Capacity, Learning, Teaching

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

jerryzhu@cs.wisc.edu

2013

Machine learning $\leftrightarrow$ human learning

- Learning capacity and generalization bounds
- Beyond supervised learning: semi-supervised, active
- Beyond learning: teaching

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 $1-\delta$,

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

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

Capacity

Rademacher complexity

- $\sigma_1, \ldots, \sigma_n : P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$
- 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 $1-\delta$,

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

Machine learning $\to$ human learning

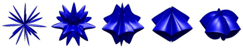
- f: you categorize x by $f(x)$
- F: all the classifiers in your mind
- $R_n(f)$: how did you do in class
- $R(f)$: how well can you do outside class
- Capacity: can we measure it in humans?
  - $VC(F)$: too brittle (find <u>one</u> dataset of size n) and combinatorial (verify shattering)
  - Others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

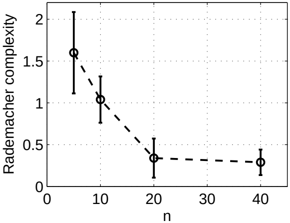
Measuring human Rademacher complexity

"learning random labels" $(x_1, \sigma_1) \ldots (x_n, \sigma_n)$, e.g., (grenade, B), (skull, A), (conflict, A), (meadow, B), (queen, B)

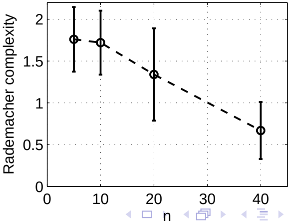
$$Rad_n(F) \approx \frac{1}{m} \sum_{j=1}^{m} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i^{(j)} \hat{f}^{(j)}(x_i^{(j)}) \right|$$

- $\hat{f}$ mnemonics: "a queen was sitting in a meadow and then a grenade was thrown (B = before), then this started a conflict ending in bodies & skulls (A = after)."
- $\hat{f}$ wrong rules: (daylight, A), (hospital, B), (termite, B), (envy, B), (scream, B), "anything related to omitting[sic] light"

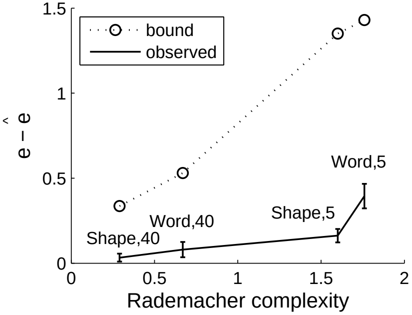




rape killer funeral $\cdots$ fun laughter joy



Overfitting indicator



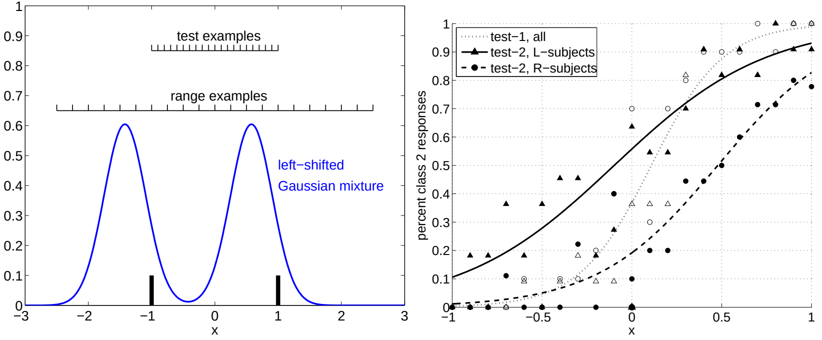
- e test set error, $\hat{e}$ training set error
- generalization error bound holds
- actual overfitting tracks bound (nice but <u>not</u> predicted by theory)

The study of capacity may

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

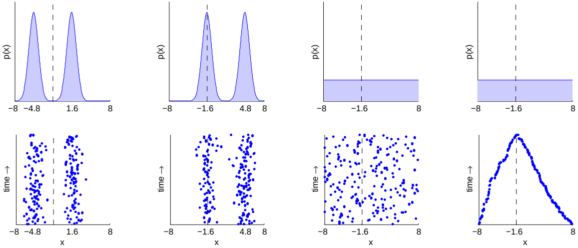
Human semi-supervised learning

- 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



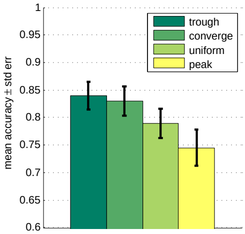
Human semi-supervised learning, the other way around

Human unsupervised learning first

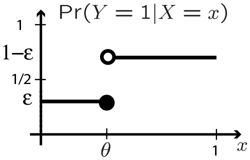


trough peak uniform converge

. . . influences subsequent (identical) supervised learning task



Active learning



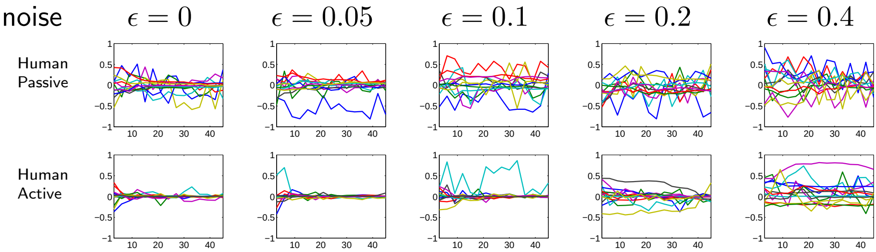
Passive learning (slow)

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

Active learning (fast)

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

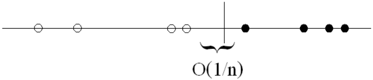
Active learning → humans



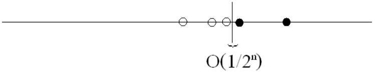
Machine teaching

Example: a threshold classifier in 1D

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

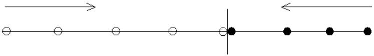


- active learning risk $\approx \frac{1}{2^n}$

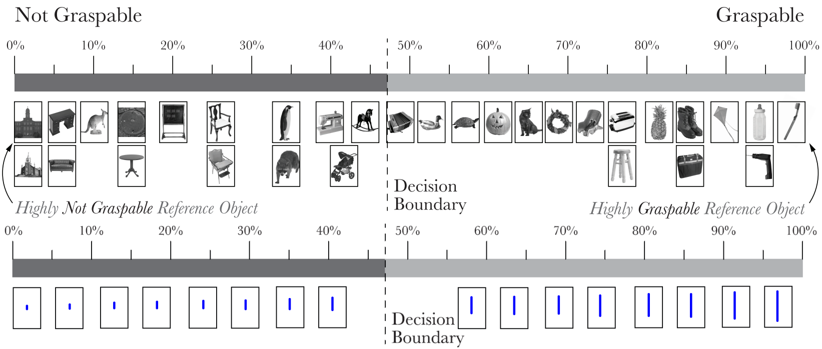


- taught: $n = 2$. Teaching dimension

- curriculum learning



Human teacher behaviors



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

A framework for teaching a Bayesian learner

1. World: $p(x, y \mid \theta^*)$, loss function $\ell(f(x), y)$
2. Learner: Bayesian.
   - prior over Θ ($\theta^* \in \Theta$), likelihood $p(x, y \mid \theta)$
   - maintains posterior $p(\theta \mid \mathrm{data})$ by Bayesian update
   - makes prediction $f(x \mid \mathrm{data})$ using the posterior
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):

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

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

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.

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. Rogers, R. Qian, and C. Kalish.
Humans perform semi-supervised classification too.
In Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07), 2007.

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