)\in D} \ell(x,y,\theta) + \lambda\|\theta\|^2$$

Machine teaching (θ^* given, learn D)

$$\begin{aligned} \min_{D,\hat{\theta}} \quad & \|\hat{\theta} - \theta^*\|^2 + \eta\|D\|_0 \\ \text{s.t.} \quad & \hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{(x,y)\in D} \ell(x,y,\theta) + \lambda\|\theta\|^2 \end{aligned}$$

D usually not $i.i.d.$

Why bother if we already know θ^*?

- education
- adversarial attacks
- ...

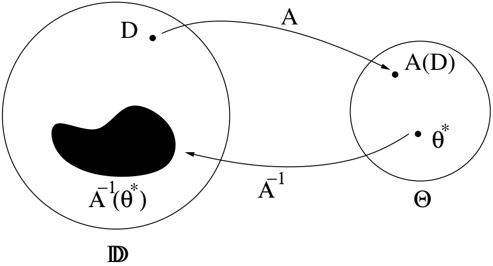
Machine teaching generic form

$$\begin{aligned} \min_{D,\hat{\theta}} \quad & \text{TeachingRisk}(\hat{\theta}) + \eta \text{TeachingCost}(D) \\ \text{s.t.} \quad & \hat{\theta} = \text{MachineLearning}(D) \end{aligned}$$

- exact vs. approximate teaching
- parameter vs. generalization error
- cost not always number of items
- any of the constraint forms

The coding view

message=θ^*, decoder=learning algorithm A, language=$\mathbb{D}$



In other words

teach·ing
/'teCHiNG/
noun

1. controlling
2. shaping
3. persuasion
4. influence maximization
5. attacking
6. poisoning

Characterizing the space of teaching tasks

The friend vs. foe dimension

friend

- education
- improving cognitive models
- fast classifier training
- debugging machine learners
- ...
- training-set poisoning attacks (not test-time adversarial attacks)

foe

The human vs. machine dimension

- machine teacher, machine learner: attacks
- machine teacher, human learner: education
- human teacher, machine learner: fast classifier training
- human teacher, human learner: (not this workshop)

The batch vs. sequential dimension

Batch teaching

- batch learners

Sequential teaching

- stochastic gradient descent learner
- multiarmed bandits
- reinforcement learning (e.g. teaching by demonstration)

The learner anticipation dimension

- does not anticipate teaching (standard learners)
  - version space learner $\{\theta \text{ agrees with } D\}$
  - Bayesian learner $p(\theta \mid D)$
  - convex regularized empirical risk minimizer

$$\min_{\theta} \sum_{(x,y) \in D} \ell(x, y, \theta) + \lambda \|\theta\|^2$$

  - deep neural network
  - cognitive model
- anticipates teaching
  - especially when human is involved

The learner anticipation dimension (cont.)

A smart learner can

- educate the (suboptimal human) teacher in the structure of the optimal teaching set
- detect suboptimal teaching and switch to active learning
- translate the teaching set aimed at a different learner

Recursive teaching dimension RTD (vs. classical TD)

The learner transparency dimension

clearbox

- bilevel optimization

$$
\begin{aligned}
\min_{D,\hat{\theta}} \quad & \text{TeachingRisk}(\hat{\theta}) + \eta \text{TeachingCost}(D) \\
\text{s.t.} \quad & \hat{\theta} = \text{MachineLearning}(D)
\end{aligned}
$$

- ...
- some learning hyper-parameters unknown. Probe and teach
- ...
- evaluation function on D. Bayesian optimization.

blackbox

The number of learners dimension

- one lecture, one student (standard)
- one lecture, many students
  - worst-case guarantee: minimax risk
  - average-case guarantee: Bayes risk

The "what's in your training set" dimension

Items in D can be:

- pool-based teaching: subset (multiset) of a given pool $\{(x_i, y_i)\}_{1:N}$
- synthetic / constructive teaching:
  - honest: $x \in \mathcal{X}, y = \theta^*(x)$
  - liar: $(x, y) \in \mathcal{X} \times \mathcal{Y}$; fake rewards. Ethics questions
- alter existing training set $D_0 + (\delta_x, \delta_y)$
- features, pairwise comparisons (requires corresponding learner)

The theory vs. empirical dimension

- teaching dimension TD, recursive teaching dimension RTD, preference-based TD, etc.

$$\begin{aligned} TD(\theta^*) \equiv \min_{D} \quad & \|D\|_0 \\ \text{s.t.} \quad & \{\theta^*\} = \text{VersionSpace}(D). \end{aligned}$$

- ...
- improve student test scores

A few open problems

Information complexity of teaching and learning

- relationship between recursive teaching dimension (RTD) and VC-dimension (VCD)
  - RTD is upper bounded by $O(VCD)$?
- connections between above question and the long-standing open "sample compression conjecture" [Littlestone and Warmuth, 1986]

New notions of teaching dimension (TD)

- complexity of teaching with new signals
  - e.g. features, pairwise comparison queries
- complexity of teaching with additional constraints
  - e.g. interpretable signals

Solving the combinatorial, bilevel optimization problem

- bilevel optimization

$$\begin{aligned} \min_{\color{red}{D},\hat{\theta}} \quad & \text{TeachingRisk}(\hat{\theta}) + \eta \text{TeachingCost}(D) \\ \text{s.t.} \quad & \hat{\theta} = \text{MachineLearning}(D) \end{aligned}$$

- mixed integer nonlinear programming
  - even simple instances are NP-Hard (e.g. set-cover, subset sum)
- requires new approximation algorithms
  - e.g. via characterizing submodularity properties of the problem

Machine teaching for reinforcement learning

- optimal demonstrations for inverse reinforcement learning
- teaching humans how to teach robots

The need for a good cognitive model

- model-based vs. model-free approaches
  - 3 workshop papers on spaced repetition technique
- can machine teaching guide the search for better models?

Novel applications and industry insights

- machine teaching for debugging machine learning?
- program synthesis, social robotics, etc.

Workshop preview

Cluster 1: Teacher who optimizes training data

Posters:

3 Optimizing Human Learning

5 Interpretable Machine Teaching via Feature Feedback

7 Interpretable and Pedagogical Examples

11 Accelerating Human Learning with Deep Reinforcement Learning

12 Program2Tutor: Combining Automatic Curriculum Generation with Multi-Armed Bandits for Intelligent Tutoring Systems

15 Predicting Recall Probability to Adaptively Prioritize Study

Talks:

- Emma Brunskill
- Burr Settles
- Le Song

Cluster 2: Student who appreciates teaching

Posters:

2 Machine Education - The Way Forward for Achieving Trust-Enabled Machine Agents

4 Model Distillation with Knowledge Transfer from Face Classification to Alignment and Verification

6 Pedagogical Learning

13 Generative Knowledge Distillation for General Purpose Function Compression

14 Explainable Artificial Intelligence via Bayesian Teaching

Cluster 3: Related Topics

Posters:

1 Generative Adversarial Active Learning

8 Faster Reinforcement Learning Using Active Simulators

9 Gradual Tuning: A Better Way of Fine Tuning the Parameters of a Deep Neural Network

10 Machine Teaching: A New Paradigm for Building Machine Learning Systems

Talks:

- Shay Moran
- Patrice Simard