Three Assertions about Interactive Machine Learning

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

jerryzhu@cs.wisc.edu

2013

Assertion 1: Humans can be modeled with statistical learning theory

- Unifying math behind cognitive science and machine learning

Example 1a: Human Rademacher Complexity

(grenade, A), (meadow, A), (skull, B), (conflict, B), (queen, B)

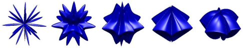
Example 1a: Human Rademacher Complexity

(grenade, A), (meadow, A), (skull, B), (conflict, B), (queen, B)

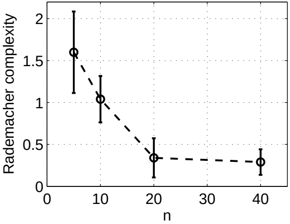
- "learning random labels" $(x_1, \sigma_1) \ldots (x_n, \sigma_n)$
- Rademacher complexity (similar to VC dimension)

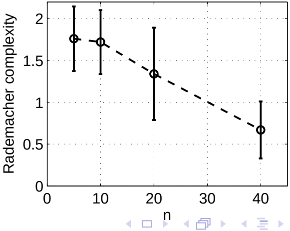
$$Rad_n(F) \approx \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i \hat{f}(x_i) \right|$$

  … of our mind!
- Larger Rademacher complexity → worse generalization error bound (overfitting) [ZRG NIPS09]



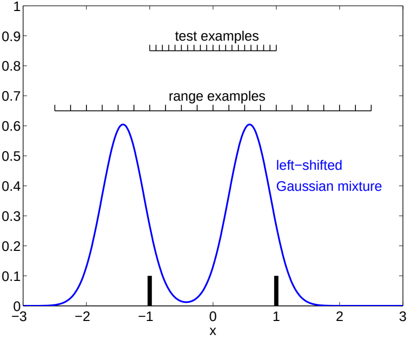
rape killer funeral ⋯ fun laughter joy

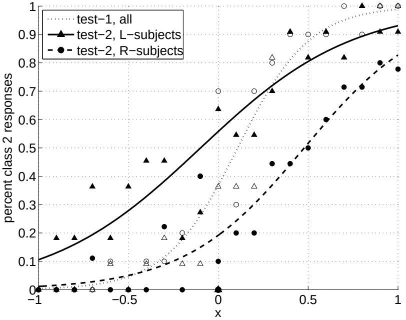




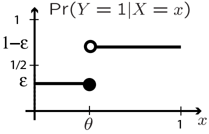
Example 1b: Human Semi-Supervised Learning

- Humans learn supervised first, then
- . . . decision boundary shifts to distribution trough in test data
- Can be explained by a variety of semi-supervised machine learning models [GRZ ToCS13]





Example 1c: Human Active Learning



[CKNQRZ NIPS08]

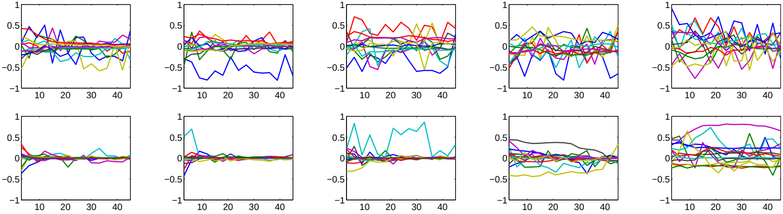
Passive learning $\inf_{\hat{\theta}_n} \sup_{\theta\in[0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \geq \frac{1}{4}\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{2\epsilon} \frac{1}{n+1}$

Active learning $\sup_{\theta\in[0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \leq 2\left(\sqrt{\frac{1}{2} + \sqrt{\epsilon(1-\epsilon)}}\right)^n$

noise $\epsilon = 0$ $\epsilon = 0.05$ $\epsilon = 0.1$ $\epsilon = 0.2$ $\epsilon = 0.4$

Human Passive

Human Active

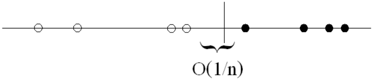


Assertion 2: There is a theoretically optimal way to teach

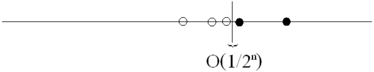
Human teaches machine (interactive ML)
Machine teaches human (education)

Example 2: 1D threshold function

- Passive learning $(x_i, y_i) \overset{iid}{\sim} p$, risk $\approx O(\frac{1}{n})$

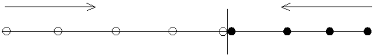


- Active learning risk $\approx \frac{1}{2^n}$



- Minimum teaching: $n = 2$. (teaching dimension)

- Alternatively: easy to hard (curriculum learning, fading, parentese)



A formula for optimal teaching

1. World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
2. Learner: makes prediction $f(x \mid \text{data})$
3. Teacher:
    - clairvoyant, knows everything above
    - can only teach by examples (x, y)
    - goal: choose the least-effort teaching set $D = (x, y)_{1:n}$ to minimize the learner's future loss (risk):

$$\min_{D} \quad \mathbb{E}_{\theta^*}[\ell(f(x \mid D), y)] + \text{effort}(D)$$

    - if the future loss approaches Bayes risk, D is a teaching set and n is the (generalized) teaching dimension

[KZM NIPS11, Z arXiv13]

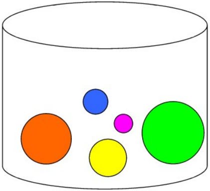
Assertion 3: Even when human teachers are not optimal, they are not iid

... and machine learners should take advantage of that non-iidness.

Example 3: Feature Volunteering (Interactive ML)

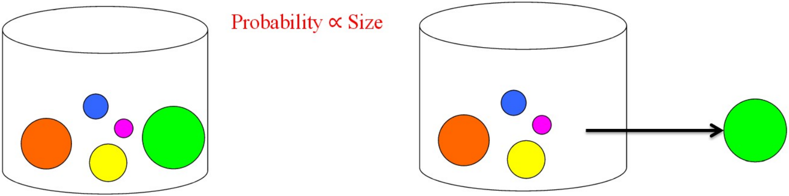
[JZSR ICML13]

Example 3: Sampling with Reduced Replacement

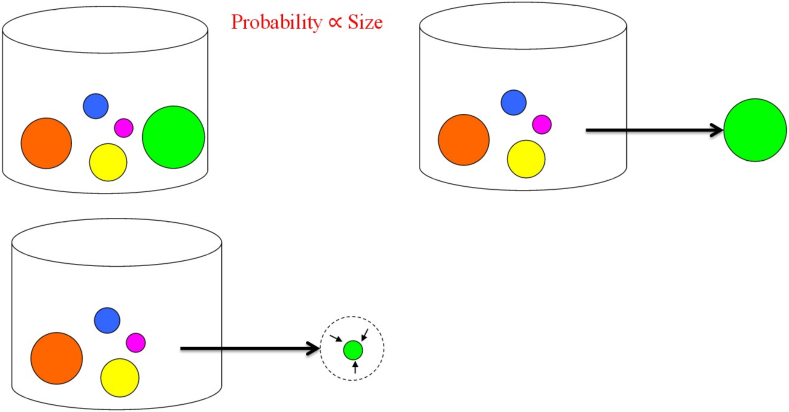


Probability $\propto$ Size

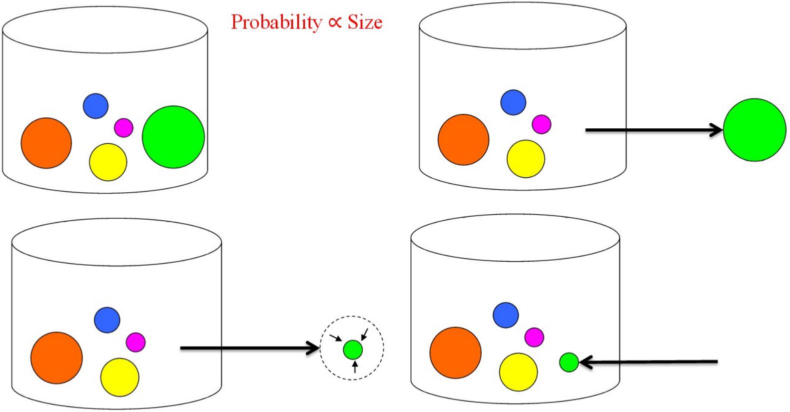
Example 3: Sampling with Reduced Replacement



Example 3: Sampling with Reduced Replacement



Example 3: Sampling with Reduced Replacement



Example 3: Sampling with Reduced Replacement

Domain	Reference Distributions		
	SWIRL	Equal	Schapire
sports	**0.865**	0.847	0.795
movies	**0.733**	**0.733**	**0.725**
webkb	**0.463**	0.444	0.429

References

R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu.
Human active learning.
In Advances in Neural Information Processing Systems (NIPS) 22. 2008.

B. R. Gibson, T. T. Rogers, and X. Zhu.
Human semi-supervised learning.
Topics in Cognitive Science, 5(1):132–172, 2013.

K.-S. Jun, X. Zhu, B. Settles, and T. Rogers.
Learning from human-generated lists.
In The 30th International Conference on Machine Learning (ICML), 2013.

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

X. Zhu, T. T. Rogers, and B. Gibson.
Human Rademacher complexity.
In Advances in Neural Information Processing Systems (NIPS) 23. 2009.

Three Assertions

1. Humans can be modeled with statistical learning theory.
2. There is a theoretically optimal way to teach.
3. Even when human teachers are not optimal, they are not iid.

Capacity

VC-dimension

- F: a family of binary classifiers
- VC-dimension $VC(F)$: size of the largest set that F can shatter
- With probability at least $1-\delta$,

$$\sup_{f\in F} R(f) - R_n(f) \leq 2\sqrt{2\frac{VC(F)\log n + VC(F)\log\frac{2e}{VC(F)} + \log\frac{2}{\delta}}{n}}.$$

- $R(f)$: error of f in the future
- $R_n(f)$: error of f on a training set of size n

Capacity

Rademacher complexity

- $\sigma_1, \dots, \sigma_n : P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$
- Rademacher complexity

$$Rad_n(F) = \mathbb{E}_{\sigma,x}\left(\sup_{f \in F}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right).$$

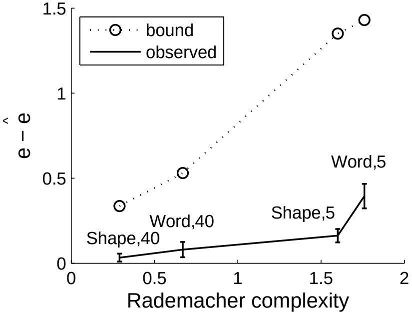
- With probability at least $1-\delta$,

$$\sup_{f \in F}|R_n(f) - R(f)| \leq 2Rad_n(F) + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Machine learning $\rightarrow$ human learning

- f: you categorize x by $f(x)$
- F: all the classifiers in your mind
- $R_n(f)$: how did you do in class
- $R(f)$: how well can you do outside class
- Capacity: can we measure it in humans?
  - $VC(F)$: too brittle (find one dataset of size n) and combinatorial (verify shattering)
  - Others may behave better, e.g., $Rad_n(F)$

Overfitting indicator



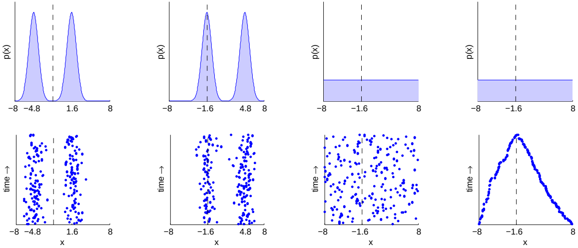
- e test set error, $\hat{e}$ training set error
- generalization error bound holds
- actual overfitting tracks bound (nice but <u>not</u> predicted by theory)

The study of capacity may

- constrain cognitive models
- understand groups differ in age, health, education, etc.

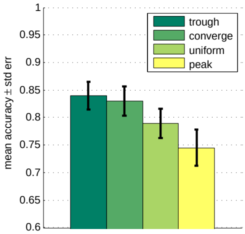
Human semi-supervised learning, the other way around

Human unsupervised learning first

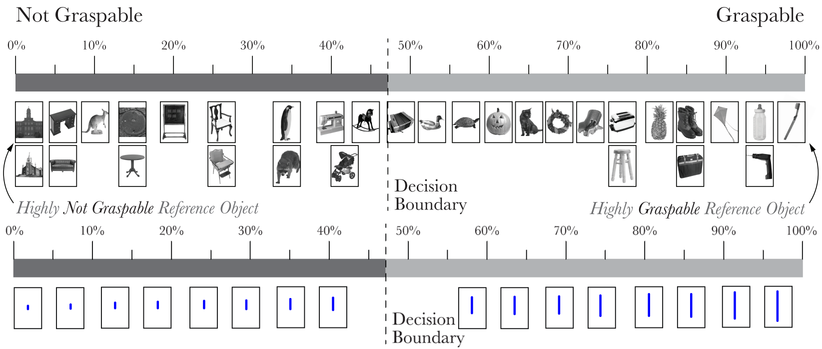


trough peak uniform converge

. . . influences subsequent (identical) supervised learning task



Human teacher behaviors



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