Machine Teaching and its Applications

Jerry Zhu

University of Wisconsin-Madison

Jan. 8, 2018

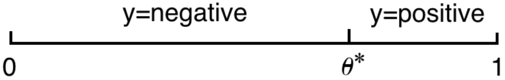
Introduction

Machine teaching

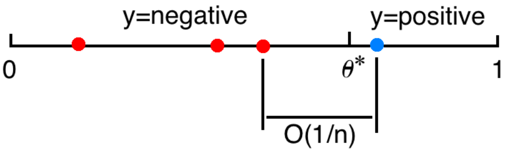
Given target model θ^*, learner A
Find the best training set D so that

$$A(D) \approx \theta^*$$

Passive learning, active learning, teaching

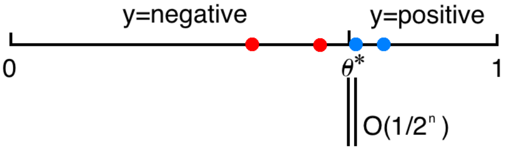


Passive learning



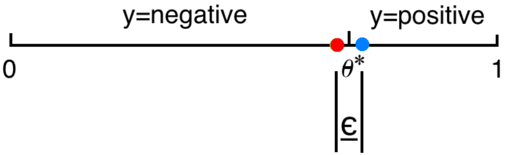
with large probability $|\hat{\theta} - \theta^*| = O(n^{-1})$

Active learning



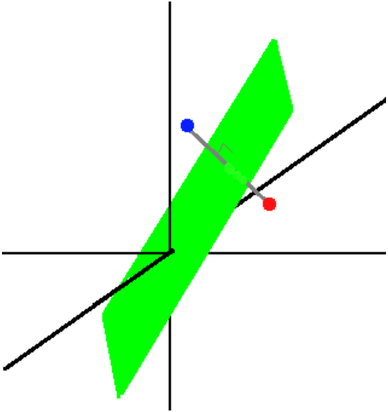
$$|\hat{\theta} - \theta^*| = O(2^{-n})$$

Machine teaching



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

Another example: teaching hard margin SVM



$TD = 2$ vs. $VC = d + 1$

Machine learning vs. machine teaching

- learning (D given, learn $\hat{\theta}$)

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{(x,y)\in D} \ell(x, y, \theta) + \lambda\|\theta\|^2$$

- 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 not $i.i.d.$
  - synthetic or pool-based

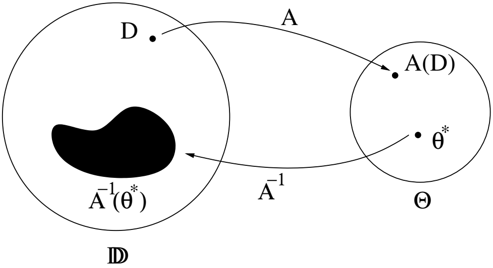
Why bother if we already know θ^*?

teach·ing
/'teCHiNG/
noun

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

The coding view

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



Machine teaching generic form

$$\min_{D,\hat{\theta}} \quad \text{TeachingRisk}(\hat{\theta}) + \eta \text{TeachingCost}(D)$$

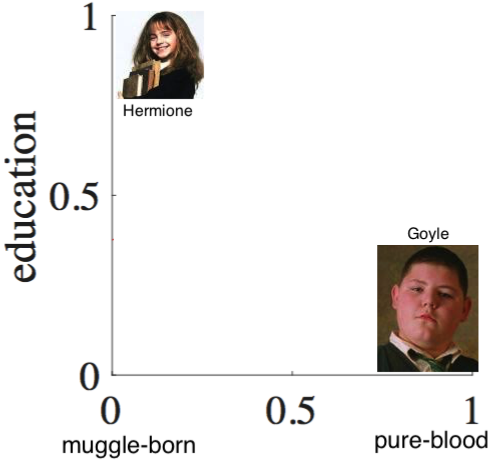
$$\text{s.t.} \quad \hat{\theta} = \text{MachineLearning}(D)$$

Fascinating things I will not discuss today

- probing graybox learners
- teaching by features, pairwise comparisons
- learner anticipates teaching
- reward shaping, reinforcement learning, optimal control

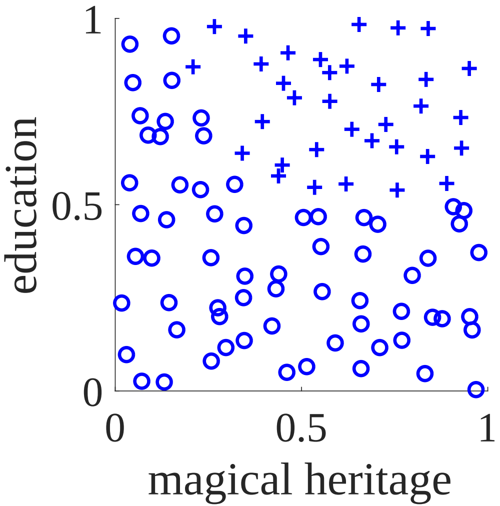
Machine learning debugging

Harry Potter toy example



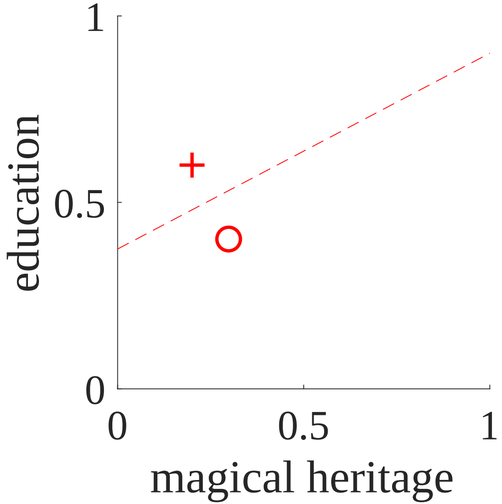
Labels y contain historical bias

- $+$ hired by the Ministry of Magic
- o no

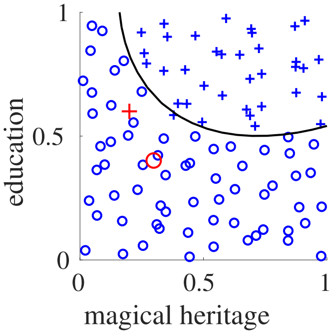


Trusted items $(\tilde{x}, \tilde{y})$

- expensive
- insufficient to learn



Idea

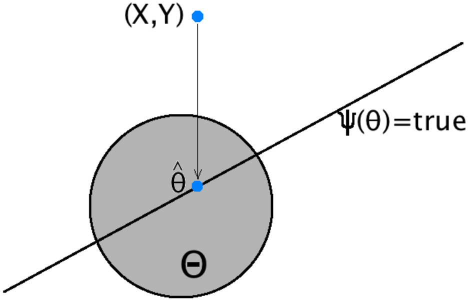


Flip training labels and re-train model to agree with trusted items.

$$\Psi(\hat{\theta}) := [\hat{\theta}(\tilde{x}) = \tilde{y}]$$

Not our goal: only to learn a better model

$$\min_{\theta \in \Theta} \quad \ell(X, Y, \theta) + \lambda \|\theta\|$$
$$\text{s.t.} \quad \Psi(\theta) = \text{true}$$

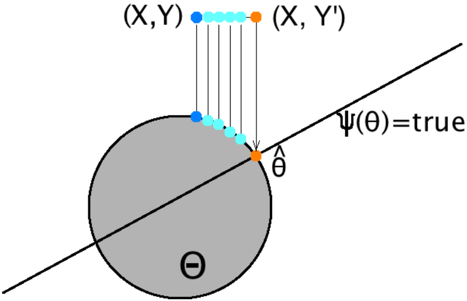


Our goal: To find bugs and learn a better model

$$\min_{Y', \hat{\theta}} \quad \|Y - Y'\|$$

$$\text{s.t.} \quad \Psi(\hat{\theta}) = \text{true}$$

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \ell(X, {\color{red}Y'}, \theta) + \lambda \|\theta\|$$



Solving combinatorial, bilevel optimization

(Stackelberg game)

step 1. label to probability simplex

$$y_i' \to \delta_i \in \Delta$$

step 2. counting to probability mass

$$\|Y' - Y\| \to \frac{1}{n}\sum_{i=1}^{n}(1 - \delta_{i,y_i})$$

step 3. soften postcondition

$$\hat{\theta}(\tilde{X}) = \tilde{Y} \to \frac{1}{m}\sum_{i=1}^{m}\ell(\tilde{x}_i, \tilde{y}_i, \theta)$$

Continuous now, but still bilevel

$$\begin{aligned} \operatorname*{argmin}_{\delta \in \Delta^n, \hat{\theta}} \quad & \frac{1}{m} \sum_{i=1}^{m} \ell(\tilde{x}_i, \tilde{y}_i, \hat{\theta}) + \gamma \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_{i,y_i}) \\ \textsf{s.t.} \quad & \hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_{ij} \ell(x_i, j, \theta) + \lambda \|\theta\|^2 \end{aligned}$$

Removing the lower level problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_{ij} \ell(x_i, j, \theta) + \lambda \|\theta\|^2$$

step 4. the KKT condition

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_{ij} \nabla_{\theta} \ell(x_i, j, \theta) + 2\lambda\theta = 0$$

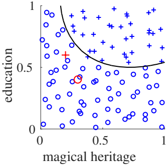
step 5. plug implicit function $\theta(\delta)$ into upper level problem

$$\operatorname*{argmin}_{\delta} \quad \frac{1}{m} \sum_{i=1}^{m} \ell(\tilde{x}_i, \tilde{y}_i, \theta(\delta)) + \gamma \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_{i,y_i})$$

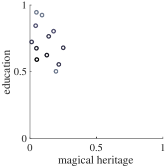
step 6. compute gradient ∇_{δ} with implicit function theorem

Software available.

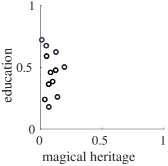
Harry Potter Toy Example



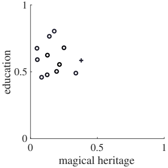
data



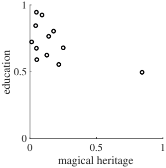
our debugger



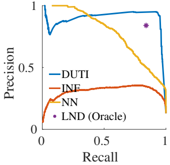
influence function



nearest neighbor



label noise detection



average PR

Adversarial Attacks

Level 1 attack: test item $(\tilde{x}, \tilde{y})$ manipulation

$$\begin{aligned} \min_{x} \quad & \|\tilde{x} - x\|_p \\ \text{s.t.} \quad & \hat{\theta}(x) \neq \tilde{y}. \end{aligned}$$

Model $\hat{\theta}$ fixed.

Level 2 attack: training set poisoning

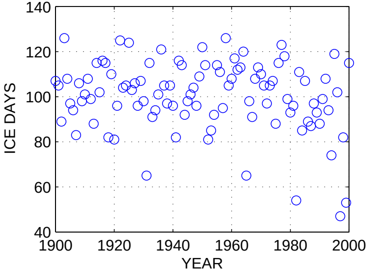
$$\begin{aligned} \min_{D} \quad & \|D_0 - D\|_p \\ \text{s.t.} \quad & \Psi(A(D)) \end{aligned}$$

e.g. $\Psi(\theta) := [\theta(\tilde{x} + \epsilon) = y']$

Level 2 attack on regression

Lake Mendota, Wisconsin

(x, y)

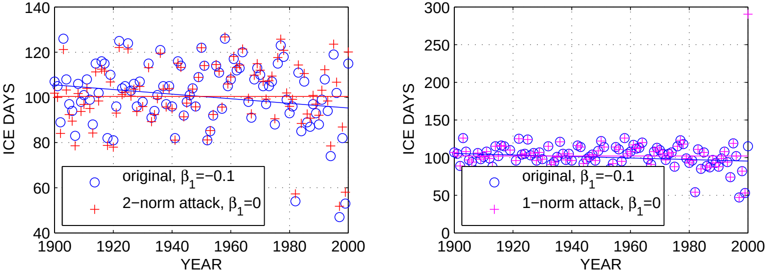


Level 2 attack on regression

$$\min_{\delta, \tilde{\beta}} \quad \|\delta\|_p$$

$$\text{s.t.} \quad \tilde{\beta}_1 \geq 0$$

$$\tilde{\beta} = \operatorname*{argmin}_{\beta} \|(\mathbf{y} + \delta) - X\beta\|^2$$



minimize $\|\delta\|_2^2$ minimize $\|\delta\|_1$

[Mei, Z 15a]

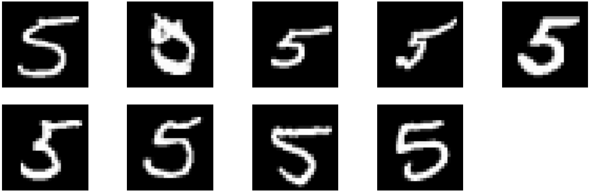
Level 2 attack on latent Dirichlet allocation

[Mei, Z 15b]

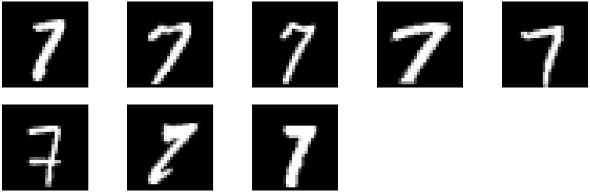
Guess the classification task

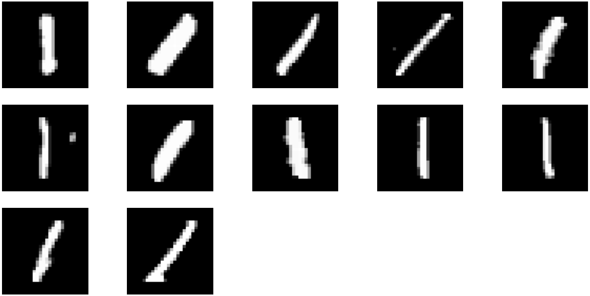
Ready?

Guess the classification task (1)



Guess the classification task (2)

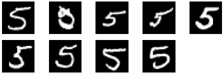




Guess the classification task (3)

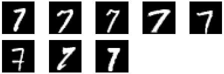
+	The Angels won their home opener against the Brewers today before 33,000+ at Anaheim Stadium, 3-1 on a 3-hitter by Mark La
+	I'm *very* interested in finding out how I might be able to get two tickets for the All Star game in Baltimore this year.
+	I know there's been a lot of talk about Jack Morris' horrible start, but what about Dennis Martinez. Last I checked he's 0-3 with 6+ I ...
-	Where are all the Bruins fans??? Good point - there haven't even been any recent posts about Ulf!
-	I agree thouroughly!! Screw the damn contractual agreements! Show the exciting hockey game. They will lose fans of ESPN
-	TV Coverage - NHL to blame! Give this guy a drug test, and some Ridalin whale you are at it.
	...

Did you get it right? (1)

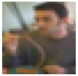


 gun vs. phone 

Did you get it right? (2)



woman vs. man

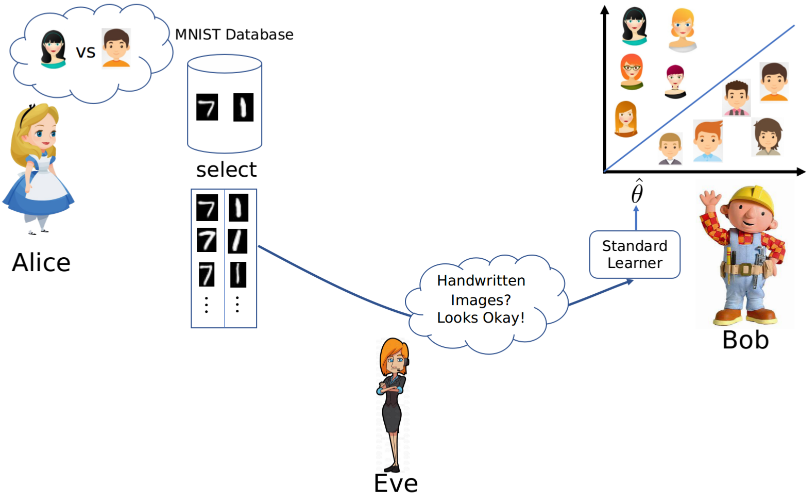


Did you get it right? (3)

20Newsgroups soc.religion.christian vs. alt.atheism

+	: T H E W I T N E S S & P R O O F O F : : J E S U S C H R I S T ' S R E S U R R E C T I O N : : F R O M T H E D E A D :
+	I've heard it said that the accounts we have of Christs life and ministry in the Gospels were actually written many years after
-	An Introduction to Atheism by mathew <mathew@mantis.co.uk>
-	Computers are an excellent example... of evolution without "a" creator.

Camouflage attack



Social engineering against Eve

Camouflage attack

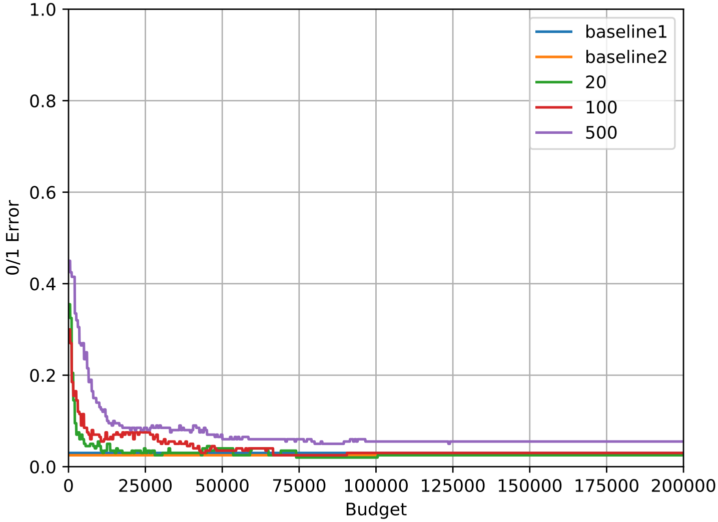
Alice knows

- S (e.g. women, men)
- C (e.g. 7, 1)
- A
- Eve's inspection function MMD (maximum mean discrepancy)

finds

$$\begin{aligned} \underset{D \subseteq C}{\operatorname{argmin}} \quad & \sum_{(x,y) \in S} \ell(A(D), x, y) \\ \text{s.t.} \quad & \mathrm{MMD}(D, C) \leq \alpha \end{aligned}$$

Test set error



(Gun vs. Phone) camouflaged as (5 vs. 2)

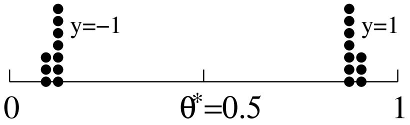
Enhance human learning

"Hedging"

1. Find D^* to maximize accuracy on cognitive model A
2. Give humans D^*
   - either human performance improved
   - or cognitive model A revised

Human learning example 1

[Patil et al. 2014]



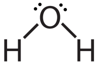
$A =$ kernel density estimator

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

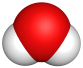
(statistically significant)

Human learning example 2

[Sen et al. in preparation]



Lewis



space-filling

A = neural network

human trained on	human test error
random	28.6%
expert	28.1%
D^*	**25.1%**

(statistically significant)

Human learning example 3

[Nosofsky & Sanders, Psychonomics 2017]

$A =$ Generalized Context Model (GCM)

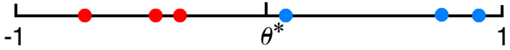
human trained on	human accuracy
random	67.2%
coverage	71.2%
D^*	69.3%

D^* not better on humans (experts revising the model)

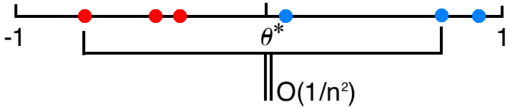
Super Teaching

Super teaching example 1

Let $D \overset{iid}{\sim} U(0,1)$, $A(D) =$SVM.



whole training set $O(n^{-1})$



most symmetrical pair $O(n^{-2})$

(Not training set reduction)

Super teaching example 2

Let $D \overset{iid}{\sim} N(0,1)$, $A(D) = \frac{1}{|D|}\sum_{x \in D} x$.

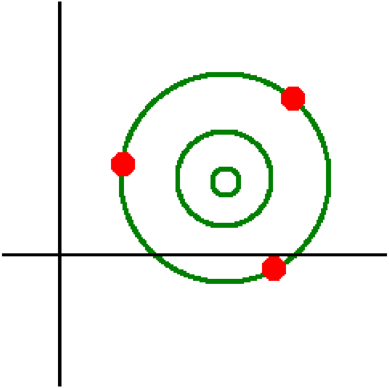
Theorem: Fix k. For n sufficiently large, with large probablity

$$\min_{S \subset D, |S|=k} |A(S)| \leq \frac{k^{k-\epsilon}}{\sqrt{k}} n^{-k+\frac{1}{2}+2\epsilon} |A(D)|$$

Thank you

- email me for "Machine Teaching Tutorial"
- `http://www.cs.wisc.edu/~jerryzhu/machineteaching/`
- Collaborators:
  - **Security**: Scott Alfeld, Paul Barford
  - **HCI**: Saleema Amershi, Bilge Mutlu, Jina Suh
  - **Programming language**: Aws Albarghouthi, Loris D'Antoni, Shalini Ghosh
  - **Machine learning**: Ran Gilad-Bachrach, Manuel Lopes, Yuzhe Ma, Christopher Meek, Shike Mei, Robert Nowak, Gorune Ohannessian, Philippe Rigollet, Ayon Sen, Patrice Simard, Ara Vartanian, Xuezhou Zhang
  - **Optimization**: Ji Liu, Stephen Wright
  - **Psychology**: Bradley Love, Robert Nosofsky, Martina Rau, Tim Rogers

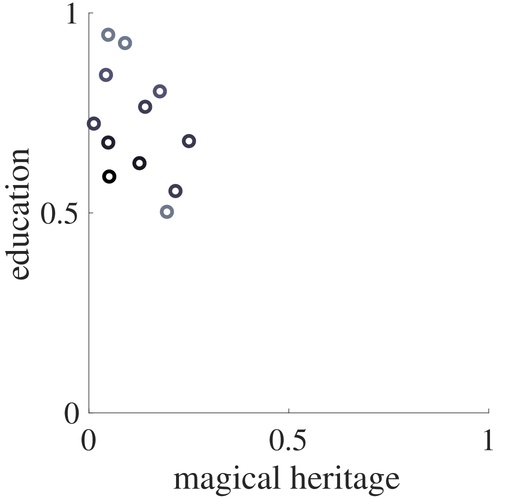
Yet another example: teach Gaussian density



$TD = d + 1$: tetrahedron vertices

Proposed bugs

- flipping them makes re-trained model agree with trusted items
- given to experts to interpret



The ML pipeline

$\boxed{\text{data } (X, Y)} \rightarrow \boxed{\text{learner } \ell} \rightarrow \boxed{\text{parameters } \lambda} \rightarrow \boxed{\text{model } \hat{\theta}}$

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \ell(X, Y, \theta) + \lambda ||\theta||$$

Postconditions

$$\Psi(\hat{\theta})$$

Examples:

- "the learned model must correctly predict an important item $(\tilde{x}, \tilde{y})$"
$$\hat{\theta}(\tilde{x}) = \tilde{y}$$
- "the learned model must satisfy individual fairness"
$$\forall x, x', |p(y = 1 \mid x, \hat{\theta}) - p(y = 1 \mid x', \hat{\theta})| \leq L\|x - x'\|$$

Bug Assumptions

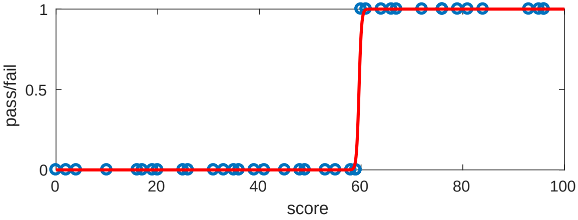
- Ψ satisfied if we were to train through "clean pipeline"
- bugs are changes to the clean pipeline
- Ψ violated on the dirty pipeline

Debugging formulation

$$\begin{aligned}
\min_{Y'} \quad & \|Y' - Y\| \\
\text{s.t.} \quad & \hat{\theta}(\tilde{X}) = \tilde{Y} \\
& \hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i', \theta) + \lambda \|\theta\|^2
\end{aligned}$$

- bilevel optimization (Stackelberg game)
- combinatorial

Another special case: bug in regularization weight



(logistic regression)

Postcondition violated

$\Psi(\hat{\theta})$: Individual fairness (Lipschitz condition)

$$\forall x, x', |p(y = 1 \mid x, \hat{\theta}) - p(y = 1 \mid x', \hat{\theta})| \leq L\|x - x'\|$$

Bug assumption

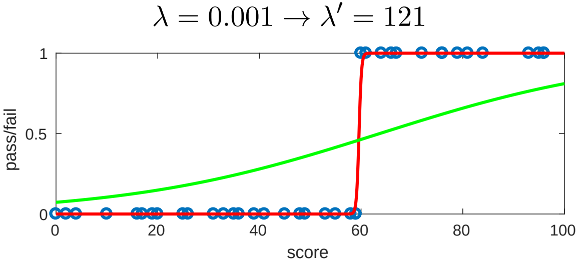
Learner's regularization weight $\lambda = 0.001$ was inappropriate

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \ell(X, Y, \theta) + \lambda \|\theta\|^2$$

Debugging formulation

$$
\begin{aligned}
\min_{\lambda', \hat{\theta}} \quad & (\lambda' - \lambda)^2 \\
\text{s.t.} \quad & \Psi(\hat{\theta}) = \text{true} \\
& \hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \ell(X, Y, \theta) + \lambda' \|\theta\|^2
\end{aligned}
$$

Suggested bug



Guaranteed defense?

Let

$$A(D_0)(\tilde{x}) = \tilde{y}$$

Attacker can use the debug formulation

$$\begin{aligned} D_1 := \operatorname*{argmin}_D \quad & \|D_0 - D\|_p \\ \text{s.t.} \quad & \Psi_1(A(D)) := A(D)(\tilde{x}) \neq \tilde{y} \end{aligned}$$

Defender can use the debug formulation, too

$$\begin{aligned} D_2 := \operatorname*{argmin}_D \quad & \|D_1 - D\|_p \\ \text{s.t.} \quad & \Psi_2(A(D)) := A(D)(\tilde{x}) = \tilde{y} \end{aligned}$$

When does $D_2 = D_0$?