Machine Teaching in Interactive Learning

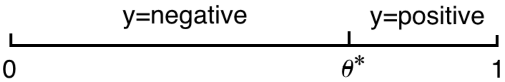
Jerry Zhu

University of Wisconsin-Madison

Simons Institute Workshop on Interactive Learning 2017

What do we want from interactivity?

Example: learn a 1D threshold classifier

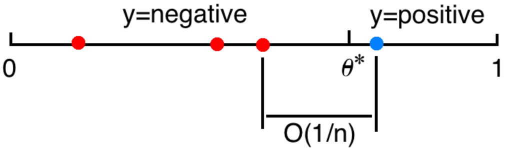


- item $x \in [0,1]$, label $y \in \{-1, 1\}$
- hypothesis space $\mathcal{H} = \{\theta \in [0,1] : \hat{y} = 1_{[x \geq \theta]}\}$
- target $\theta^* \in \mathcal{H}$

PAC (passive) learning

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

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



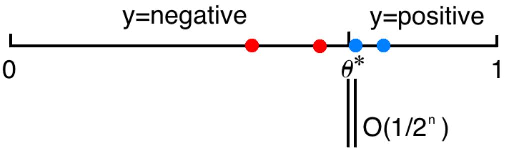
With large probability

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

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

Active learning

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

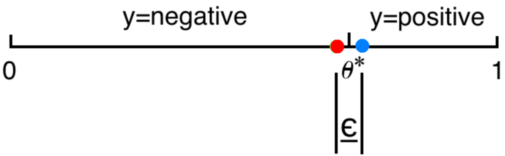


$$|\hat{\theta} - \theta^*| = O(2^{-n}) \leq \epsilon$$

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

An ideal human teacher

- passive learner: picks any $\hat{\theta}$ in version space
- teacher knows the learner
- designs an optimal training set!



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

Talk plan

1. Machine teaching: what can we expect from an ideal teacher?
2. The real world is not ideal

Part I

Humans are teachers, not annotators.

What can an ideal teacher do?

Machine teaching assumptions

- teacher knows $\theta^* \in \mathcal{H}$
- teacher can give a training set D, but not θ^*, to the learner

$$\text{constructive teaching (can lie) } D \in \mathbb{D} = \cup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n$$

$$\text{constructive teaching (honest) } D \in \mathbb{D} = (\cup_{n=1}^{\infty} (\mathcal{X})^n, Y = \theta^*(X))$$

$$\text{pool-based teaching } D \in \mathbb{D} = 2^{\{(x_i, y_i)\}_{1:N}}$$

- teacher knows the learning algorithm / estimator / student A

$$A : \mathbb{D} \mapsto 2^{\mathcal{H}}$$

  - e.g. version space learner

$$A(D) = \{\theta \in \mathcal{H} : \theta(x_i) = y_i, i = 1 \ldots n\}$$

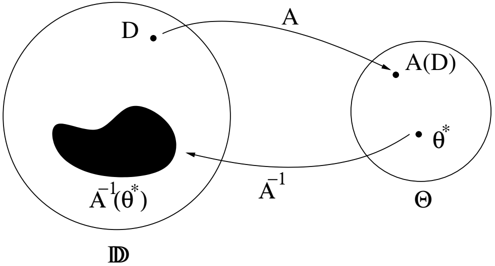
  - e.g. regularized empirical risk minimizer

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

(Special) machine teaching

$$\begin{aligned} \min_{D \in \mathbb{D}} \quad & \|D\|_0 \\ \text{s.t.} \quad & \{\theta^*\} = A(D) \end{aligned}$$

- Inverse machine learning view:



- Coding view: message$=\theta^*$, decoder$=A$, language$=\mathbb{D}$

(Special special) machine teaching

- i.e. classic optimal teaching (e.g. [Goldman+Kearns'95])
- further restrictions:
  - A is a version space learner
  - $\mathcal{X}, \mathbb{D}, \mathcal{H}$ often finite
- main concern: teaching dimension (TD)

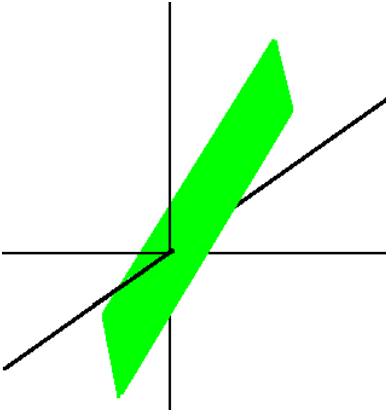
$$
\begin{aligned}
TD(\theta^*) \equiv \min_{D \in \mathbb{D}} \quad & \|D\|_0 \\
\text{s.t.} \quad & \{\theta^*\} = A(D).
\end{aligned}
$$

$$
TD(\mathcal{H}) \equiv \sup_{\theta^* \in \mathcal{H}} TD(\theta^*)
$$

- TD known for intervals, hypercubes, etc.
- TD $\neq$ VC-dim

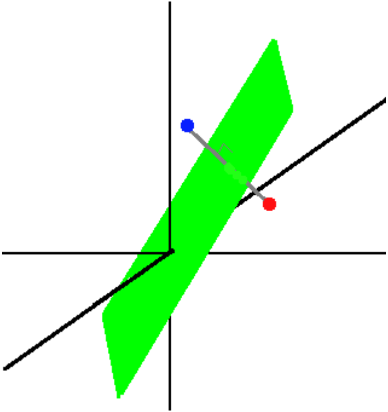
(Special) machine teaching: other A's

Example: teach target hyperplane $\mathbf{x}^\top \theta^* = 0$ in $\mathbb{R}^d$ to a hard margin SVM



(Special) machine teaching: other A's

Optimal (non-iid) training set with $\|D\|_0 = 2$ items

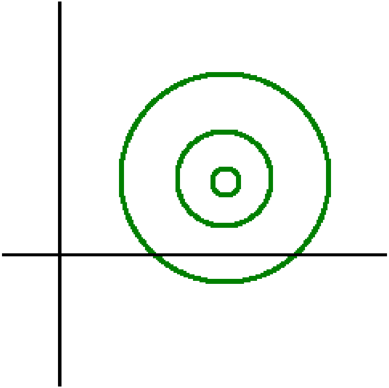


Note: TD=2 but VC=$d + 1$

(Special) machine teaching: other A's

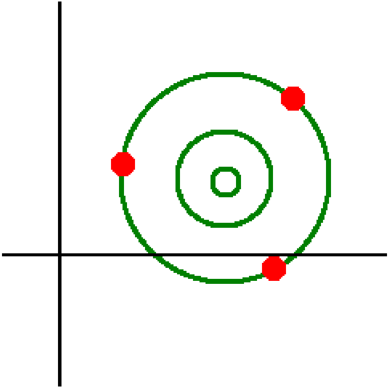
Example: teach a d-dim Gaussian $N(\mu^*, \Sigma^*)$ to the Maximum Likelihood Estimator

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



(Special) machine teaching: other A's

$TD = d + 1$: tetrahedron vertices



(Special) machine teaching: other A's

Example: teach linear learners (ridge regression, soft-margin SVM, logistic regression) [Liu+Z'16]

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

TD:

goal ↓	loss $\ell()$		
	squared	hinge	logistic
$A(D) = \theta^*$	1	$\lceil \lambda \|\theta^*\|^2 \rceil$	$\left\lceil \frac{\lambda \|\theta^*\|^2}{\tau_{\max}} \right\rceil$
boundary	-	1	1

Note: sometimes 1 training item suffices, even for classification.

Example: Ridge regression

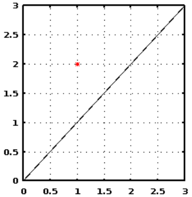
$$A(D) = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \frac{1}{2}(\theta x_i - y_i)^2 + \frac{\lambda}{2}\theta^2$$

Optimal teaching sets $n = 1$ ($\forall a \neq 0$):

$$x_1 = a\theta^*, \quad y_1 = \frac{\lambda + \|x_1\|^2}{a}$$

To teach a $\lambda = 1$ student the target $\theta^* = 1$, the teacher lies:

$$x_1 = 1, \quad y_1 = 2$$



TD as "Speed of Light"

[Goldman+Kearns'95, Angluin'04, Cakmak+Thomaz'11, Suh+Z+Amershi'16]

Unavoidable Effort in Interactive Machine Learning

$n \geq$ TD

- ideal teacher achieves $n =$TD
- can be much faster than active learning (recall 2 vs. $\log \frac{1}{\epsilon}$)
- must allow teacher-initiated items (unlike active learning)

(General) machine teaching

[Alfeld+Z+Barford'16,17, Mei+Z'15]

- learner risk $f(A(D), \theta^*)$, e.g. $\|A(D) - \theta^*\|^2$
- teacher effort $g(D)$, e.g. $\sum_{z \in D} \mathrm{cost}(z)$
- constrained forms:

$$\min_{D \in \mathbb{D}} g(D), \quad \text{s.t.} \quad f(A(D), \theta^*) \leq \text{Tolerance}$$

$$\min_{D \in \mathbb{D}} f(A(D), \theta^*), \quad \text{s.t.} \quad g(D) \leq \text{Budget}$$

- Lagrangian form:

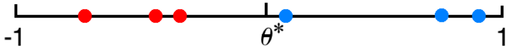
$$\min_{D \in \mathbb{D}} f(A(D), \theta^*) + \eta g(D)$$

- extends to sequential learners A

Example: Pool-based teaching

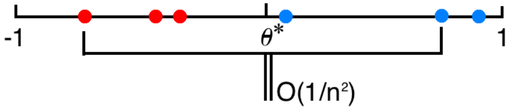
$$\min_{D \in \mathbb{D}} |A(D) - \theta^*|$$

- $x_1 \ldots x_n \sim U[-1, 1]$ fixed, $\mathbb{D} = 2^{\{(x_i, y_i)\}_{1:N}}$
- $\theta^* = 0$
- A=hard-margin SVM



Example: Pool-based teaching

Most symmetrical pair



With large probability,

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

Recall using the whole pool only gets $O(n^{-1})$
(Not training set reduction, nor sample compression)

Part I recap

Humans are teachers, not annotators.
What can an ideal teacher do?

- achieve TD, beat active learning
- passive learners just sit and wait for optimal training set

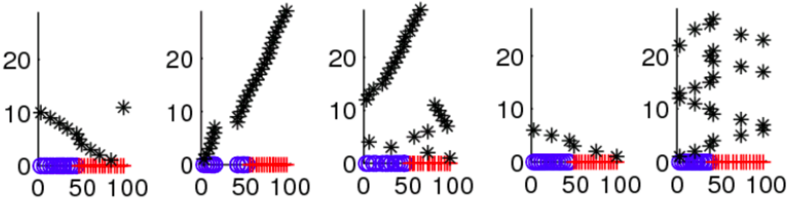
Part II

Most humans are not ideal teachers

How do real humans teach?

[Khan+Z+Mutlu'11]

1D pool-based task



Part II

Most humans are not ideal teachers

Ideas:

1. control them with mixed-initiative learning

The mixed-initiative algorithm

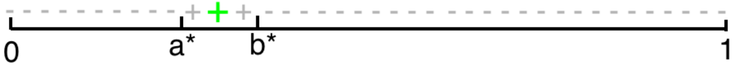
[Suh+Z+Amershi'16]

Data $D = \emptyset$
for $i = 1$ to TD **do**
    if human no longer wants to lead **then**
        break;
    else
        human chooses (x_i, y_i)
        append (x_i, y_i) to D
    end if
end for
run active learning starting from D until completion

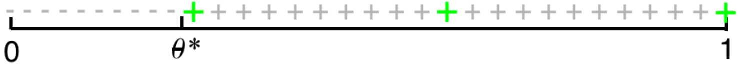
The guarantee

teacher →	ideal	seed	naive
active learning	AL	AL	AL
human-initiative	TD	∞	∞
mixed-initiative	TD	$TD + AL -$ blind search	$TD + AL$

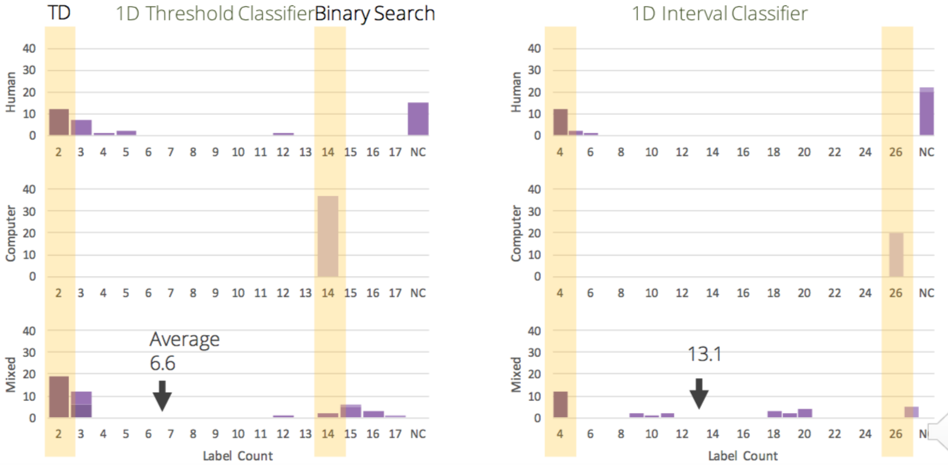
- Seed teacher: provides one point per positive region



- Naive teacher: can be arbitrarily bad



Human experiments: learn from 481 MTurkers



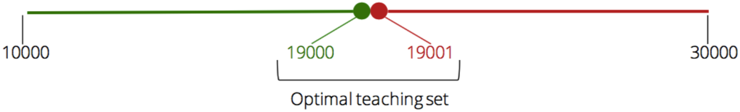
Part II

Most humans are not ideal teachers

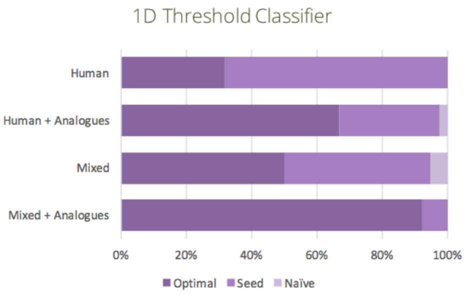
Ideas:

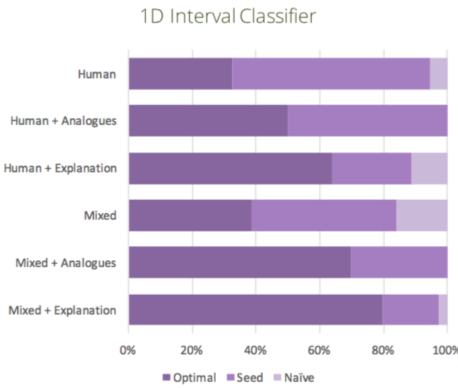
1. control them with mixed-initiative learning
2. educate them with analogues: automatically generated optimal training sets for arbitrary $\theta' \in \mathcal{H}$

   "If your price threshold was \$19000, you could show your robot these 2 examples: \$19000 is acceptable, \$19001 is unacceptable."



Human experiments





Part II

Most humans are not ideal teachers

Ideas:

1. control them with mixed-initiative learning
2. educate them with analogues
3. translate them: are they teaching for a different learner?

Translate the teacher

- human teaches $\theta^* = 1$ to ridge regression

$$A(D) = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{2}(\theta x_i - y_i)^2 + \frac{\lambda}{2}\theta^2$$

- human assumes wrong $\lambda^w = 1$, constructs teaching set $(x = \theta^*, y = \lambda^w + x^2) = (1, 2)$
- learner actually has $\lambda^* = 2$, will learn wrong $\theta = \frac{xy}{x^2+\lambda^*} = \frac{2}{3}$
- if a translator-in-the-middle knows λ^w, λ^*:

$$\tilde{x} = \frac{xy}{x^2 + \lambda^w}, \quad \tilde{y} = \lambda^* + \tilde{x}^2$$

- learner receives $(1, 3)$, learns correct $\theta = 1$

Summary

- Humans are teachers, not annotators
- Ideal teachers achieve TD, beat active learning
- Interactive learner can work with less ideal humans
  - control them
  - educate them
  - translate them

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