Machine Teaching as a Probe for Learning Mechanism in Humans

Jerry Zhu

University of Wisconsin-Madison

Tsinghua Laboratory of Brain and Intelligence
Workshop on Brain and Artificial Intelligence
Dec. 27, 2017

$\boxed{\text{input}} \rightarrow \boxed{\text{human}} \rightarrow \boxed{\text{performance}}$

Examples:

- human categorization
- human memorization

$\boxed{\text{input}} \rightarrow \boxed{\text{human}} \rightarrow \boxed{\text{performance}}$

$\boxed{\text{input}} \rightarrow \boxed{\text{mathematical model } A} \rightarrow \boxed{\text{performance 2}}$

The usual research flow:

1. run human experiments
2. tweak model A so that "performance 2" $\approx$ "performance"
3. publish

How to improve "already good" models?

1. no obvious improvements
2. multiple equally good models

Idea: feed atypical input to model A

The most interesting atypical input

Input D^* that, according to model A, maximizes performance:

$$\begin{array}{ll} \max\limits_{D} & \text{performance(A(D))} \\ \text{s.t.} & \text{constraints on } D \end{array}$$

We call D^* the optimal teaching input

A little logic

E1=A is a faithful model of human learning
E2=D^* maximizes performance on A
E3=D^* maximizes performance on humans

$$E1 \wedge E2 \Rightarrow E3$$

contraposition

$$\neg E3 \Rightarrow (\neg E1 \vee \neg E2)$$

If humans do not perform well on D^* ($\neg E3$), and since Jerry has confidence in how he optimizes D^* ($E2$), then the only logical conclusion is $\neg E1$.

The new research flow

1. find the optimal teaching input D^* that maximizes performance for model A
2. run human experiments with input D^*
3. if human performance improved
   - great! retain model A, publish

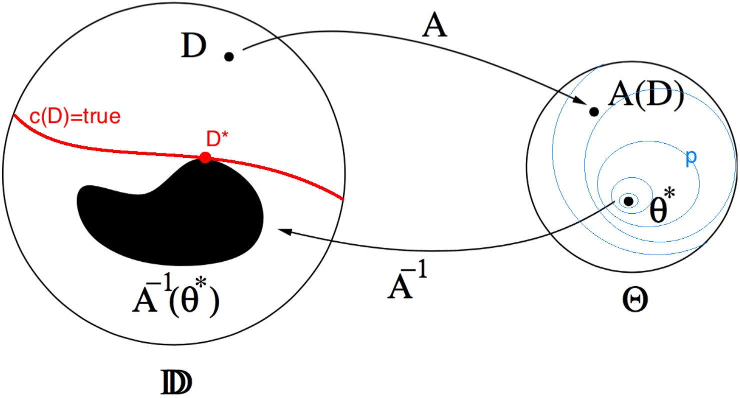
   else
   - D^* exposes problems, revise model A, publish

"Hedging"

Introducing machine teaching

Machine teaching: Finding the optimal teaching input D^*

- Given:
  - model $A : \mathbb{D} \mapsto \Theta$
  - performance measure $p(\theta), \theta \in \Theta$
  - constraints $c(D)$
- Optimize: input D^* that maximizes $p(A(D^*))$



Case study: humans categorization

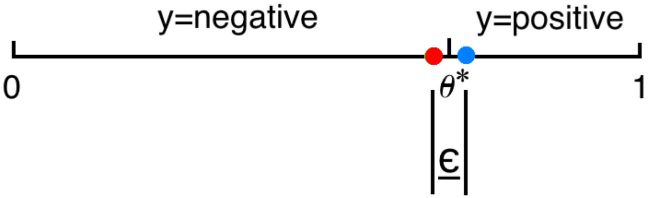
- training data $D = (x_1, y_1), \ldots, (x_n, y_n)$
- x_i: feature vector, y_i: class label
- cognitive model A: a (machine) learning algorithm $\mathbb{D} \mapsto \Theta$
- classifier $\theta : X \mapsto Y$
- performance measure $p(\theta)$: test set accuracy w.r.t. θ^* (This requires us to know the target model, or have a labeled test set)
- example constraints $c(D)$:
  - $x_i \in$ finite candidate pool (vs. R^d)
  - $|D| \leq n$

Machine teaching: Finding the optimal teaching input

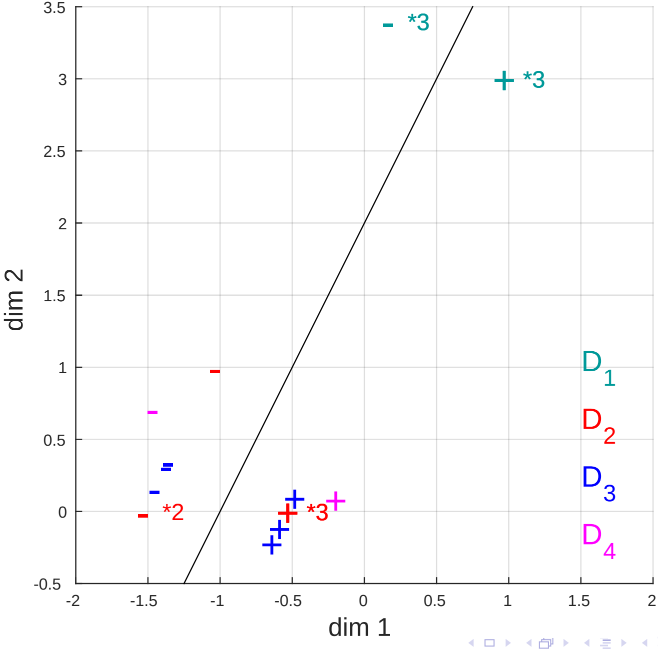
$$
\begin{aligned}
D^* := \operatorname*{argmax}_{D,\theta} \quad & p(\theta) \\
\text{s.t.} \quad & \theta = A(D) \\
& |D| \le n
\end{aligned}
$$

- first constraint = empirical risk minimization = optimization by itself
- bilevel combinatorial optimization
  - simple A (e.g. linear regression): closed-form D^*
  - convex A (e.g. logistic regression): KKT+implicit function $\rightarrow$ nonlinear optimization, or mixed-integer nonlinear program
  - complex A (e.g. neural networks): hill climbing etc.
- D^* usually not $i.i.d.$

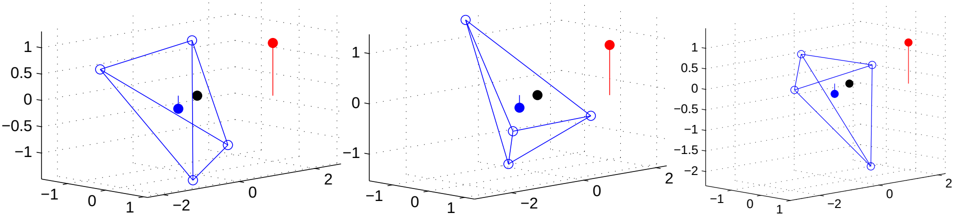
Machine teaching example 1



Machine teaching example 2



Machine teaching example 3



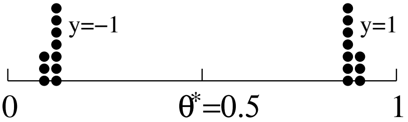
Recent machine teaching research

- NIPS 2017 Workshop on Teaching Machines, Robots, and Humans (my tutorial `http://pages.cs.wisc.edu/~jerryzhu/pub/NIPS17WStutorial.pdf`)
- Applications:
  - education
  - adversarial attacks
  - human robot interaction
  - interactive machine learning
  - algorithmic fairness
  - machine learning debugging

Human Categorization Example 1

[Patil et al. 2014]

- Human categorization task: line length
- 1D threshold $\theta^* = 0.5$
- A: kernel density estimator
- Optimal D^*:



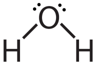
human trained on	human test accuracy
random items	69.8%
optimal D^*	72.5%

(statistically significant)

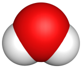
Human Categorization Example 2

[Sen et al. in preparation]

- Human categorization task: same or different molecules



Lewis representation



Space-filling representation

- A: neural network
- Optimal $D^*(n = 60)$:

human trained on	human pre-test error	post-test error
random input	31.7%	28.6%
expert input	28.7%	28.1%
D^*	30.6%	25.1%

(statistically significant)

Human Categorization Example 3

[Nosofsky & Sanders, Psychonomics 2017]

- Human categorization task: rock type

- Model A: Generalized Context Model (GCM)
- Optimal D^* does not work better on humans

human trained on	human accuracy
random input	67.2%
coverage input	71.2%
D^*	69.3%

- Experts are revising the model

Summary

1. Find D^* that maximizes performance for model A
2. Run human experiments with input D^*
   - either human performance improved
   - or model A revised

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