Machine Teaching

- Kevin, Justin, Zilong, Kaikai

Caltech

Contents

- Introduction
- Problem Formulation
- Application on Computer Security
- Application on Education
- Open Questions

Caltech

Introduction

	Active Teacher	Passive Teacher
Active Learner	Interactive Machine Teaching	Active Learning
Passive Learner	**Machine Teaching**	Passive Learning

Machine Teaching

- 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 θ^*, for example, a specific hyper-plane in SVM.
- The teacher knows θ^* and the student’s learning algorithm A, and teaches by giving the student training examples.

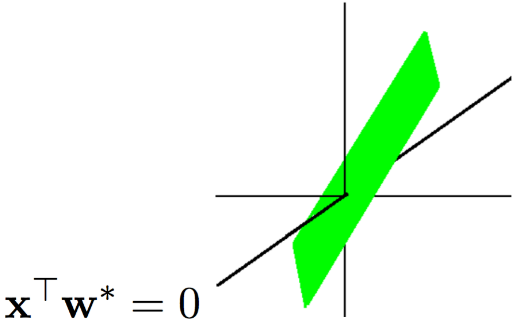
- **Machine teaching aims to design the optimal training set D**

Caltech

Example One

- Steve (the student) runs a linear SVM
- Given a training set with n items $x_i \in R^d$, $y_i \in \{-1, 1\}$, Steve learns $w \in R^d$

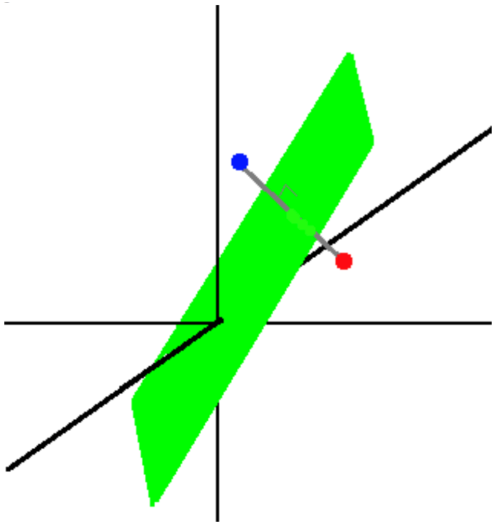
- Tina (the teacher) wants Steve to learn a target w^*



- What is the smallest training set Tina can give Steve?

Example One

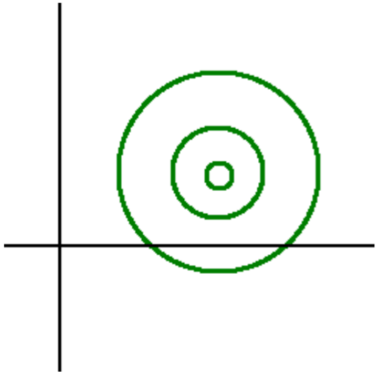
- Tina's $non - iid$ training set with n = 2 items
  One positive sample + One negative sample



Example Two

- Steve wants to estimate a Gaussian density
- Given $x_1 \cdots x_n \in R^d$
- Steve learns

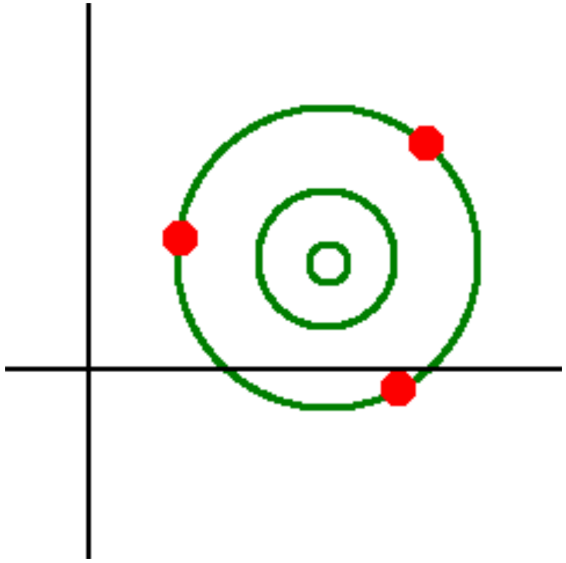
$$\hat{\mu} = \frac{1}{n}\sum x_i$$
$$\hat{\Sigma} = \frac{1}{n-1}\sum (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$



- Tina wants Steve to learn a target Gaussian with (μ^*, Σ^*)

Example Two

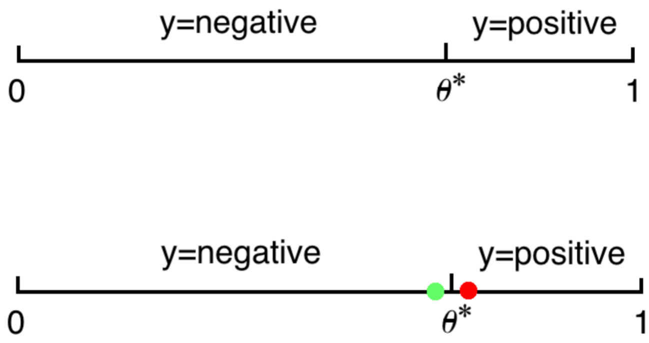
- Tina's minimal training set of $n = d + 1$ tetrahedron vertices



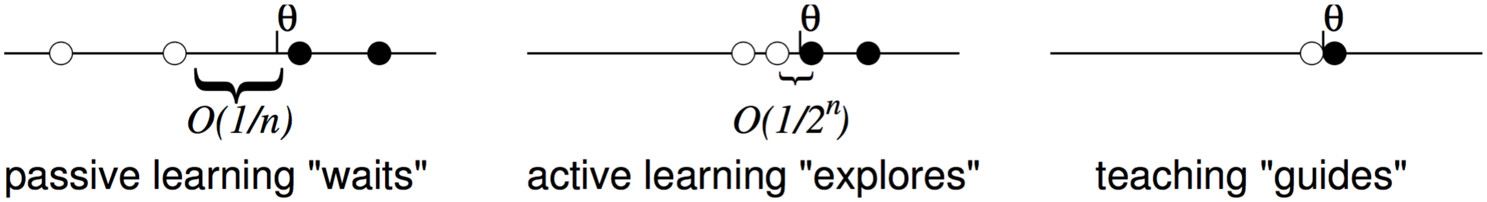
Comparison between 3 algorithms

- Teach a 1D threshold classifier
- For example, given: A = SVM, θ^*=0.6

- What is the smallest training set D?
  - Passive Learning?
  - Active Learning?
  - Machine Teaching?



Comparison between 3 algorithms

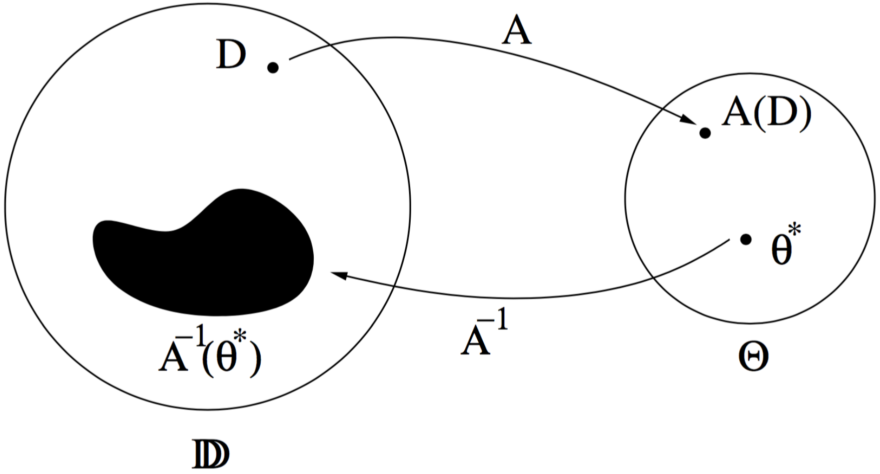


Sample complexity to achieve ϵ error:

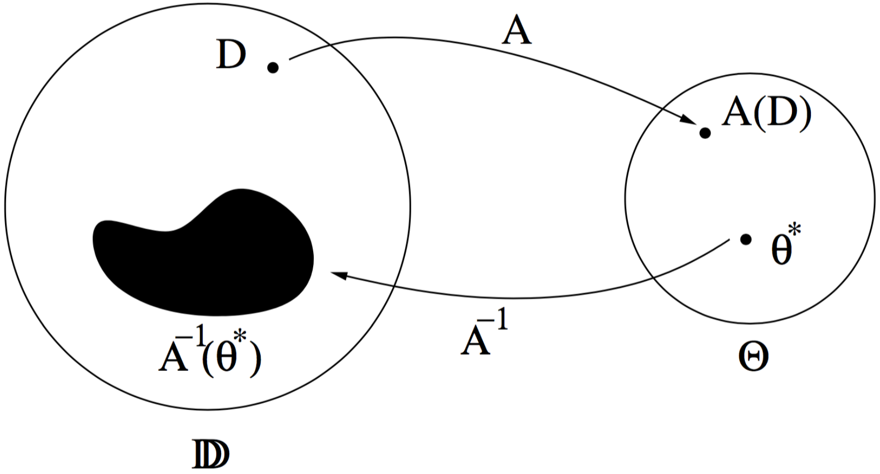
- **Passive learning** $n = O\left(\frac{1}{\epsilon}\right)$
- **Active learning** $n = O\left(\log(\frac{1}{\epsilon})\right)$: needs binary search
- **Machine teaching** $n = O(1)$: teaching dimension [Goldman + Kearns 1995], the teacher knows θ^*, only need two samples

$$n_1 = \left(\theta^* - \frac{\epsilon}{2}, -1\right), n_2 = \left(\theta^* + \frac{\epsilon}{2}, +1\right)$$

Machine Teaching, an inverse problem of Machine Learning



Machine Teaching, an inverse problem of Machine Learning



- Teacher wants Student to learn a target model θ^*
  - not machine learning: Teacher already knows θ^*
  - 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

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

- $\epsilon(D)$: “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
- θ^*: Target model

How to define "Best"?

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

- **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 $D = (x_1, y_1), \cdots, (x_n, y_n)$, we may define
$$\epsilon(D) = \sum_{i,j:y_i \neq y_j} \left\| x_i - x_j \right\|^{-1}$$

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

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

General Machine Teaching Framework

$$\min_{D \in \mathbb{D}, \hat{\theta} \in \Theta} \quad R_T(\hat{\theta}) + \lambda E_T(D)$$

$$s.t. \quad \hat{\theta} = A(D)$$

- $R_T()$: **Teaching risk function** e.g. $\left\|\hat{\theta} - \theta^*\right\|_2^2$
- $E_T()$: **Teaching effort function** e.g. different item costs
- Teacher's search space $\mathbb{D}$: 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]

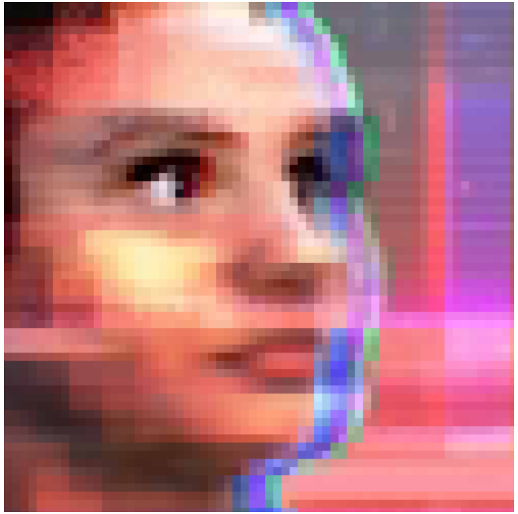
Machine Teaching on Education and Computer Security

$$\min_{D\in\mathbb{D},\hat{\theta}\in\Theta} \quad R_T(\hat{\theta}) + \lambda E_T(D)$$

$$s.t. \quad \hat{\theta} = A(D)$$

- **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

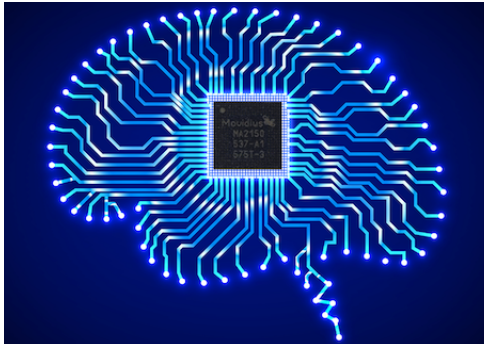
Adversarial Machine Teaching



"[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."

Caltech

Application - Teaching the Wrong Model



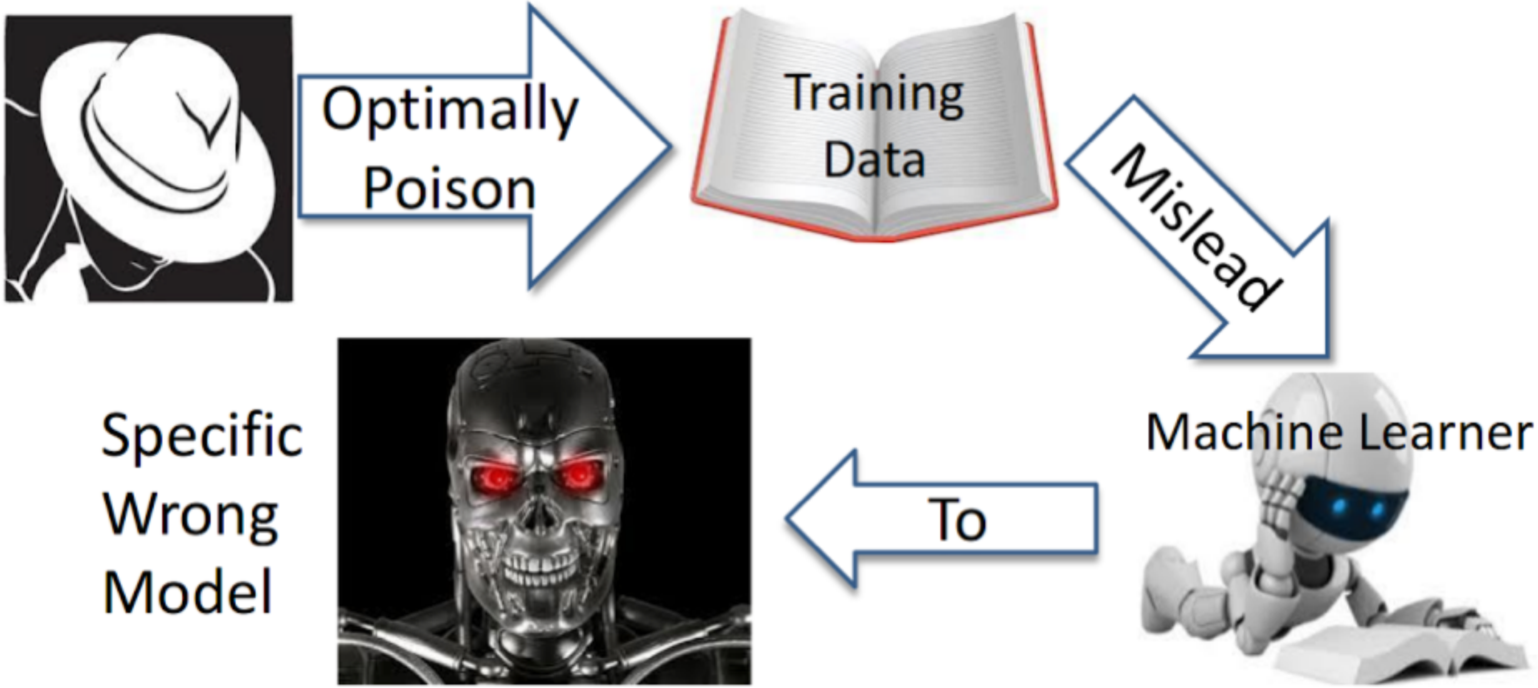
Why do we want to teach the wrong model?

How can we poison the training data?

What is the goal?

Contaminating Training Data

Hacking

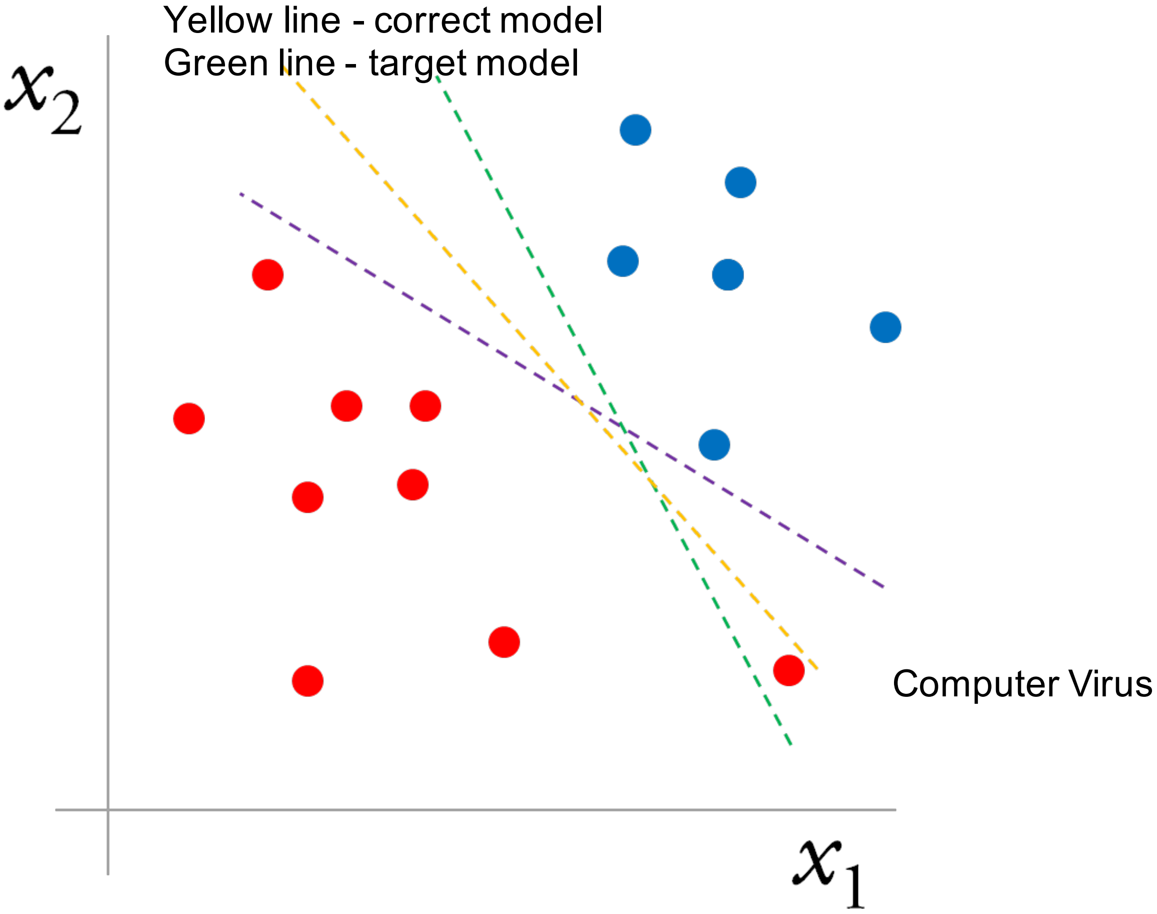


Data Poisoning Attack

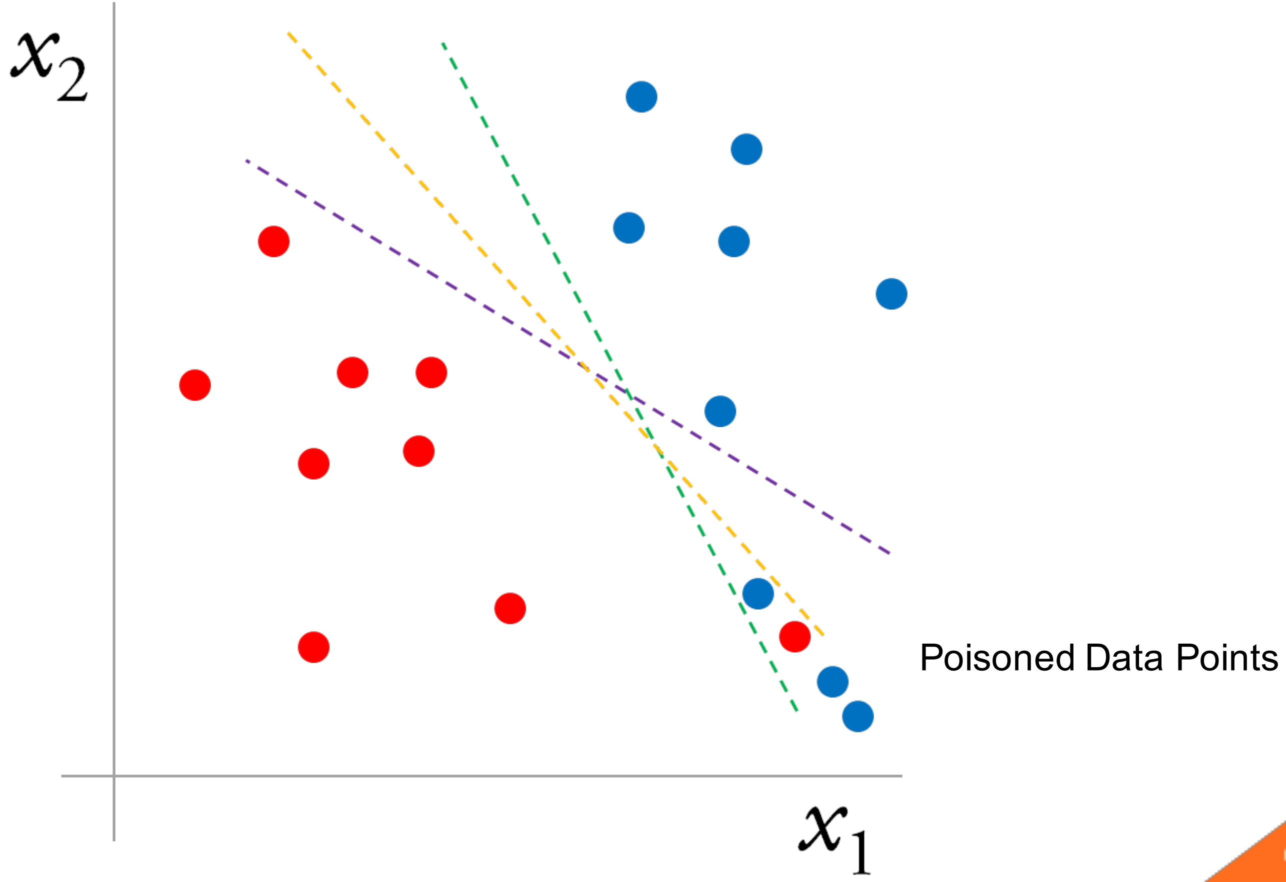
Given: learner A, attack target θ^*, clean training data D_0

Find: the minimum "poison" δ such that $A(D_0 + \delta) = \{\theta^*\}$

Contaminating Training Data



Contaminating Training Data



Training Set Attack Algorithm

$$
\begin{aligned}
& min_{D \in \mathbf{D}, \hat{\theta}_D} && O_A(D, \hat{\theta}_D) \\
& s.t. && \hat{\theta}_D \in \arg\min_{\theta \in \Theta} O_L(D, \theta) \\
& && s.t.\ \mathbf{g}(\theta) \leq 0, \mathbf{h}(\theta) = 0
\end{aligned}
$$

Overall attacker objective function

Learner's objective

Bilevel optimization problem

Training Set Attack Algorithm

Bilevel optimization problems are NP-hard in general.

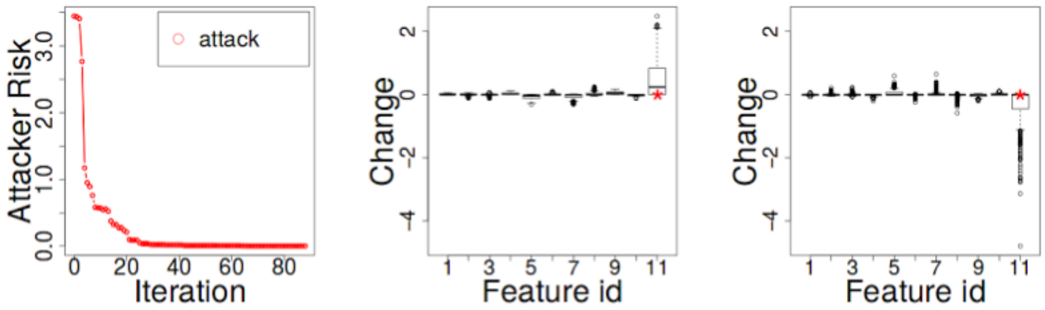
Assume attack space is differentiable.

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)

$$
\begin{aligned}
min_{D \in \mathbf{D}, \theta, \lambda, \mu} \quad & O_A(D, \theta) \\
s.t. \quad & \partial_\theta (O_L(D, \theta) + \lambda^T \mathbf{g}(\theta) + \mu^T \mathbf{h}(\theta)) = 0 \\
& \lambda_i g_i(\theta) = 0, i = 1 \ldots m \\
& \mathbf{g}(\theta) \leq 0, \mathbf{h}(\theta) = 0, \lambda \geq 0
\end{aligned}
$$

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



(a) attacker risk R_A

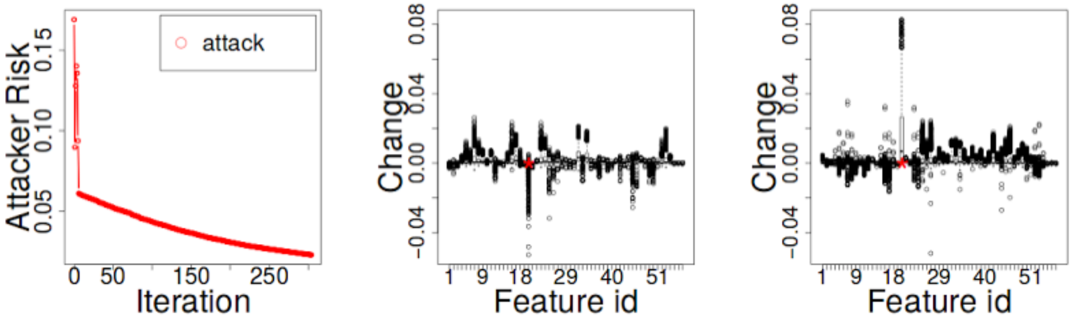
(b) feature changes on positive data

(c) feature changes on negative data

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

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 $E_A = 232$

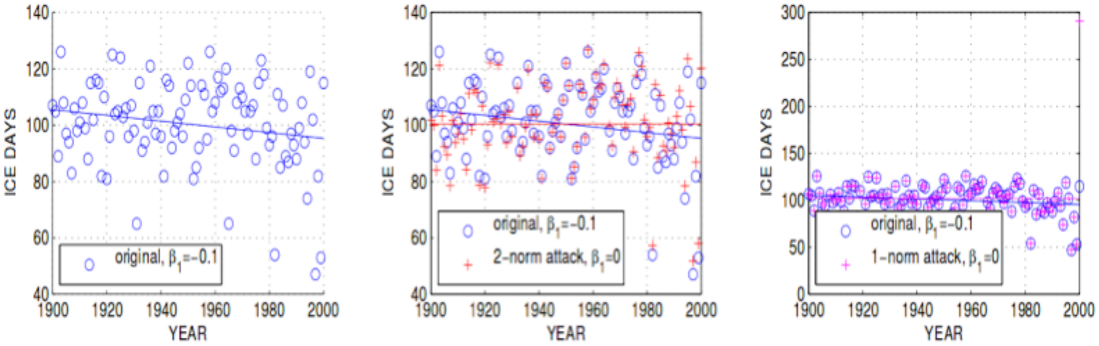


(a) attacker risk R_A (b) feature changes on positive data (c) feature changes on negative data

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

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

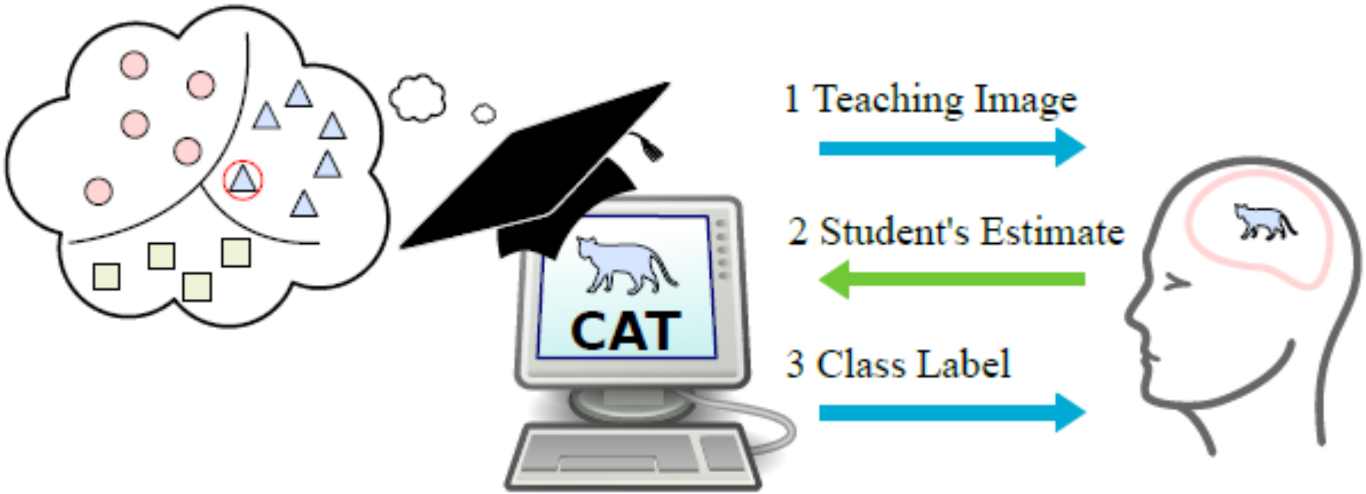


OLS on D_0 | ℓ_2 attacker effort | ℓ_1 attacker effort

Figure 3: Training-set attack on OLS

Machine Teaching on Education

Education System Overview



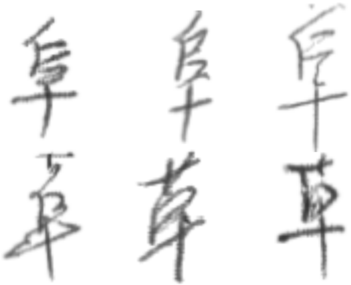
Two fundamental questions

- 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



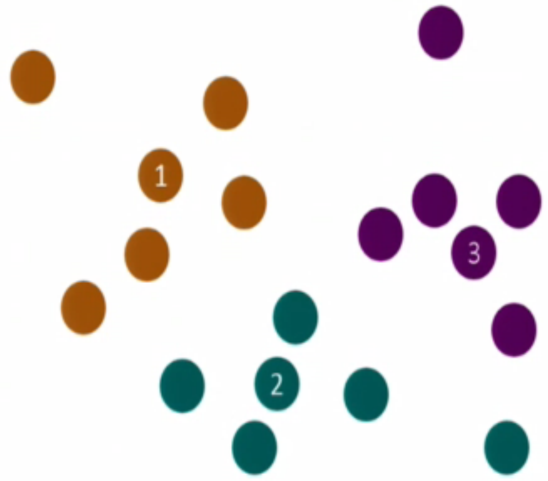
- The goal is to choose teaching images that will maximize the student's classification ability in the minimum amount of teaching time

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

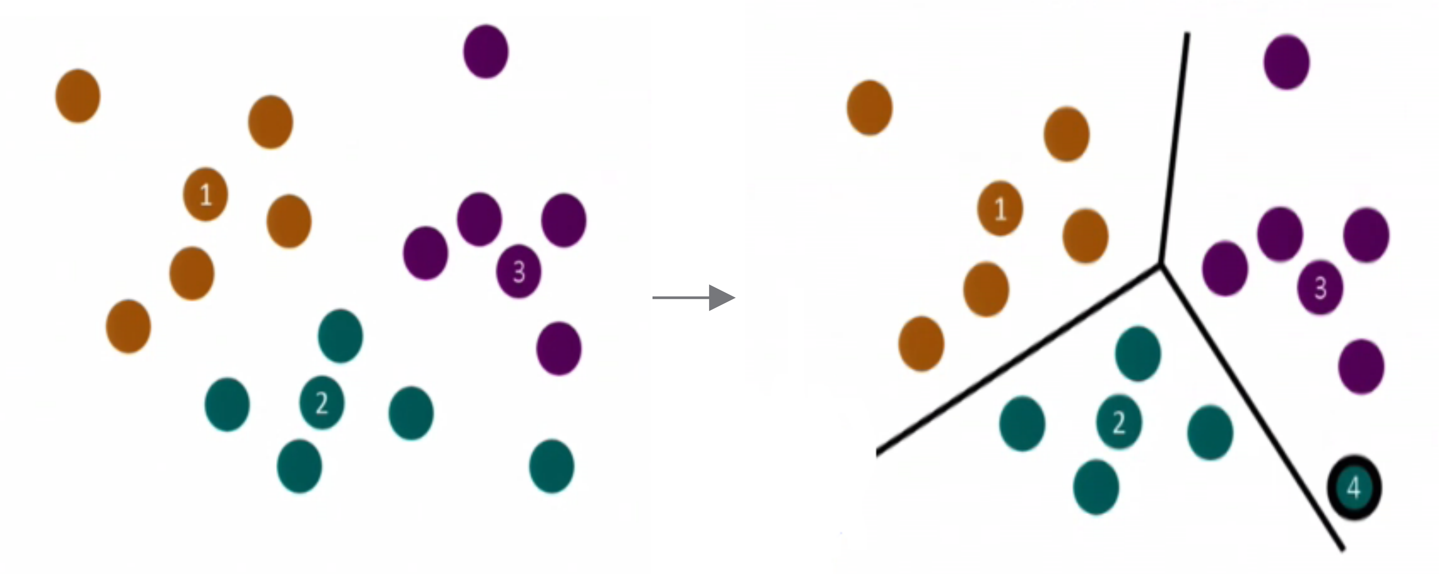
Teaching Strategy

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



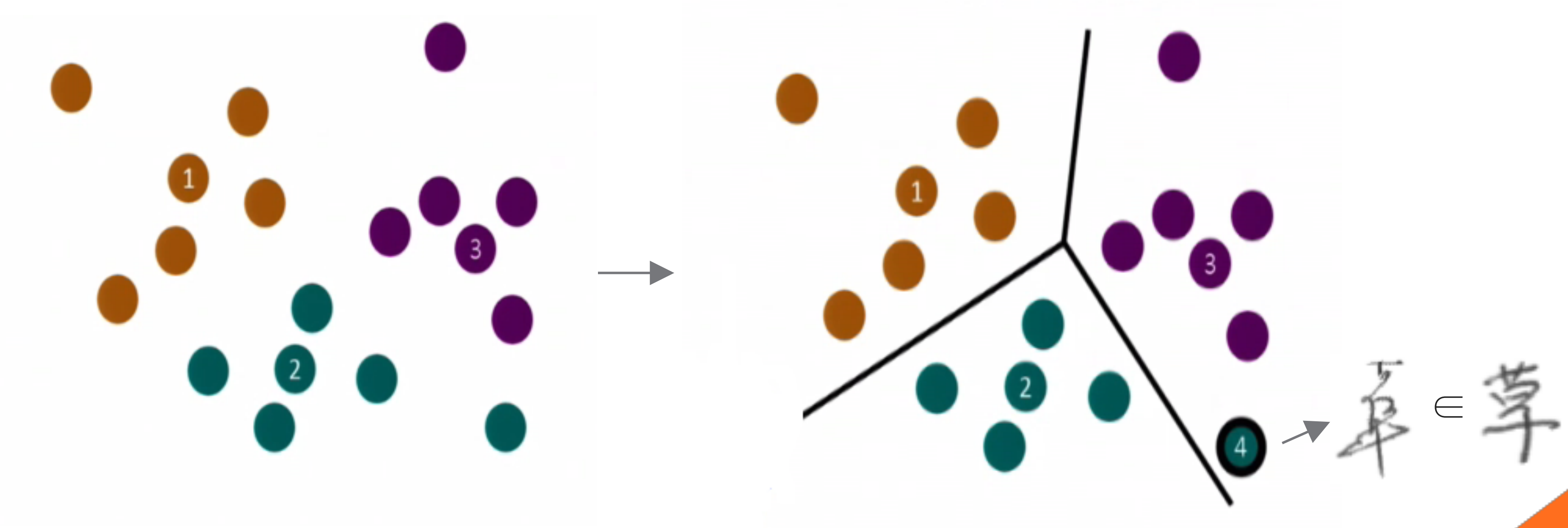
Teaching Strategy

- “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



Teaching Strategy

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

all unlabeled images

probability of ground truth class if u was labeled correctly

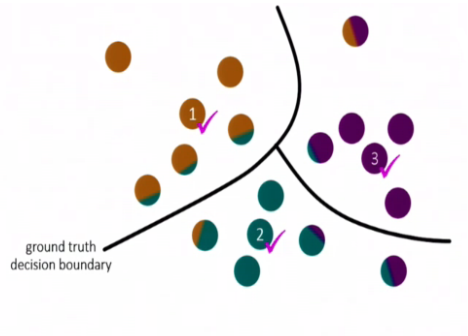
$$u = \underset{u}{\operatorname{argmin}} \sum_{x_i} \left[1 - p^{u+}(\bar{y}_i \mid x_i)\right]$$

next image

all unlabeled images except for u

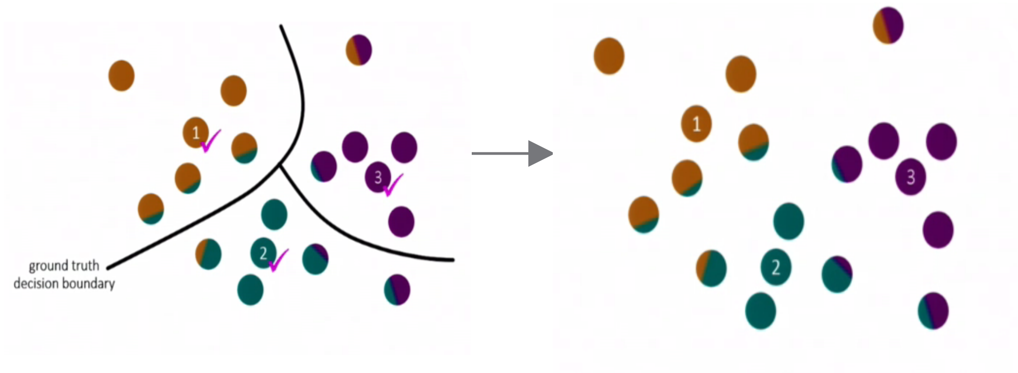
Teaching Strategy

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



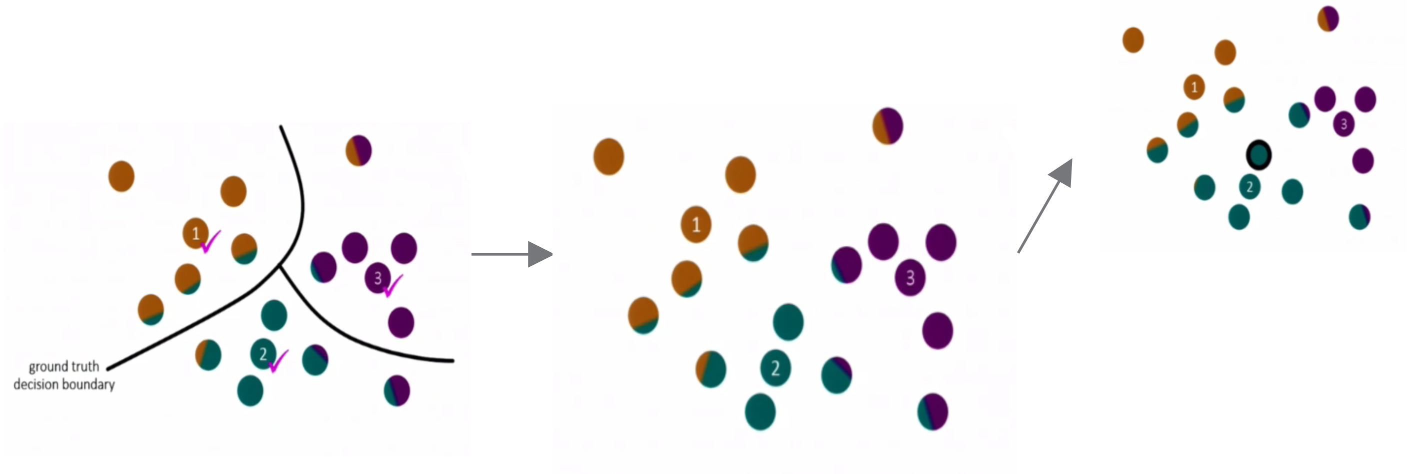
Teaching Strategy

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



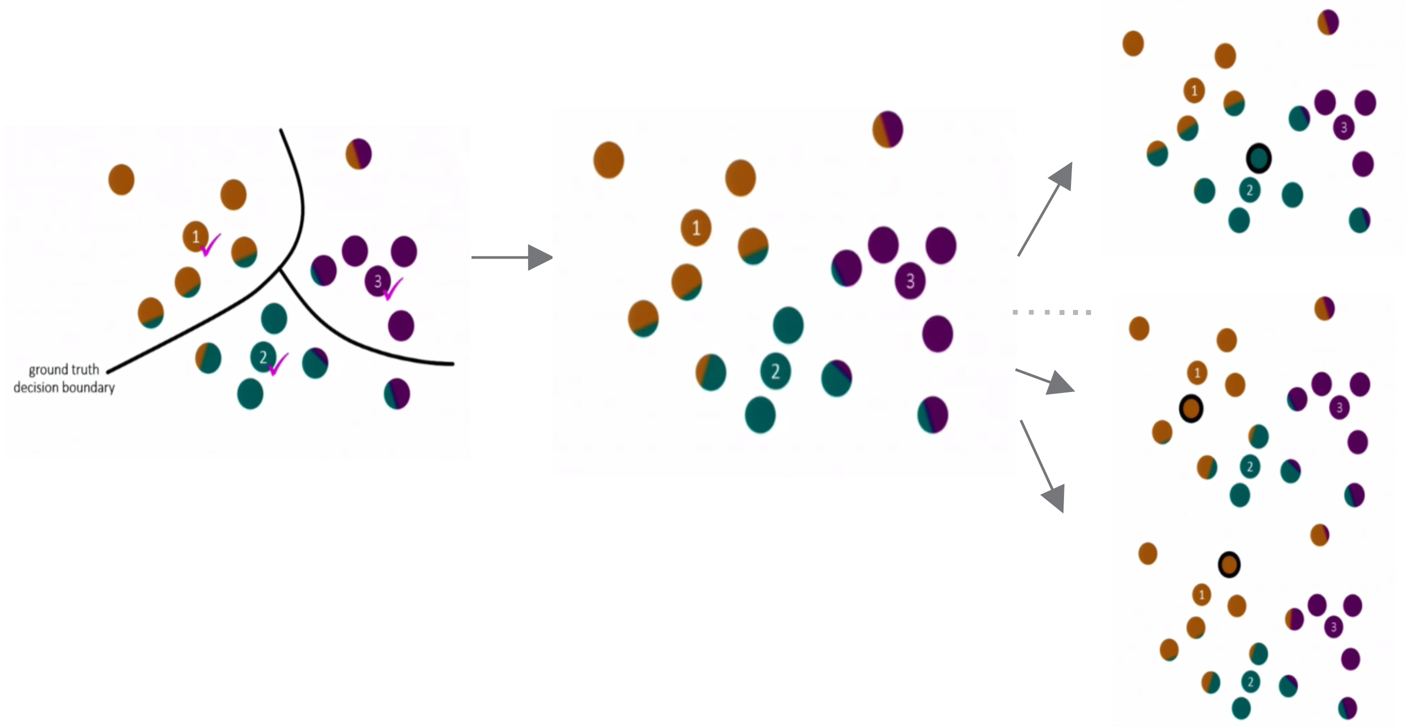
Teaching Strategy

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



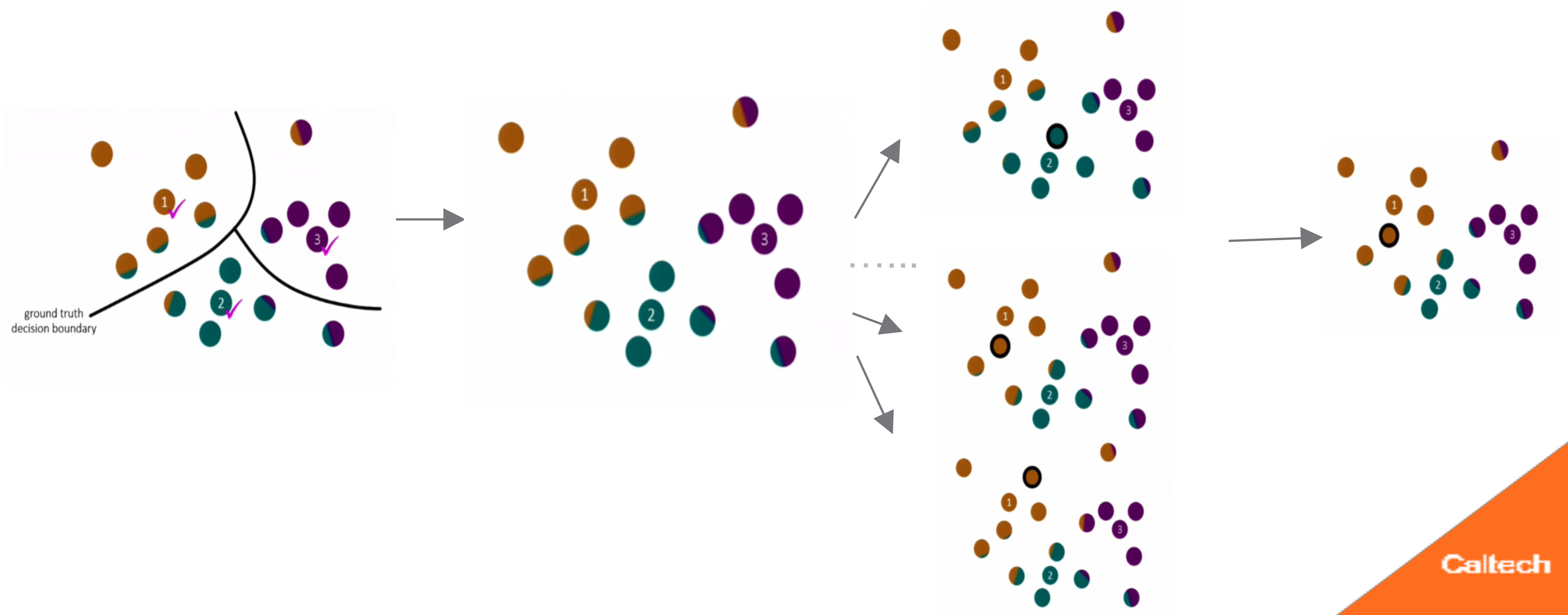
Teaching Strategy

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

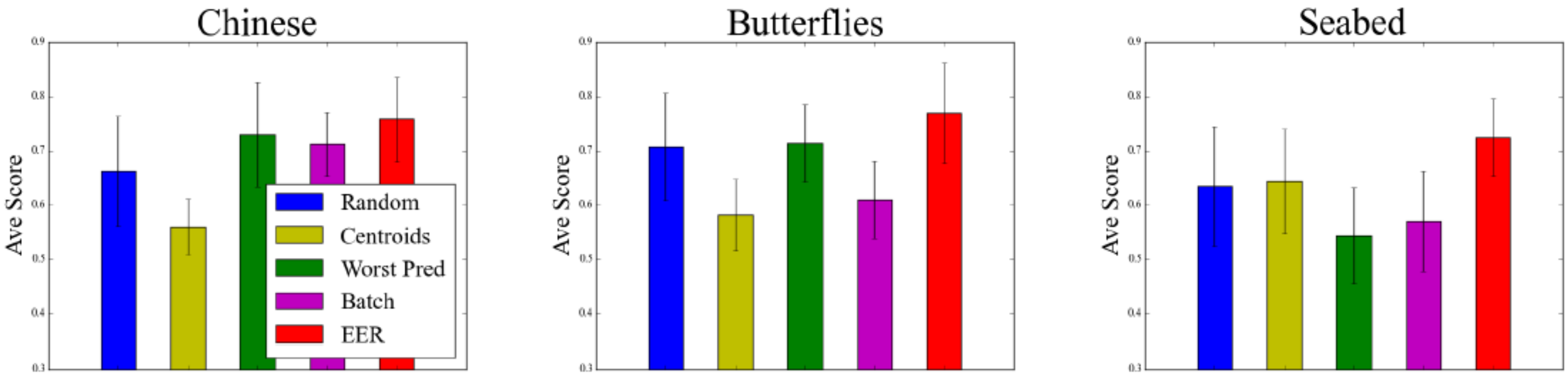


Teaching Strategy

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



Performance comparison



Teaching process

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?

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

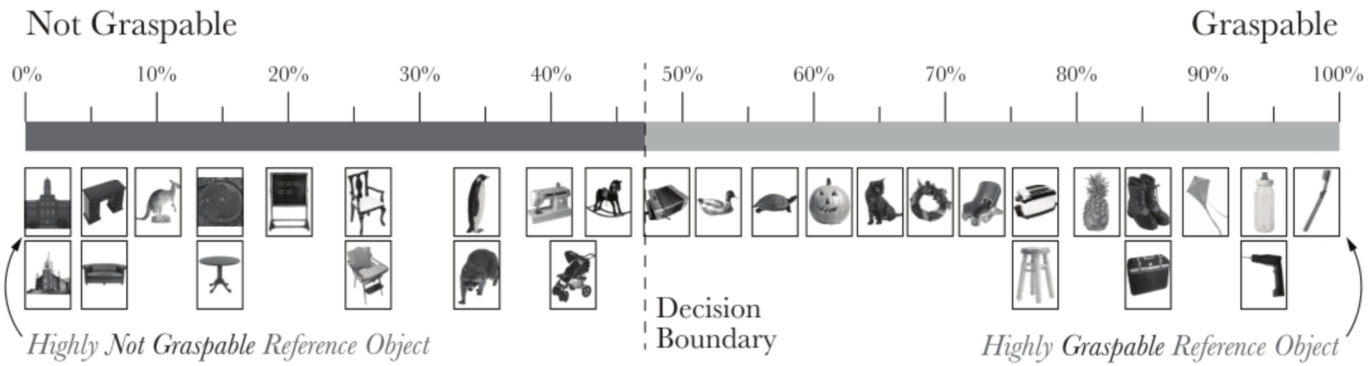
Alternate easy to hard $(x_1, 0), (x_n, 1), (x_2, 0), (x_{n-1}, 1), \ldots, (x_j, 0), (x_{j+1}, 1)$

Human Behavior studies

31 Volunteers were given the task of teaching whether an object in a picture is graspable or not

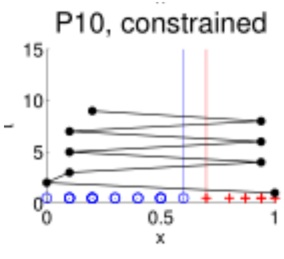
Target was a robot that simply followed motion in the room and did not learn anything

Each participant had their own labeling of graspable or not graspable

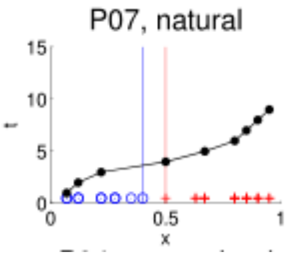


Three Major Human Teaching Strategies

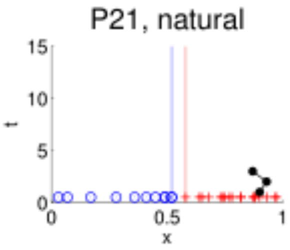
1. **Extreme strategy** - starts with objects at extremes and moves towards decision boundary (14/31)
2. **Linear Strategy** - moves from one side to the other (14/31)
3. **Positive-only strategy** - only gave examples of objects that were graspable. (3/31)
4. None used the boundary strategy, and people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



Extreme



Linear



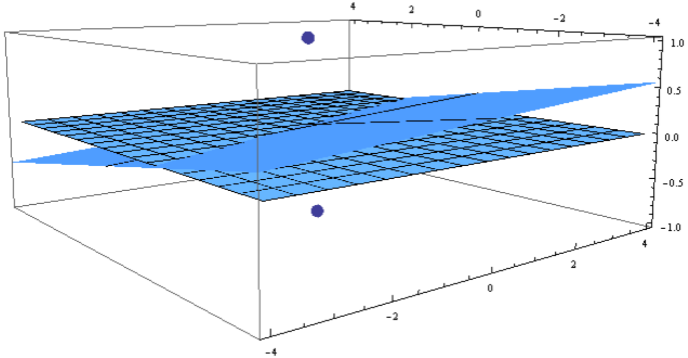
Positive-only

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!

Starting from Extreme Teaching is Asymptotically Optimal

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 $R(2) = \frac{(\frac{1}{2} - b)^2 + (a - \frac{1}{2})^2 + c}{2(a - b + c)}$ where c is the sum over the non-relevant dimmensions of the difference of x_{1j} and x_{2j}

Risk is achieved when $a = \frac{\sqrt{c^2 + 2c} - c + 1}{2}$

Minimizer is a=1, b=0 when $d \to \infty$

Teaching Sequence should Gradually Approach Boundary

Assumption: Teacher only cares about the 1st dimmension.

Collorary: All other dimmensions can be treated as random variables

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

Choosing extreme points makes sure that the other V_k for k != 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 hypothes left are good

Potential Takeaways on Human Learning

Humans utilize a multidimmensional 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*

3. Optimal Teaching for Limited-Capacity Human Learners, *Kaustubh Patil, Xiaojin Zhu, Lukasz Kopec, and Bradley Love.*

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

5. *http://pages.cs.wisc.edu/~jerryzhu/machineteaching/ (presentation slides & papers)*

Thank you!

Q&A