How Do Humans Teach: On Curriculum Learning and Teaching Dimension

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

NIPS 2011
The teaching dimension [Goldman and Kearns 1995]

Optimal teaching should start around the decision boundary.
The teaching dimension [Goldman and Kearns 1995]

- items $\mathcal{X}$

- Teaching in humans and machines

(NIPS 2011)
The teaching dimension [Goldman and Kearns 1995]

- items $\mathcal{X}$
- $H$ threshold functions

Optimal teaching should start around the decision boundary.
The teaching dimension [Goldman and Kearns 1995]

- items $\mathcal{X}$
- $H$ threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with $h$ only
The teaching dimension [Goldman and Kearns 1995]

- items $\mathcal{X}$
- $H$ threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with $h$ only
- $TD(h)$: size of the smallest teaching set of $h$, 1 or 2

$T_D(h^*)$, $TD(H)$: size of the smallest teaching set of $h^*$, 1 or 2
The teaching dimension [Goldman and Kearns 1995]

- items $\mathcal{X}$
- $H$ threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with $h$ only
- $TD(h)$: size of the smallest teaching set of $h$, 1 or 2
- $TD(H)$: $TD(h^*)$ for the hardest $h^* \in H$, 2
The teaching dimension [Goldman and Kearns 1995]

- items $\mathcal{X}$
- $H$ threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with $h$ only
- $TD(h)$: size of the smallest teaching set of $h$, 1 or 2
- $TD(H)$: $TD(h^*)$ for the hardest $h^* \in H$, 2

Optimal teaching should start around the decision boundary.
Teaching should start from easy to hard, i.e., outside to inside.
You teach robot ...
... graspability

Two conditions:
1. Teacher can say anything
2. Teacher can only say “graspable” or “not graspable”
... graspability

Two conditions:

1. teacher can say anything
2. teacher can only say “grasppable” or “not grasppable”
Observed human teaching strategy 1

P31, natural

<table>
<thead>
<tr>
<th>V</th>
<th>1</th>
</tr>
</thead>
</table>

(NIPS 2011)
Observed human teaching strategy 2

P08, constrained
Observed human teaching strategy

P21, natural

(NIPS 2011)
Teaching in humans and machines
All results

“curriculum learning”

“linear”

“positive example only”
Humans represent objects by many dimensions!

- squirrel = (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 ...)
Idealized assumptions

- available teaching items $x_1, \ldots, x_n \sim \text{unif}[0, 1]^d$
Idealized assumptions

- available teaching items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \text{unif}[0, 1]^d$
- first dim determines label $p(y_i = 1 \mid \mathbf{x}_i) = 1\{x_{i1} > \frac{1}{2}\}$

(NIPS 2011) Teaching in humans and machines
Idealized assumptions

- available teaching items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \text{unif}[0, 1]^d$
- first dim determines label $p(y_i = 1 \mid \mathbf{x}_i) = \mathbb{1}_{x_{i1} > \frac{1}{2}}$
- learner’s version space $V$: axis-parallel decision boundaries
Idealized assumptions

- available teaching items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \text{unif}[0, 1]^d$
- first dim determines label $p(y_i = 1 \mid \mathbf{x}_i) = 1\{x_{i1} > \frac{1}{2}\}$
- learner’s version space $\mathcal{V}$: axis-parallel decision boundaries
  - after two teaching items
    - $\mathbf{x}_1 = (x_{11}, \ldots, x_{1d}), y_1 = 1$
    - $\mathbf{x}_2 = (x_{21}, \ldots, x_{2d}), y_2 = 0$
One more assumption

learner is a Gibbs classifier (uniformly select a hypothesis from $V$)

$a \equiv x_{11}, b \equiv x_{21}$

if hypothesis selected from dim 1, error $= |\theta_1 - \frac{1}{2}|$

if from dim 2, error $= \frac{1}{2}$
Risk minimization leads to teaching extremes

- 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 leads to teaching extremes

- learner’s risk

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

- teacher chooses \( a, b \) to minimize \( R \) (trade off)
Risk minimization leads to teaching extremes

- learner’s risk

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

- teacher chooses \( a, b \) to minimize \( R \) (trade off)

**Theorem**

*The risk \( R \) is minimized by \( a^* = \frac{\sqrt{c^2 + 2c - c + 1}}{2} \) and \( b = 1 - a^* \), where \( c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}| \).*
Risk minimization leads to teaching extremes

- learner's risk

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

- teacher chooses \(a, b\) to minimize \(R\) (trade off)

**Theorem**

The risk \(R\) is minimized by \(a^* = \frac{\sqrt{c^2 + 2c - c + 1}}{2}\) and \(b = 1 - a^*\), where \(c \equiv \sum_{k=2}^d |x_{1k} - x_{2k}|\).

\(c\) is the sum of \(d - 1\) Beta\((1, 2)\) random variables.
Risk minimization leads to teaching extremes

- learner’s risk

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

- teacher chooses \( a, b \) to minimize \( R \) (trade off)

**Theorem**

The risk \( R \) is minimized by \( a^* = \frac{\sqrt{c^2 + 2c - c + 1}}{2} \) and \( b = 1 - a^* \), where

\[ c = \sum_{k=2}^{d} |x_{1k} - x_{2k}|. \]

\( c \) is the sum of \( d - 1 \) Beta(1, 2) random variables.

**Corollary**

When \( d \to \infty \), the minimizer of \( R \) is \( a^* = 1, b^* = 0 \).
Risk minimization leads to teaching extremes

- learner’s risk

\[ R = \frac{1}{|V|} \left( \int_a^b |\theta_1 - \frac{1}{2}| d\theta_1 + \sum_{k=2}^d \int_{\min(x_{1k}, x_{2k})}^{\max(x_{1k}, x_{2k})} \frac{1}{2} d\theta_k \right) \]

- teacher chooses \( a, b \) to minimize \( R \) (trade off)

**Theorem**

The risk \( R \) is minimized by \( a^* = \sqrt{\frac{c^2 + 2c - c + 1}{2}} \) and \( b = 1 - a^* \), where \( c \equiv \sum_{k=2}^d |x_{1k} - x_{2k}| \).

\( c \) is the sum of \( d - 1 \) Beta(1, 2) random variables.

**Corollary**

When \( d \to \infty \), the minimizer of \( R \) is \( a^* = 1, b^* = 0 \).

In practice, \( d = 10, a^* = 0.94; d = 100, a^* = 0.99 \)
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

\[
\alpha_k \sim \text{Bernoulli}\left(\frac{2}{t/0}, 1 - \frac{2}{t/0}\right)
\]

\[
\beta_k \sim \text{Beta}(1, t)
\]

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

\[
\alpha_k \sim \text{Bernoulli} \left( \frac{2}{t} \left( \frac{t}{t_0} \right), 1 - \frac{2}{t} \left( \frac{t}{t_0} \right) \right)
\]

\[
\beta_k \sim \text{Beta}(1, t)
\]

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

(NIPS 2011)
Conclusion

- People teach like curriculum learning
- Can extend teaching theory to explain it
Conclusion

- People teach like curriculum learning
- Can extend teaching theory to explain it

Acknowledgments
- Faisal Khan, Bilge Mutlu
- NSF CAREER Award IIS-0953219
- AFOSR FA9550-09-1-0313
- The Wisconsin Alumni Research Foundation