Is Machine Learning the Wrong Name?

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

October 2010

Iris Learns “Cow”

Iris Learns "Cow"

Think machine learning

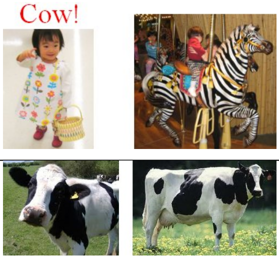
- supervised learning
- given stimulus feedback pairs $(x_1, y_1), \ldots (x_n, y_n) \sim p(x, y)$
- learn classifier $f : \mathcal{X} \mapsto \mathcal{Y}$

Think More Machine Learning

Think More Machine Learning

overfitting

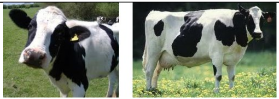
Think More Machine Learning



overfitting

Think More Machine Learning

overfitting

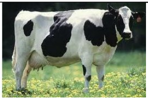


manifold learning

Think More Machine Learning

Cow!

overfitting

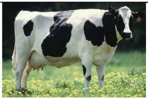


manifold learning

What's that?

Think More Machine Learning

overfitting



manifold learning

active learning

Outline

1. Overfitting in Humans
2. Human Manifold Learning
3. Active Learning in Humans

Bounding Overfitting in Humans [NIPS 2009]

- binary classifier $f : \mathcal{X} \mapsto \pm 1$

Bounding Overfitting in Humans [NIPS 2009]

- binary classifier $f : \mathcal{X} \mapsto \pm 1$
- training error $\hat{e}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i \neq f(x_i))$

Bounding Overfitting in Humans [NIPS 2009]

- binary classifier $f : \mathcal{X} \mapsto \pm 1$
- training error $\hat{e}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i \neq f(x_i))$
- generalization error $e(f) = \mathbb{E}_{(x,y) \overset{iid}{\sim} P_{XY}}\left[(y \neq f(x))\right]$
  - unknowable as the World P_{XY} is unknown

Bounding Overfitting in Humans [NIPS 2009]

- binary classifier $f : \mathcal{X} \mapsto \pm 1$
- training error $\hat{e}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i \neq f(x_i))$
- generalization error $e(f) = \mathbb{E}_{(x,y) \overset{iid}{\sim} P_{XY}} [(y \neq f(x))]$
  - unknowable as the World P_{XY} is unknown
- overfitting $e(f) - \hat{e}(f)$
  - usually estimated using a test set
  - the nature of overfitting unclear

Generalization Error Bounds in Machine Learning

Review:

- Though P_{XY} is unknown, computational learning theory can bound overfitting
- Key idea: f comes from a function family $\mathcal{F}$ with *limited capacity* R

Generalization Error Bounds in Machine Learning

Review:

- Though P_{XY} is unknown, computational learning theory can bound overfitting
- Key idea: f comes from a function family $\mathcal{F}$ with *limited capacity* R

Theorem. Let $\mathcal{F} : \mathcal{X} \mapsto \pm 1$. Let $\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} P_{XY}$ be a training sample of size n. $\forall \delta > 0$, with probability at least $1 - \delta$, every function $f \in F$ satisfies

$$e(f) - \hat{e}(f) \leq \frac{R(\mathcal{F}, \mathcal{X}, P_X, n)}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

Rademacher Complexity

Review:

$$R(\mathcal{F}, \mathcal{X}, P_X, n) = \mathbb{E}_{\boldsymbol{x\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right]$$

where the expectation is over $\mathbf{x} = x_1, \ldots, x_n \stackrel{iid}{\sim} P_X$, and $\sigma = \sigma_1, \ldots, \sigma_n \stackrel{iid}{\sim} \mathrm{Bernoulli}(\frac{1}{2}, \frac{1}{2})$ with values ± 1.

Rademacher Complexity

Review:

$$R(\mathcal{F}, \mathcal{X}, P_X, n) = \mathbb{E}_{\boldsymbol{x\sigma}} \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| \right]$$

where the expectation is over $\mathbf{x} = x_1, \ldots, x_n \overset{iid}{\sim} P_X$, and $\sigma = \sigma_1, \ldots, \sigma_n \overset{iid}{\sim} \mathrm{Bernoulli}(\frac{1}{2}, \frac{1}{2})$ with values ± 1.

- intuition: if for any random data $(x_1, \sigma_1) \ldots (x_n, \sigma_n)$, $\exists f \in \mathcal{F}$ which correlates the random labels, then $\mathcal{F}$ has high capacity
- R can be estimated from samples of $\mathbf{x}, \sigma$

Estimating Human Rademacher Complexity

$\mathcal{F}$ is all the classifiers in our mind!

Estimating Human Rademacher Complexity

$\mathcal{F}$ is all the classifiers in our mind!

1. Participant shown paper with $\{(x_i, \sigma_i\}_{i=1}^n$, asked to learn rule

Estimating Human Rademacher Complexity

$\mathcal{F}$ is all the classifiers in our mind!

1. Participant shown paper with $\{(x_i, \sigma_i\}_{i=1}^n$, asked to learn rule
2. filler task

Estimating Human Rademacher Complexity

$\mathcal{F}$ is all the classifiers in our mind!

1. Participant shown paper with $\{(x_i, \sigma_i\}_{i=1}^n$, asked to learn rule
2. filler task
3. Shown $\{x_i\}_{i=1}^n$ again, predict labels $\hat{f}(x_j)$. Order scrambled, not told the items are the same.

Estimating Human Rademacher Complexity

$\mathcal{F}$ is all the classifiers in our mind!

1. Participant shown paper with $\{(x_i, \sigma_i\}_{i=1}^n$, asked to learn rule
2. filler task
3. Shown $\{x_i\}_{i=1}^n$ again, predict labels $\hat{f}(x_j)$. Order scrambled, not told the items are the same.

Key approximation:

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

Estimating Human Rademacher Complexity

$\mathcal{F}$ is all the classifiers in our mind!

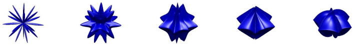
1. Participant shown paper with $\{(x_i, \sigma_i\}_{i=1}^n$, asked to learn rule
2. filler task
3. Shown $\{x_i\}_{i=1}^n$ again, predict labels $\hat{f}(x_j)$. Order scrambled, not told the items are the same.

Key approximation:

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

Average over m participants $R \approx \frac{1}{m} \sum_{j=1}^{m} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i^{(j)} \hat{f}^{(j)}(x_i^{(j)}) \right|$

Estimated Human Rademacher Complexity

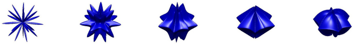


the Shape domain

rape killer funeral $\cdots$ fun laughter joy

the Word domain

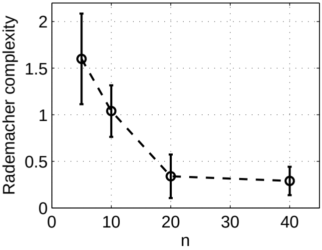
Estimated Human Rademacher Complexity



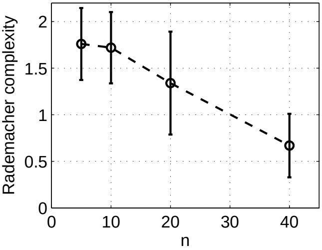
the Shape domain

rape killer funeral $\cdots$ fun laughter joy

the Word domain



$R(\mathcal{F}, \text{Shape}, \text{uniform}, n)$



$R(\mathcal{F}, \text{Word}, \text{uniform}, n)$

Human Generalization Error Bounds

$$e(f) \leq \hat{e}(f) + \frac{R(\mathcal{F}, \mathcal{X}, P_X, n)}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

condition	subject	$\hat{e}$	RHS	e
WordEmotion	101	0.00	1.43	0.58
n=5	102	0.00	1.43	0.46
	103	0.00	1.43	0.04
	104	0.00	1.43	0.03
	105	0.00	1.43	0.31
WordEmotion	106	0.70	1.23	0.65
n=40	107	0.00	0.53	0.04
	108	0.00	0.53	0.00
	109	0.62	1.15	0.53
	110	0.00	0.53	0.05

Human Overfitting Behaviors

Wrong rules learned by humans:

Human Overfitting Behaviors

Wrong rules learned by humans:

- whether the shape faces downward

Human Overfitting Behaviors

Wrong rules learned by humans:

- whether the shape faces downward
- whether the word contains the letter T

Human Overfitting Behaviors

Wrong rules learned by humans:

- whether the shape faces downward
- whether the word contains the letter T
- things you can go inside

Human Overfitting Behaviors

Wrong rules learned by humans:

- whether the shape faces downward
- whether the word contains the letter T
- things you can go inside
- odd or even number of syllables

Human Overfitting Behaviors

Wrong rules learned by humans:

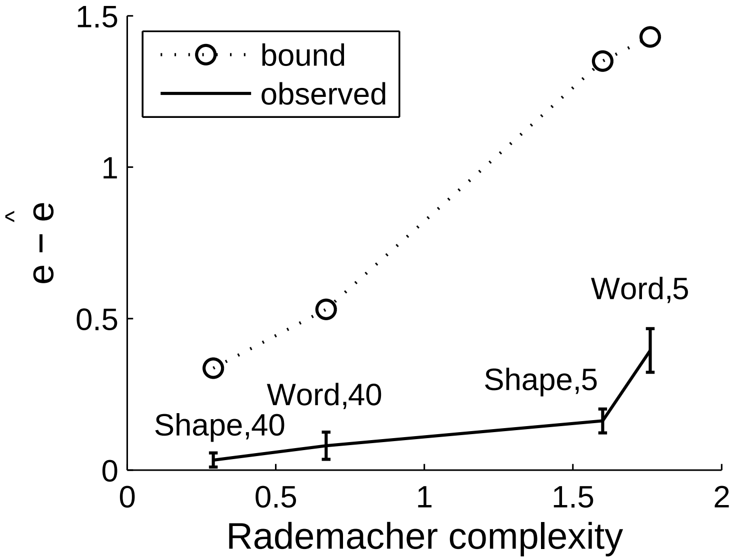
- whether the shape faces downward
- whether the word contains the letter T
- things you can go inside
- odd or even number of syllables
- training items (grenade, B), (skull, A), (conflict, A), (meadow, B), (queen, B) $\Rightarrow$ story: a queen was sitting in a meadow and then a grenade was thrown (B = before), then this started a conflict ending in bodies & skulls (A = after).

Human Overfitting Behaviors

Wrong rules learned by humans:

- whether the shape faces downward
- whether the word contains the letter T
- things you can go inside
- odd or even number of syllables
- training items (grenade, B), (skull, A), (conflict, A), (meadow, B), (queen, B) ⇒ story: 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).
- training items (daylight, A), (hospital, B), (termite, B), (envy, B), (scream, B) ⇒ class A is anything related to omitting[sic] light

Rademacher Complexity Predicts Overfitting



Mini Summary

- overfitting = true error - training error
- computational learning theory bounds overfitting
- Rademacher complexity: "capacity" of learner

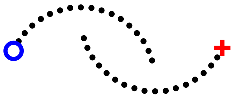
Outline

1. Overfitting in Humans
2. Human Manifold Learning
3. Active Learning in Humans

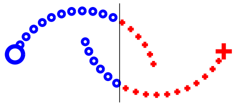
Human Manifold Learning [NIPS 2010]

Classification with

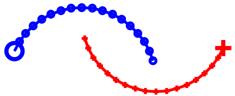
- labeled items $x_1, \ldots, x_l \in \mathbb{R}^d$ and labels $y_1, \ldots, y_l \in \{-1, 1\}$
- unlabeled items $x_{l+1}, \ldots, x_{l+u} \in \mathbb{R}^d$ without labels



(a) the data



(b) supervised learning



(c) manifold learning

An electric network interpretation

Review:

- Edges (constructed by ϵ-NN) are resistors with conductance w_{ij}

An electric network interpretation

Review:

- Edges (constructed by ϵ-NN) are resistors with conductance w_{ij}
- 1 volt battery connects to labeled points $y = 0, 1$

An electric network interpretation

Review:

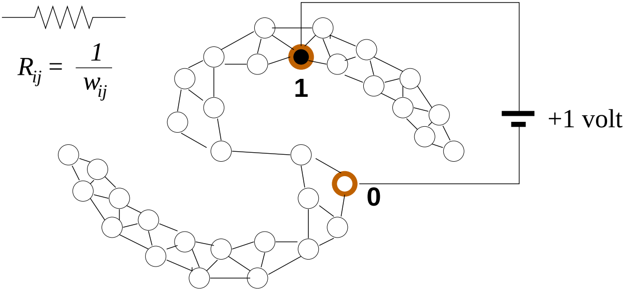
- Edges (constructed by ϵ-NN) are resistors with conductance w_{ij}
- 1 volt battery connects to labeled points $y = 0, 1$
- The voltage at the nodes is the harmonic function $f_u = -\Delta_{uu}{}^{-1}\Delta_{ul}Y_l$

An electric network interpretation

Review:

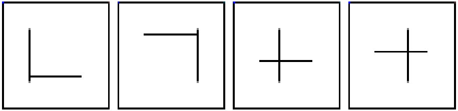
- Edges (constructed by ϵ-NN) are resistors with conductance w_{ij}
- 1 volt battery connects to labeled points $y = 0, 1$
- The voltage at the nodes is the harmonic function $f_u = -\Delta_{uu}{}^{-1}\Delta_{ul}Y_l$

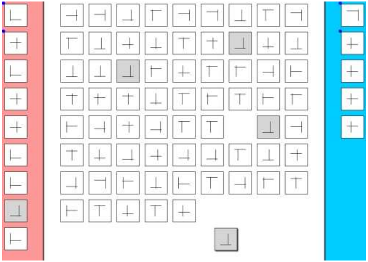
Implied similarity: similar voltage if many paths exist



Human Behavioral Experiments

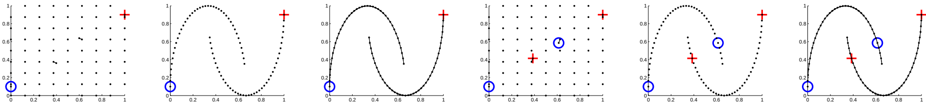
$$x_1 = (0, 0.1), x_2 = (1, 0.9), x_3 = (0.39, 0.41), x_4 = (0.61, 0.59)$$





(demo)

Six Tasks

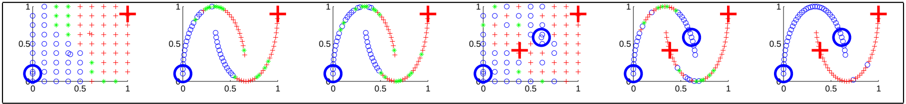


$2^l\,\text{grid}^{\mathrm{U}}$ $\quad$ $2^l\,\text{moons}^{\mathrm{U}}$ $\quad$ $2^l\,\text{moons}^{\mathrm{U}}\,\text{h}$ $\quad$ $4^l\,\text{grid}^{\mathrm{U}}$ $\quad$ $4^l\,\text{moons}^{\mathrm{U}}$ $\quad$ $4^l\,\text{moons}^{\mathrm{U}}\,\text{h}$

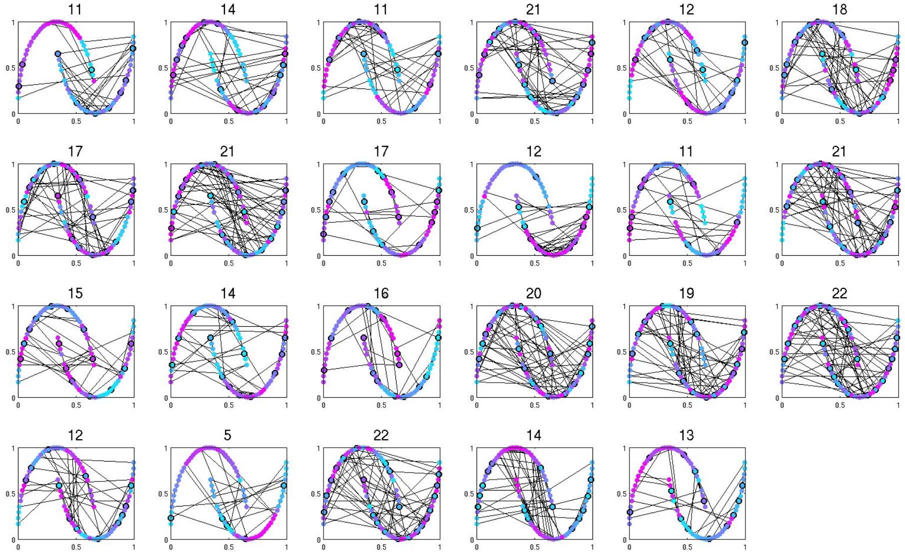
Human Behaviors (Majority Vote)

Majority vote

2^lgrid$^{\mathrm{U}}$	2^lmoons$^{\mathrm{U}}$	2^lmoons$^{\mathrm{U}}$h	4^lgrid$^{\mathrm{U}}$	4^lmoons$^{\mathrm{U}}$	4^lmoons$^{\mathrm{U}}$h



Humans are Probably Not *Just* Following Highlighting



Human Model Selection

$\boxed{\text{axis-parallel} \gg \text{graph (with highlighting)} > \text{other} > \text{graph (no highlighting)}}$

Can be explained by Bayesian model selection...

Bayesian Model Selection

- 7 Gaussian Process models: kernel (covariance matrix) $k_1 \dots k_7$
- Our model is a convex combination

$$k(\lambda) = \sum_{i=1}^{7} \lambda_i k_i, \quad \text{s.t. } \lambda_i \geq 0, \sum_{i=1}^{7} \lambda_i = 1$$

- The best weights can be found via *evidence maximization* (assume uniform prior over λ):

$$\begin{aligned} \max_{\lambda} \quad & p(y_{1:l} \mid x_{1:l}, \lambda) \\ \text{s.t.} \quad & \lambda_i \geq 0, \sum_{i=1}^{7} \lambda_i = 1 \end{aligned}$$

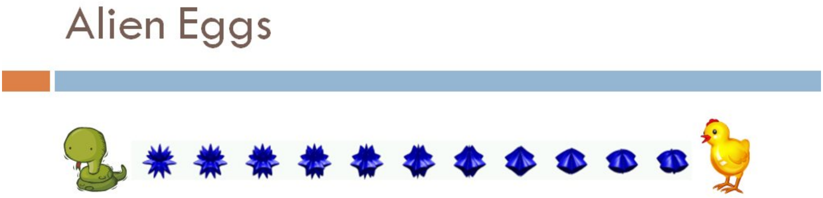
Bayesian Model Selection Explanations

- no manifold learning without highlighting: people don't have k_{graph}
- no manifold learning in $2^{l}_{\text{moons}}{}^{\mathrm{U}}\text{h}$
  - many optimal λ with evidence 0.25, mean is $(0, 0.27, 0.25, 0.22, 0.26, 0, 0)$
  - "manifold learning" $\lambda = (1, 0, 0, 0, 0, 0, 0)$ has inferior evidence 0.249
- yes in $4^{l}_{\text{moons}}{}^{\mathrm{U}}\text{h}$
  - "manifold learning" $\lambda = (1, 0, 0, 0, 0, 0, 0)$ has largest evidence 0.0626
  - all other λ's have inferior evidence

Outline

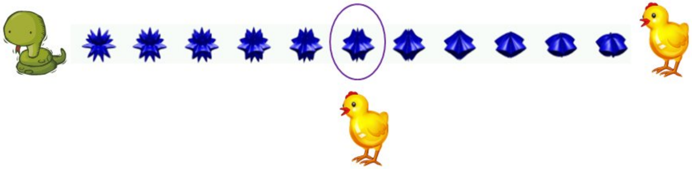
1. Overfitting in Humans
2. Human Manifold Learning
3. Active Learning in Humans

Active Learning in Humans [NIPS 2008]

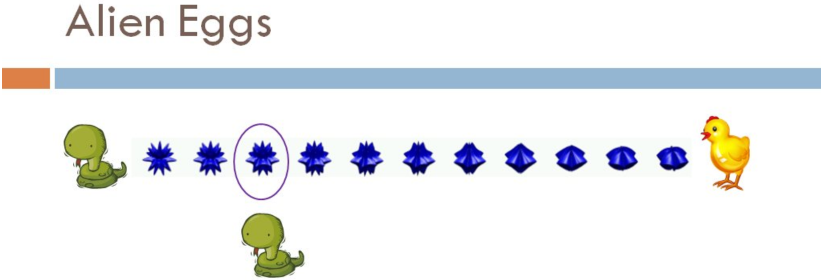


Phenomenon 2: Active Learning [NIPS 2008]

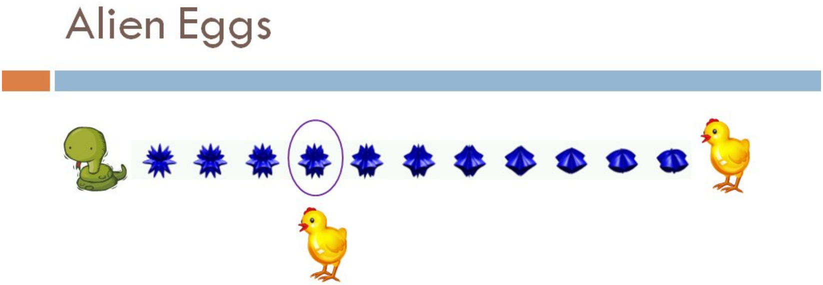
Alien Eggs



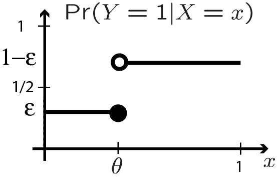
Phenomenon 2: Active Learning [NIPS 2008]



Phenomenon 2: Active Learning [NIPS 2008]

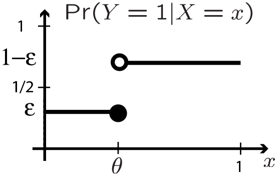


Active Learning



- $\mathcal{X} = [0, 1], \mathcal{Y} = \pm 1$
- unknown threshold $\theta \in [0, 1]$
- label noise $\epsilon > 0$ (no longer binary search!)

Active Learning



- $\mathcal{X} = [0, 1], \mathcal{Y} = \pm 1$
- unknown threshold $\theta \in [0, 1]$
- label noise $\epsilon > 0$ (no longer binary search!)
- goal: learn θ from training data $(x_1, y_1), (x_2, y_2) \dots$
  - passive learning: x_i uniform random
  - active learning: learner selects x_i

  in either case, the world produces $y_i \sim P(y|x_i)$
- main question: how fast does $|\hat{\theta}_n - \theta|$ decrease?

Theory

Passive learning: the minimax lower bound decreases polynomially

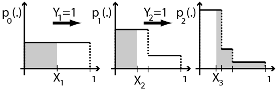
$$\inf_{\hat{\theta}_n} \sup_{\theta \in [0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \geq \frac{1}{4} \left(\frac{1+2\epsilon}{1-2\epsilon} \right)^{2\epsilon} \frac{1}{n+1}$$

Theory

Passive learning: the minimax lower bound decreases polynomially

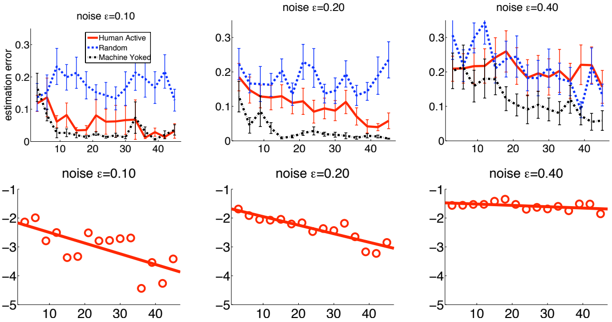
$$\inf_{\hat{\theta}_n} \sup_{\theta \in [0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \geq \frac{1}{4}\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{2\epsilon} \frac{1}{n+1}$$

Active learning: there is a probabilistic bisection algorithm with exponential rate



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

Human Experiment



- human active learning better than passive
- achieves exponential rate (but worse decay constant than theory)
- label noise makes learning harder

Mini Summary

- active learning convergence rate: exponential
- humans can achieve that

Conclusion

Machine learning is not just for machines

Conclusion

Machine learning is not just for machines

- overfitting in humans (Rademacher complexity)

Conclusion

Machine learning is not just for machines

- overfitting in humans (Rademacher complexity)
- manifold learning in humans (Bayesian model selection)

Conclusion

Machine learning is not just for machines

- overfitting in humans (Rademacher complexity)
- manifold learning in humans (Bayesian model selection)
- active learning in humans (exponential rate)

Conclusion

Machine learning is not just for machines

- overfitting in humans (Rademacher complexity)
- manifold learning in humans (Bayesian model selection)
- active learning in humans (exponential rate)
- ...

Conclusion

Machine learning is not just for machines

- overfitting in humans (Rademacher complexity)
- manifold learning in humans (Bayesian model selection)
- active learning in humans (exponential rate)
- ...

Next step: bring insights from humans to machine learning.

Conclusion

Machine learning is not just for machines

- overfitting in humans (Rademacher complexity)
- manifold learning in humans (Bayesian model selection)
- active learning in humans (exponential rate)
- ...

Next step: bring insights from humans to machine learning.

Acknowledgment: I thank my collaborators Rui Castro, Bryan Gibson, Joe Harrison, Chuck Kalish, Rob Nowak, Richard Qian, and Tim Rogers. Research supported by NSF CAREER award IIS-0953219 and IIS-0916038, AFOSR FA9550-09-1-0313, and the Wisconsin Alumni Research Foundation.