Can Machine Learning Rationalize Simple Human Teaching Behaviors?

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

May 2012

Outline

1. Teaching as a machine learning problem

2. Human teaching behaviors in a 1D task
   - "Graspability"
   - "lines"

3. Our computational rationalize of the human teaching behaviors

Binary classification

- Input space $\mathcal{X} \subseteq \mathbb{R}^d$. item $x \in \mathcal{X}$

Binary classification

- Input space $\mathcal{X} \subseteq \mathbb{R}^d$. item $x \in \mathcal{X}$
- Output space $\mathcal{Y} = \{-1, 1\}$. label $y \in \mathcal{Y}$

Binary classification

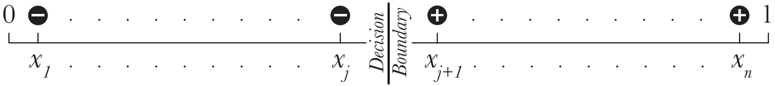
- Input space $\mathcal{X} \subseteq \mathbb{R}^d$. item $x \in \mathcal{X}$
- Output space $\mathcal{Y} = \{-1, 1\}$. label $y \in \mathcal{Y}$
- Unknown test distribution $(x, y) \overset{iid}{\sim} p$

Binary classification

- Input space $\mathcal{X} \subseteq \mathbb{R}^d$. item $x \in \mathcal{X}$
- Output space $\mathcal{Y} = \{-1, 1\}$. label $y \in \mathcal{Y}$
- Unknown test distribution $(x, y) \stackrel{iid}{\sim} p$
- Goal: pick classifier $f \in \mathcal{F}, f : \mathcal{X} \mapsto \mathcal{Y}$ to minimize $\mathbb{E}_p[f(x) \neq y]$

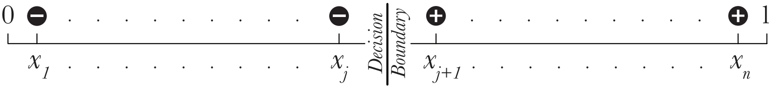
Binary classification

- Input space $\mathcal{X} \subseteq \mathbb{R}^d$. item $x \in \mathcal{X}$
- Output space $\mathcal{Y} = \{-1, 1\}$. label $y \in \mathcal{Y}$
- Unknown test distribution $(x, y) \overset{iid}{\sim} p$
- Goal: pick classifier $f \in \mathcal{F}, f : \mathcal{X} \mapsto \mathcal{Y}$ to minimize $\mathbb{E}_p[f(x) \neq y]$
- Example: $d = 1$, $\mathcal{X} = [0, 1]$, $\mathcal{F} = \{1_{x \geq \theta} \mid \theta \in [0, 1]\}$ threshold functions



Binary classification

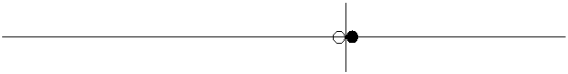
- Input space $\mathcal{X} \subseteq \mathbb{R}^d$. item $x \in \mathcal{X}$
- Output space $\mathcal{Y} = \{-1, 1\}$. label $y \in \mathcal{Y}$
- Unknown test distribution $(x, y) \overset{iid}{\sim} p$
- Goal: pick classifier $f \in \mathcal{F}, f : \mathcal{X} \mapsto \mathcal{Y}$ to minimize $\mathbb{E}_p[f(x) \neq y]$
- Example: $d = 1$, $\mathcal{X} = [0, 1]$, $\mathcal{F} = \{1_{x \geq \theta} \mid \theta \in [0, 1]\}$ threshold functions



- Teaching/learning by labeled examples only!

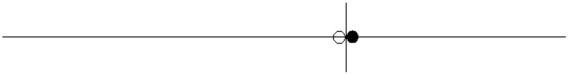
Optimal computational teaching theory

- The teacher picks two items $(x_1, y_1 = -1), (x_2, y_2 = 1)$ next to the decision boundary



Optimal computational teaching theory

- The teacher picks two items $(x_1, y_1 = -1), (x_2, y_2 = 1)$ next to the decision boundary



- Assuming the learner knows $\mathcal{F}$

Optimal computational teaching theory

- The teacher picks two items $(x_1, y_1 = -1), (x_2, y_2 = 1)$ next to the decision boundary

- Assuming the learner knows $\mathcal{F}$
- $n = 2$, teaching accomplished!

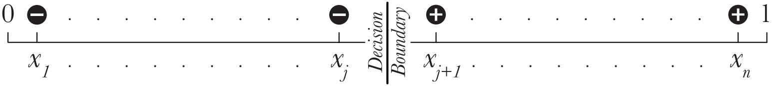
Optimal computational teaching theory

- The teacher picks two items $(x_1, y_1 = -1), (x_2, y_2 = 1)$ next to the decision boundary

- Assuming the learner knows $\mathcal{F}$
- $n = 2$, teaching accomplished!
- Formalized by the notion of teaching dimension

The teaching dimension [Goldman and Kearns 1995]

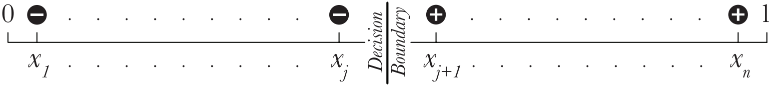
$\mathcal{X} = \{x_1, \ldots, x_n\}$



- teaching set of f with respect to $\mathcal{F}$: subset of $\mathcal{X}$ consistent with only f, not any other $f' \in \mathcal{F}$

The teaching dimension [Goldman and Kearns 1995]

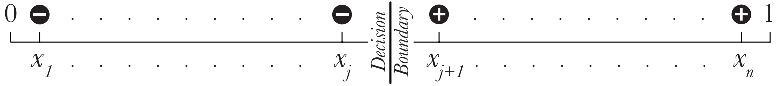
$\mathcal{X} = \{x_1, \ldots, x_n\}$



- teaching set of f with respect to $\mathcal{F}$: subset of $\mathcal{X}$ consistent with only f, not any other $f' \in \mathcal{F}$
- $TD(f)$: size of the smallest teaching set of f (1 or 2)

The teaching dimension [Goldman and Kearns 1995]

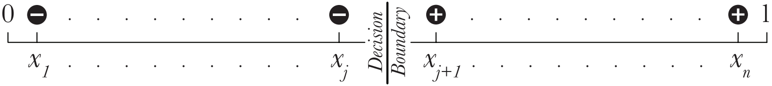
$\mathcal{X} = \{x_1, \ldots, x_n\}$



- teaching set of f with respect to $\mathcal{F}$: subset of $\mathcal{X}$ consistent with only f, not any other $f' \in \mathcal{F}$
- $TD(f)$: size of the smallest teaching set of f (1 or 2)
- $TD(\mathcal{F})$: $TD(f)$ for the hardest $f \in \mathcal{F}$ (2)

The teaching dimension [Goldman and Kearns 1995]

$\mathcal{X} = \{x_1, \ldots, x_n\}$



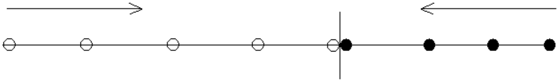
- teaching set of f with respect to $\mathcal{F}$: subset of $\mathcal{X}$ consistent with only f, not any other $f' \in \mathcal{F}$
- $TD(f)$: size of the smallest teaching set of f (1 or 2)
- $TD(\mathcal{F})$: $TD(f)$ for the hardest $f \in \mathcal{F}$ (2)
- Implication: for the 1D example optimal teaching should start around the decision boundary.

Curriculum learning [Bengio et al. 2009]

- An alternative suggestion for good teaching

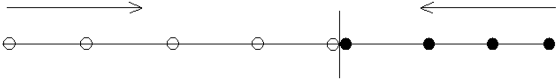
Curriculum learning [Bengio et al. 2009]

- An alternative suggestion for good teaching
- Teaching should start from easy to hard, i.e., outside to inside.



Curriculum learning [Bengio et al. 2009]

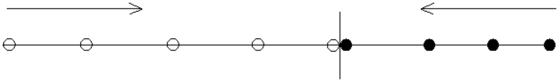
- An alternative suggestion for good teaching
- Teaching should start from easy to hard, i.e., outside to inside.



- A principle motivated by:

Curriculum learning [Bengio et al. 2009]

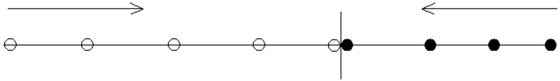
- An alternative suggestion for good teaching
- Teaching should start from easy to hard, i.e., outside to inside.



- A principle motivated by:
  - psychology

Curriculum learning [Bengio et al. 2009]

- An alternative suggestion for good teaching
- Teaching should start from easy to hard, i.e., outside to inside.



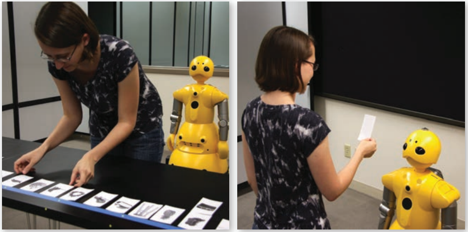
- A principle motivated by:
  - psychology
  - optimization (continuation method to avoid being trapped in bad local optima)

Outline

Teaching a robot

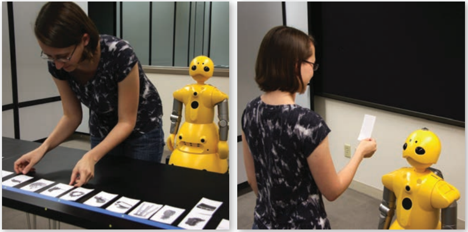
- 1D concepts to make teaching theory simple

Teaching a robot



- 1D concepts to make teaching theory simple
- robot behaviors consistent across conditions and trials (motion tracking), facilitating experimental control

Teaching a robot

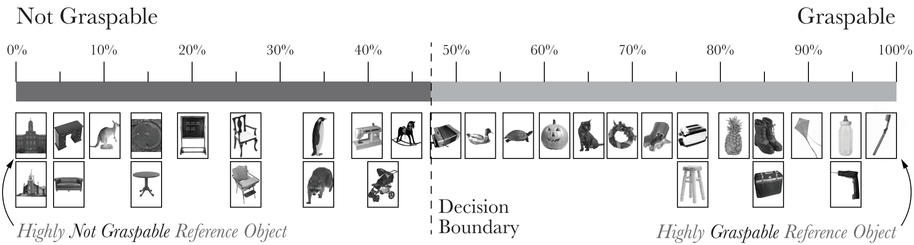


- 1D concepts to make teaching theory simple
- robot behaviors consistent across conditions and trials (motion tracking), facilitating experimental control
- Participants (human teachers): undergraduate students at Wisconsin

Outline

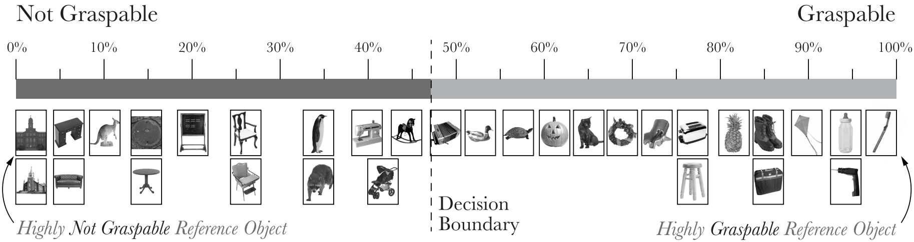
1. Teaching as a machine learning problem
2. Human teaching behaviors in a 1D task
   - "Graspability"
   - "lines"
3. Our computational rationalize of the human teaching behaviors

Materials and procedure



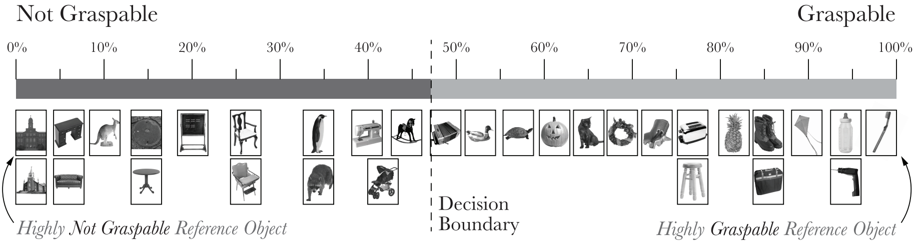
1. place cards along ruler ($x_{1:n}$)

Materials and procedure



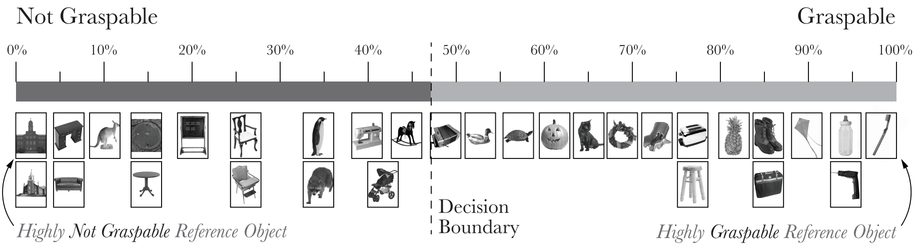
1. place cards along ruler ($x_{1:n}$)
2. label the back of each card ($y_{1:n}$)

Materials and procedure



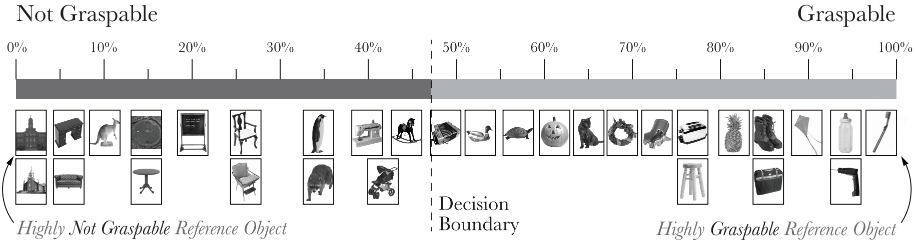
1. place cards along ruler ($x_{1:n}$)
2. label the back of each card ($y_{1:n}$)
3. leave the room, let robot inspect $x_{1:n}$

Materials and procedure



1. place cards along ruler ($x_{1:n}$)
2. label the back of each card ($y_{1:n}$)
3. leave the room, let robot inspect $x_{1:n}$
4. teach by showing one card at a time

Materials and procedure



1. place cards along ruler ($x_{1:n}$)
2. label the back of each card ($y_{1:n}$)
3. leave the room, let robot inspect $x_{1:n}$
4. teach by showing one card at a time
5. instruction: use as few cards as possible

Conditions

1. "natural": the teacher can say anything

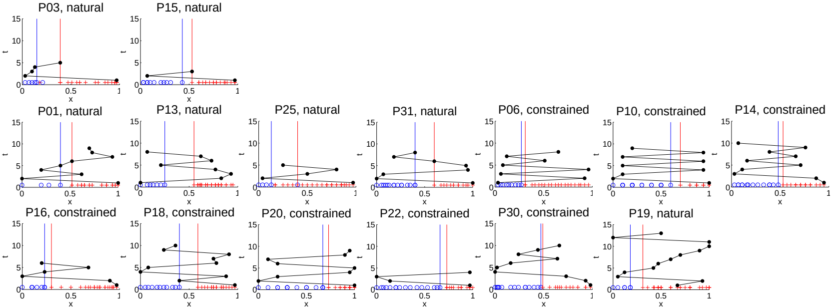
Conditions

1. "natural": the teacher can say anything
2. "constrained": the teacher can only say "graspable" or "not graspable"

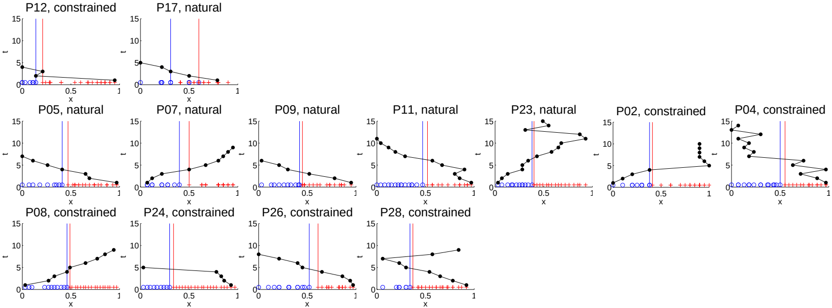
Strategy 1: "decision boundary" (0% subjects)

None

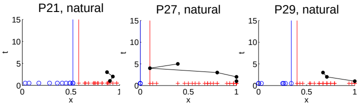
Strategy 2: “curriculum learning” (48% subjects)



Strategy 3: "linear" (42% subjects)



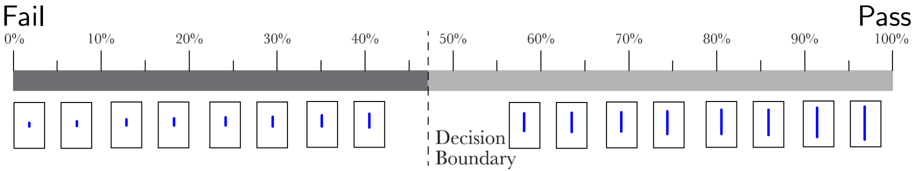
Strategy 4: "positive only" (10% subjects)



Outline

1. Teaching as a machine learning problem
2. Human teaching behaviors in a 1D task
   - "Graspability"
   - "lines"
3. Our computational rationalize of the human teaching behaviors

Materials



The master card

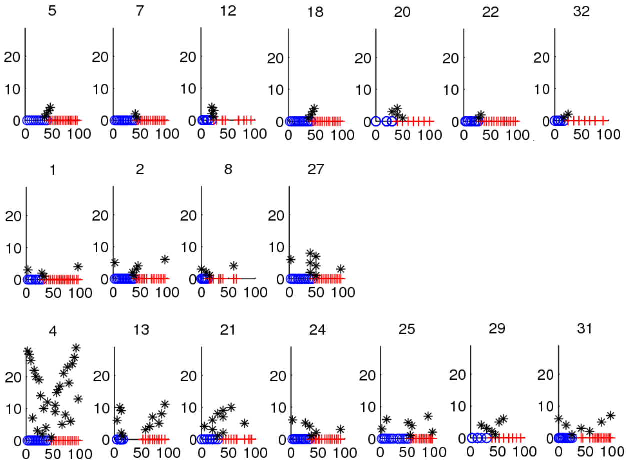
Conditions

1. with master card: the teacher can use it during sorting but not teaching (even participant IDs)

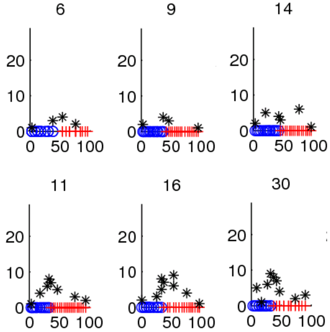
Conditions

1. with master card: the teacher can use it during sorting but not teaching (even participant IDs)
2. without master card: the teacher is shown the master card for 5 seconds at the very beginning (odd participant IDs)

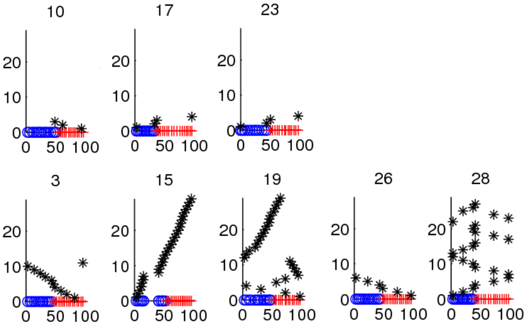
Strategy 1: "decision boundary" (56% subjects)



Strategy 2: "curriculum learning" (19% subjects)



Strategy 3: "linear" (25% subjects)



Strategy 4: "positive only" (0% subjects)

None

Comparing the two experiments

strategy	boundary	curriculum	linear	positive
"graspability" $(n = 31)$	0%	48%	42%	10%
"lines" $(n = 32)$	56%	19%	25%	0%

Outline

1. Teaching as a machine learning problem
2. Human teaching behaviors in a 1D task
   - "Graspability"
   - "lines"
3. Our computational rationalize of the human teaching behaviors

Seeking a hypothesis

- Under what assumptions is the human teaching behavior optimal?

Seeking a hypothesis

- Under what assumptions is the human teaching behavior optimal?
- Focus on decision boundary and curriculum learning

Seeking a hypothesis

- Under what assumptions is the human teaching behavior optimal?
- Focus on decision boundary and curriculum learning
- Not the linear strategy

Seeking a hypothesis

- Under what assumptions is the human teaching behavior optimal?
- Focus on decision boundary and curriculum learning
- Not the linear strategy
- Not the positive-only strategy

The hidden dimensionality

- Humans represent objects by $\mathcal{X} \subseteq \mathbb{R}^d, d \gg 1$.

The hidden dimensionality

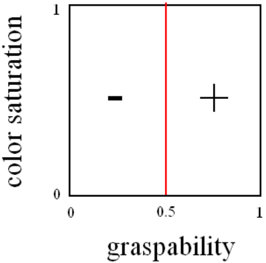
- Humans represent objects by $\mathcal{X} \subseteq \mathbb{R}^d, d \gg 1$.
- e.g., squirrel = Boolean vector (graspable, shy, store supplies for the winter, is not poisonous, has four paws, has teeth, has two ears, has two eyes, is beautiful, is brown, lives in trees, rodent, doesn't herd, doesn't sting, drinks water, eats nuts, feels soft, fluffy, gnaws on everything, has a beautiful tail, has a large tail, has a mouth, has a small head, has gnawing teeth, has pointy ears, has short paws, is afraid of people, is cute, is difficult to catch, is found in Belgium, is light, is not a pet, is not very big, is short haired, is sweet , jumps, lives in Europe, lives in the wild, short front legs, small ears, smaller than a horse, soft fur, timid animal, can't fly, climbs in trees, collects nuts, crawls up trees, eats acorns, eats plants, does not lay eggs ... $)^\top$

The hidden dimensionality

- Humans represent objects by $\mathcal{X} \subseteq \mathbb{R}^d, d \gg 1$.
- e.g., squirrel = Boolean vector (graspable, shy, store supplies for the winter, is not poisonous, has four paws, has teeth, has two ears, has two eyes, is beautiful, is brown, lives in trees, rodent, doesn't herd, doesn't sting, drinks water, eats nuts, feels soft, fluffy, gnaws on everything, has a beautiful tail, has a large tail, has a mouth, has a small head, has gnawing teeth, has pointy ears, has short paws, is afraid of people, is cute, is difficult to catch, is found in Belgium, is light, is not a pet, is not very big, is short haired, is sweet , jumps, lives in Europe, lives in the wild, short front legs, small ears, smaller than a horse, soft fur, timid animal, can't fly, climbs in trees, collects nuts, crawls up trees, eats acorns, eats plants, does not lay eggs ... $)^\top$
- "Graspability" is probably a 1D subspace in $\mathcal{X}$

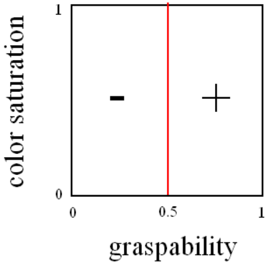
Idealized problem setting

- The first dimension determines label: $p(y_i = 1 \mid \mathbf{x}_i) = \mathbb{1}_{\{x_{i1} > \frac{1}{2}\}}$



Idealized problem setting

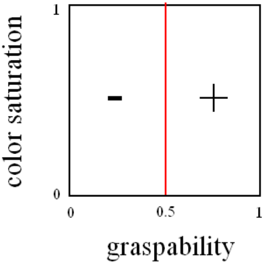
- The first dimension determines label: $p(y_i = 1 \mid \mathbf{x}_i) = \mathbb{1}_{\{x_{i1} > \frac{1}{2}\}}$



- A pool of items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{unif}[0,1]^d$ available to the teacher

Idealized problem setting

- The first dimension determines label: $p(y_i = 1 \mid \mathbf{x}_i) = \mathbb{1}_{\{x_{i1} > \frac{1}{2}\}}$



- A pool of items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \text{unif}[0,1]^d$ available to the teacher
- At time t, the teacher picks one item x_t from the pool, shows (x_t, y_t) to the learner

Sufficient conditions

The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

1. The learner has an axis-parallel version space V

Sufficient conditions

The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

1. The learner has an axis-parallel version space V
2. The learner is a Gibbs classifier

Sufficient conditions

The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

1. The learner has an axis-parallel version space V
2. The learner is a Gibbs classifier
3. The teacher is computationally limited

Sufficient conditions

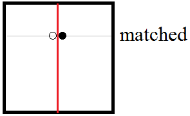
The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

1. The learner has an axis-parallel version space V
2. The learner is a Gibbs classifier
3. The teacher is computationally limited
   - only pays attention to the target dimension

Sufficient conditions

The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

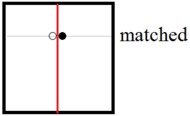
1. The learner has an axis-parallel version space V
2. The learner is a Gibbs classifier
3. The teacher is computationally limited
   - only pays attention to the target dimension
   - does not teach by matching irrelevant dimensions



Sufficient conditions

The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

1. The learner has an axis-parallel version space V
2. The learner is a Gibbs classifier
3. The teacher is computationally limited
   - only pays attention to the target dimension
   - does not teach by matching irrelevant dimensions

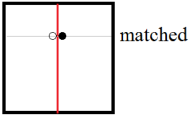


   - $\Rightarrow$ teaching items' irrelevant dimensions are random

Sufficient conditions

The following assumptions are sufficient (but not necessary) to explain human's "decision boundary" vs. "curriculum learning" behaviors:

1. The learner has an axis-parallel version space V
2. The learner is a Gibbs classifier
3. The teacher is computationally limited
   - only pays attention to the target dimension
   - does not teach by matching irrelevant dimensions



   - $\Rightarrow$ teaching items' irrelevant dimensions are random
4. The teacher sequentially minimizes the learner's risk (expected error)

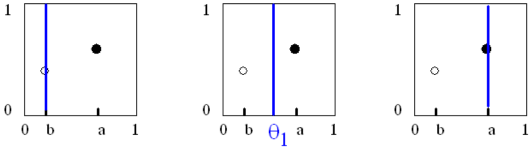
$$R = \mathbb{E}[f(x) \neq y]$$

The version space

- After two items $(x_1, y_1 = 1), (x_2, y_2 = -1)$ the version subspaces:

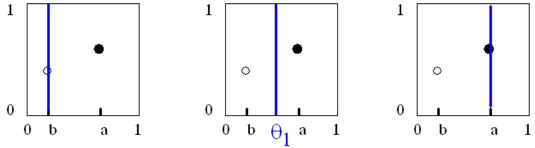
The version space

- After two items $(x_1, y_1 = 1), (x_2, y_2 = -1)$ the version subspaces:
  - $V_1 = \{1_{x_{\cdot 1} \geq \theta_1} \mid \theta_1 \in [b, a]\}$, where $b \equiv x_{11}, a \equiv x_{21}$

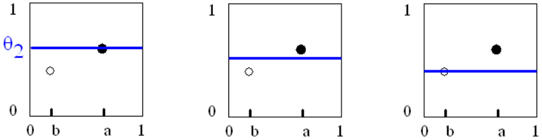


The version space

- After two items $(x_1, y_1 = 1), (x_2, y_2 = -1)$ the version subspaces:
  - $V_1 = \{1_{x_{\cdot 1} \geq \theta_1} \mid \theta_1 \in [b, a]\}$, where $b \equiv x_{11}, a \equiv x_{21}$

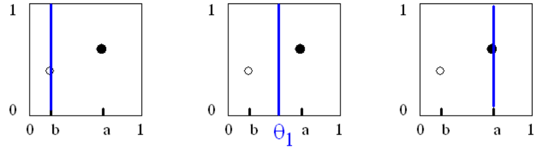


  - $V_2 = \{x_{\cdot 2} \geq \theta_2 : \theta_2 \in [\min(x_{21}, x_{22}), \max(x_{21}, x_{22})]\}$, similarly for $V_3 \ldots V_d$

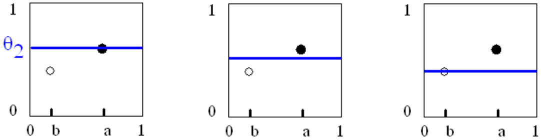


The version space

- After two items $(x_1, y_1 = 1), (x_2, y_2 = -1)$ the version subspaces:
  - $V_1 = \{1_{x_{\cdot 1} \geq \theta_1} \mid \theta_1 \in [b, a]\}$, where $b \equiv x_{11}, a \equiv x_{21}$



  - $V_2 = \{x_{\cdot 2} \geq \theta_2 : \theta_2 \in [\min(x_{21}, x_{22}), \max(x_{21}, x_{22})]\}$, similarly for $V_3 \ldots V_d$



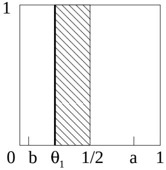
- The complete version space $V = \cup_{i=1}^{d} V_k$

The error

- The learner randomly selects one hypothesis from the version space

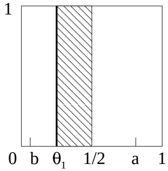
The error

- The learner randomly selects one hypothesis from the version space
- if the hypothesis is selected from dimension 1, error$=|\theta_1 - \frac{1}{2}|$

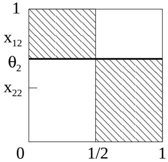


The error

- The learner randomly selects one hypothesis from the version space
- if the hypothesis is selected from dimension 1, error=$|\theta_1 - \frac{1}{2}|$



- if from dimension $2 \ldots d$, error=$\frac{1}{2}$



Risk minimization

- The learner's risk

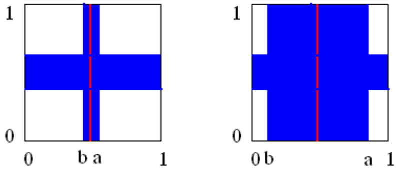
$$R = \frac{1}{|V|}\left(\int_b^a |\theta_1 - \frac{1}{2}|d\theta_1 + \sum_{k=2}^{d}\int_{\min(x_{1k},x_{2k})}^{\max(x_{1k},x_{2k})} \frac{1}{2}d\theta_k\right)$$

Risk minimization

- The learner's risk

$$R = \frac{1}{|V|}\left(\int_b^a |\theta_1 - \frac{1}{2}| d\theta_1 + \sum_{k=2}^{d}\int_{\min(x_{1k},x_{2k})}^{\max(x_{1k},x_{2k})}\frac{1}{2}d\theta_k\right)$$

- The teacher chooses a, b to minimize R. Trade off:

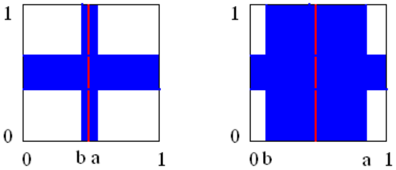


Risk minimization

- The learner's risk

$$R = \frac{1}{|V|}\left(\int_b^a |\theta_1 - \frac{1}{2}|d\theta_1 + \sum_{k=2}^{d}\int_{\min(x_{1k},x_{2k})}^{\max(x_{1k},x_{2k})}\frac{1}{2}d\theta_k\right)$$

- The teacher chooses a, b to minimize R. Trade off:
  - $a - b$ too small: learner frequently picks f in irrelevant dimensions $\Rightarrow$ large error

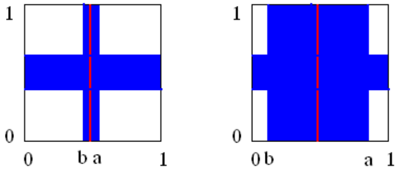


Risk minimization

- The learner's risk

$$R = \frac{1}{|V|}\left(\int_b^a |\theta_1 - \frac{1}{2}| d\theta_1 + \sum_{k=2}^{d} \int_{\min(x_{1k},x_{2k})}^{\max(x_{1k},x_{2k})} \frac{1}{2} d\theta_k\right)$$

- The teacher chooses a, b to minimize R. Trade off:
  - $a - b$ too small: learner frequently picks f in irrelevant dimensions $\Rightarrow$ large error
  - $a - b$ too large: learner picks very wrong f in the relevant dimension $\Rightarrow$ large error



Risk minimization

Theorem

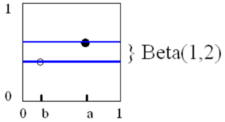
The risk R is minimized by

$$
\begin{aligned}
a^* &= \frac{\sqrt{c^2 + 2c} - c + 1}{2} \\
b^* &= 1 - a^*
\end{aligned}
$$

where $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the version subspace size in irrelevant dimensions.

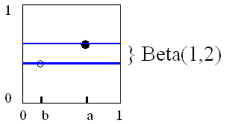
Starting teaching where? d decides

- $|x_{1k} - x_{2k}| \sim \mathrm{Beta}(1,2)$ for $k = 2, \ldots, d$ (order statistics)



Starting teaching where? d decides

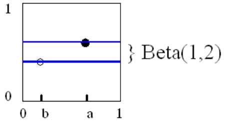
- $|x_{1k} - x_{2k}| \sim \text{Beta}(1,2)$ for $k = 2, \ldots, d$ (order statistics)



- $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the sum of $d-1$ $\text{Beta}(1,2)$ random variables.

Starting teaching where? d decides

- $|x_{1k} - x_{2k}| \sim \mathrm{Beta}(1, 2)$ for $k = 2, \ldots, d$ (order statistics)



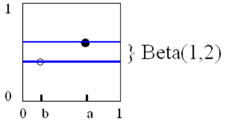
- $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the sum of $d - 1$ $\mathrm{Beta}(1, 2)$ random variables.

Corollary

When $d \to \infty$, the minimizer of R is $a^* = 1, b^* = 0$.
When $d = 1$, the minimizer of R is $a^* \to \frac{1}{2}_-, b^* \to \frac{1}{2}_+$.

Starting teaching where? d decides

- $|x_{1k} - x_{2k}| \sim \text{Beta}(1,2)$ for $k = 2, \ldots, d$ (order statistics)



- $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$ is the sum of $d-1$ $\text{Beta}(1,2)$ random variables.

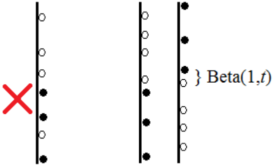
Corollary

When $d \to \infty$, the minimizer of R is $a^* = 1, b^* = 0$.
When $d = 1$, the minimizer of R is $a^* \to \frac{1}{2}_-, b^* \to \frac{1}{2}_+$.

- For example, $d = 10, a^* = 0.94$; $d = 100, a^* = 0.99$

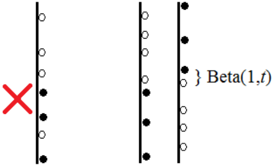
With more teaching items

- Version subspace V_k survives t teaching items if the items are linearly separable in dimension $k = 2 \ldots d$



With more teaching items

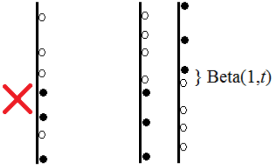
- Version subspace V_k survives t teaching items if the items are linearly separable in dimension $k = 2 \ldots d$



- This happens with probability $\frac{2}{\binom{t}{t_0}}$ where t_0 is the number of positive items

With more teaching items

- Version subspace V_k survives t teaching items if the items are linearly separable in dimension $k = 2 \ldots d$



- This happens with probability $\frac{2}{\binom{t}{t_0}}$ where t_0 is the number of positive items
- If V_k does survive, its size $\sim \mathrm{Beta}(1, t)$ (order statistics)

Teaching items should approach decision boundary

Theorem

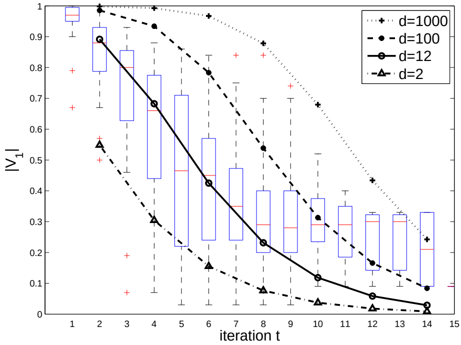
Let the teaching sequence contain t_0 *negative labels and* $t - t_0$ *positive ones. Then the version space in dim* k *has size* $|V_k| = \alpha_k \beta_k$*, where*

$$\begin{aligned} \alpha_k &\sim \text{Bernoulli}\left(2/\binom{t}{t_0}, 1 - 2/\binom{t}{t_0}\right) \\ \beta_k &\sim \text{Beta}(1, t) \end{aligned}$$

independently for $k = 2 \ldots d$*. Consequently,* $\mathbb{E}(c) = \frac{2(d-1)}{\binom{t}{t_0}(1+t)}$.

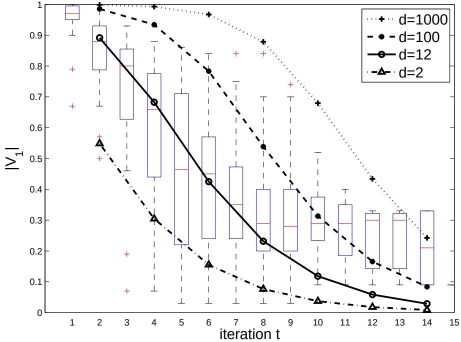
Comparing theory to behaviors

- On the "graspability" task with assumed d's:



Comparing theory to behaviors

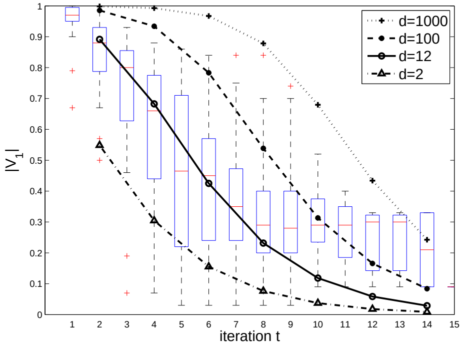
- On the "graspability" task with assumed d's:



- On the "lines" task, theory predicts $|V_1|$ at minimum in iteration 2

Comparing theory to behaviors

- On the "graspability" task with assumed d's:



- On the "lines" task, theory predicts $|V_1|$ at minimum in iteration 2
- Curriculum learning and teaching dimension are both correct: different cases of the same theory

Conclusion

- Behavioral studies of human teaching

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d
- Applications:

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d
- Applications:
  - robots that learn from grandma (and CS grads, too)

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d
- Applications:
  - robots that learn from grandma (and CS grads, too)
  - more effective educational strategies for kids

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d
- Applications:
  - robots that learn from grandma (and CS grads, too)
  - more effective educational strategies for kids
- Acknowledgments

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d
- Applications:
  - robots that learn from grandma (and CS grads, too)
  - more effective educational strategies for kids
- Acknowledgments
  - Collaborators: Kwangsung Jun, Faisal Khan, Bilge Mutlu, Burr Settles

Conclusion

- Behavioral studies of human teaching
  - "graspability": curriculum learning strategy
  - "lines": decision boundary strategy
- Potential computational teaching theory
  - sequential risk minimization
  - d controls behavior
  - justifies curriculum learning for large d
- Applications:
  - robots that learn from grandma (and CS grads, too)
  - more effective educational strategies for kids
- Acknowledgments

  - Collaborators: Kwangsung Jun, Faisal Khan, Bilge Mutlu, Burr Settles
  - NSF CAREER IIS-0953219, AFOSR FA9550-09-1-0313, The Wisconsin Alumni Research Foundation

Reference

Faisal Khan, Xiaojin Zhu, and Bilge Mutlu.
How do humans teach: On curriculum learning and teaching dimension.
In *Advances in Neural Information Processing Systems (NIPS) 25*. 2011.

Backup slides

Learning from iid data

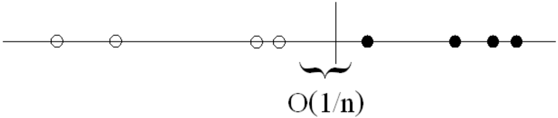
- The most common machine learning assumption

Learning from iid data

- The most common machine learning assumption
- The learner passively receives a training sample $(x_1, y_1) \ldots (x_n, y_n) \stackrel{iid}{\sim} p$

Learning from iid data

- The most common machine learning assumption
- The learner passively receives a training sample $(x_1, y_1) \ldots (x_n, y_n) \overset{iid}{\sim} p$
- Risk decreases as $O(\frac{1}{n})$



Active learning

- The learner picks x_t

Active learning

- The learner picks x_t
- The teacher answers y_t

Active learning

- The learner picks x_t
- The teacher answers y_t
- The teacher does not pick x_t!

Active learning

- The learner picks x_t
- The teacher answers y_t
- The teacher does not pick x_t!
- Risk decreases as $\frac{1}{2^n}$ (noiseless 1D case, equivalent to binary search)