Machine Learning Theory
by the People, for the People, of the People

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

2011

war	moonlight	fever	bravery	lice
A	B	A	B	A

Notation

- $\mathcal{X}$ domain, e.g., a finite set of words

Notation

- $\mathcal{X}$ domain, e.g., a finite set of words
- P_X, e.g., uniform

Notation

- $\mathcal{X}$ domain, e.g., a finite set of words
- P_X, e.g., uniform
- $f : \mathcal{X} \mapsto \{-1, 1\}$ classifier

Notation

- $\mathcal{X}$ domain, e.g., a finite set of words
- P_X, e.g., uniform
- $f : \mathcal{X} \mapsto \{-1, 1\}$ classifier
- $f \in \mathcal{F}$ hypothesis space

In-class exam

bravery fever lice moonlight war

Take home exam

cowardice daylight fun hero screech

Did you overfit?

- $(x,y) \overset{iid}{\sim} P_{XY}$

Did you overfit?

- $(x, y) \overset{iid}{\sim} P_{XY}$
- training error: $\hat{e}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i \neq f(x_i))$

Did you overfit?

- $(x, y) \overset{iid}{\sim} P_{XY}$
- training error: $\hat{e}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i \neq f(x_i))$
- true error: $e(f) = \mathbb{E}_{(x,y) \overset{iid}{\sim} P_{XY}} \left[(y \neq f(x)) \right]$

Did you overfit?

- $(x, y) \overset{iid}{\sim} P_{XY}$
- training error: $\hat{e}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i \neq f(x_i))$
- true error: $e(f) = \mathbb{E}_{(x,y) \overset{iid}{\sim} P_{XY}} \left[(y \neq f(x)) \right]$
- want a bound

$$e(f) \leq \hat{e}(f) + \text{something}$$

Rademacher bound [Bartlett and Mendelson 2002]

On a training set of size n, w.p. at least $1-\delta$, $\forall f \in \mathcal{F}$:

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

Rademacher complexity [Bartlett and Mendelson 2002]

- $\boldsymbol{x} = x_1, \ldots, x_n \overset{iid}{\sim} P_X$

Rademacher complexity [Bartlett and Mendelson 2002]

- $\boldsymbol{x} = x_1, \ldots, x_n \overset{iid}{\sim} P_X$
- $\boldsymbol{\sigma} = \sigma_1, \ldots, \sigma_n \overset{iid}{\sim} \mathrm{Bernoulli}(\frac{1}{2}, \frac{1}{2})$ with values ± 1

Rademacher complexity [Bartlett and Mendelson 2002]

- $\boldsymbol{x} = x_1, \ldots, x_n \overset{iid}{\sim} P_X$
- $\boldsymbol{\sigma} = \sigma_1, \ldots, \sigma_n \overset{iid}{\sim} \mathrm{Bernoulli}(\frac{1}{2}, \frac{1}{2})$ with values ± 1
- fit of f: $|\sum_{i=1}^{n} \sigma_i f(x_i)|$

Rademacher complexity [Bartlett and Mendelson 2002]

- $\boldsymbol{x} = x_1, \ldots, x_n \overset{iid}{\sim} P_X$
- $\boldsymbol{\sigma} = \sigma_1, \ldots, \sigma_n \overset{iid}{\sim} \mathrm{Bernoulli}(\frac{1}{2}, \frac{1}{2})$ with values ± 1
- fit of f: $|\sum_{i=1}^{n} \sigma_i f(x_i)|$
- fit of $\mathcal{F}$: $\sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \sigma_i f(x_i)|$

Rademacher complexity [Bartlett and Mendelson 2002]

- $\boldsymbol{x} = x_1, \ldots, x_n \overset{iid}{\sim} P_X$
- $\boldsymbol{\sigma} = \sigma_1, \ldots, \sigma_n \overset{iid}{\sim} \mathrm{Bernoulli}(\frac{1}{2}, \frac{1}{2})$ with values ± 1
- fit of f: $|\sum_{i=1}^{n} \sigma_i f(x_i)|$
- fit of $\mathcal{F}$: $\sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \sigma_i f(x_i)|$
- Rademacher complexity

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

Estimating human Rademacher complexity [NIPS 09]

"Learning the noise"

1. participant study $\{(x_i, \sigma_i)\}_{i=1}^n$
2. filler task
3. classify $\{x_i\}_{i=1}^n$: re-ordered; not told these were training items

At the end, we observe $\hat{f}(x_1) \ldots \hat{f}(x_n)$ from the human.

Estimating human Rademacher complexity (cont.)

Key assumption:

$$\hat{f} = \arg\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x_i)$$

Estimating human Rademacher complexity (cont.)

Key assumption:

$$\hat{f} = \arg\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x_i)$$

therefore,

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

Estimating human Rademacher complexity (cont.)

Key assumption:

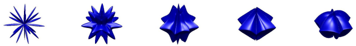
$$\hat{f} = \arg\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x_i)$$

therefore,

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

Averaging over participants gives an estimate of R.

Human Rademacher complexity

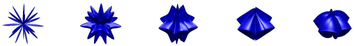


the Shape $\mathcal{X}$

rape killer funeral $\cdots$ fun laughter joy

the Word $\mathcal{X}$

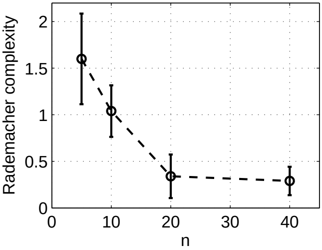
Human Rademacher complexity



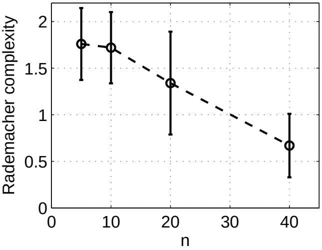
the Shape $\mathcal{X}$

rape killer funeral $\cdots$ fun laughter joy

the Word $\mathcal{X}$



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



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

Does the bound work?

"Learning any task"

1. participant study $\{(x_i, y_i)\}_{i=1}^{n}$
2. filler task
3. classify $\{x_i\}_{i=1}^{n+100}$: re-ordered; not told some were training items

Yes the bound works

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

condition	subject	$\hat{e}$	bound	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

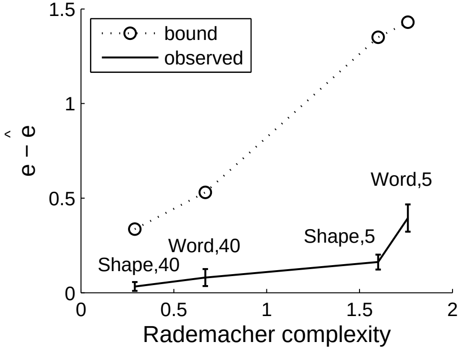
Oh how they overfit!

- mnemonics
  - (grenade, B), (skull, A), (conflict, A), (meadow, B), (queen, B)
  - "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)"

Oh how they overfit!

- mnemonics
  - (grenade, B), (skull, A), (conflict, A), (meadow, B), (queen, B)
  - "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)"
- idiosyncratic rules
  - whether the shape "faces downward"
  - whether the word "tastes good"
  - "anything related to omitting (sic) light"
  - "things you can go inside"
  - odd or even number of syllables
  - "relates to motel service"
  - "physical vs. abstract"

Smaller Rademacher complexity, less actual overfitting



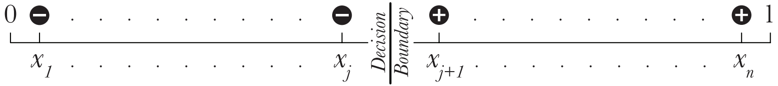
now	you	be	the	teacher
B	B	B	B	B

The teaching dimension [Goldman and Kearns 1995]

The teaching dimension [Goldman and Kearns 1995]

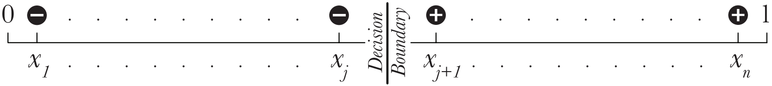
- items $\mathcal{X}$

The teaching dimension [Goldman and Kearns 1995]



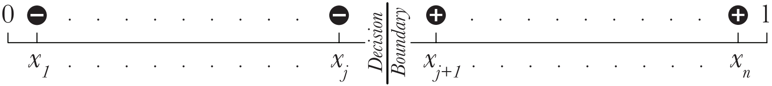
- items $\mathcal{X}$
- H threshold functions

The teaching dimension [Goldman and Kearns 1995]



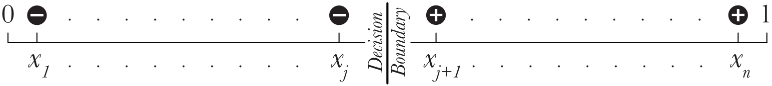
- items $\mathcal{X}$
- H threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with h only

The teaching dimension [Goldman and Kearns 1995]



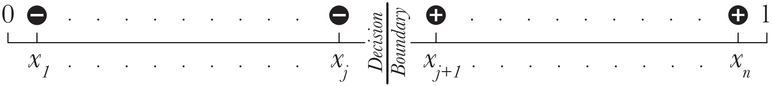
- items $\mathcal{X}$
- H threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with h only
- $TD(h)$: size of the smallest teaching set of h, 1 or 2

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]



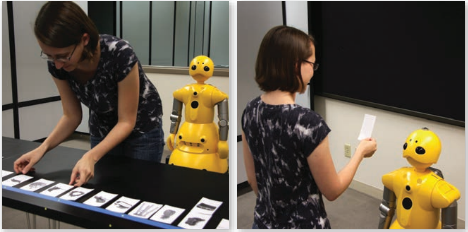
- items $\mathcal{X}$
- H threshold functions
- teaching set of $h \in \mathcal{H}$: subset of $\mathcal{X}$ consistent with h only
- $TD(h)$: size of the smallest teaching set of h, 1 or 2
- $TD(H)$: $TD(h^*)$ for the hardest $h^* \in H$, 2

Optimal teaching should start around the decision boundary.

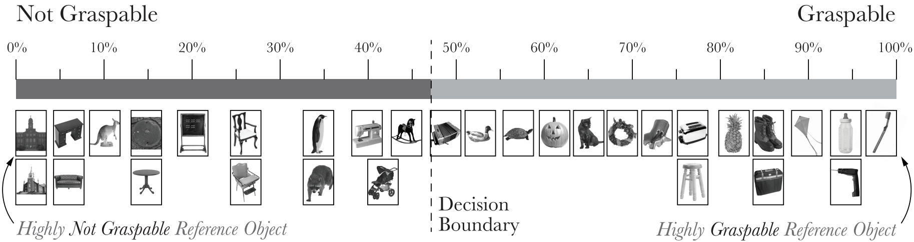
Curriculum learning [Bengio et al. 2009]

Teaching should start from easy to hard, i.e., outside to inside.

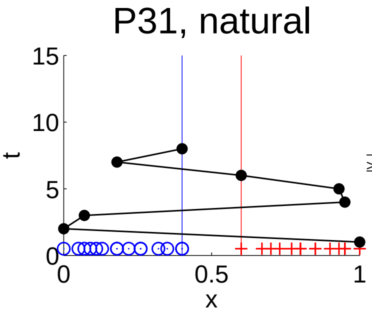
You teach robot ... [to appear at NIPS 11]

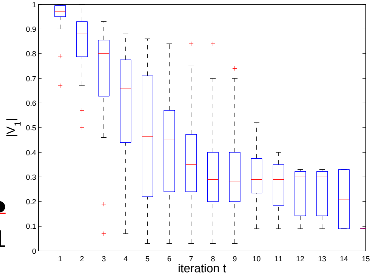


... graspability

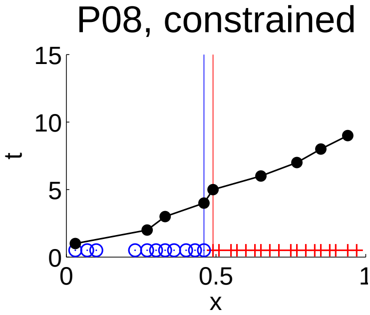


Observed human teaching strategy 1





Observed human teaching strategy 2



Extending teaching dimension for curriculum learning

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{unif}[0,1]^d$

Idealized assumptions

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

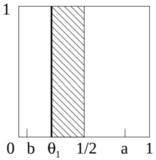
Idealized assumptions

- available teaching items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{unif}[0,1]^d$
- 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

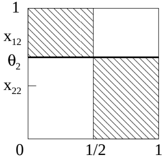
- available teaching items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \text{unif}[0,1]^d$
- first dim determines label $p(y_i = 1 \mid \mathbf{x}_i) = \mathbb{1}_{\{x_{i1} > \frac{1}{2}\}}$
- learner's version space V: axis-parallel decision boundaries
  - after two teaching items $(\mathbf{x}_1, 1), (\mathbf{x}_2,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

dim 1



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

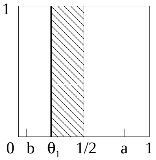
dim 2



Idealized assumptions

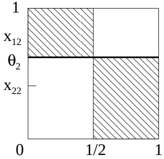
- available teaching items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathrm{unif}[0,1]^d$
- first dim determines label $p(y_i = 1 \mid \mathbf{x}_i) = \mathbb{1}_{\{x_{i1} > \frac{1}{2}\}}$
- learner's version space V: axis-parallel decision boundaries
  - after two teaching items $(\mathbf{x}_1, 1), (\mathbf{x}_2, 0)$

dim 1



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

dim 2



- learner is a Gibbs classifier

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

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

- 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 $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$.

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 $d-1$ $\mathrm{Beta}(1,2)$ 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 $c \equiv \sum_{k=2}^{d} |x_{1k} - x_{2k}|$.

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

Corollary

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

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 $d-1$ $\mathrm{Beta}(1,2)$ 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

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

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

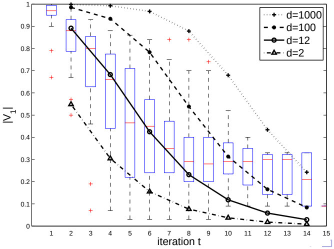
Teaching items should approach decision boundary

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

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

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



Conclusion

Machine learning and cognitive science have much to offer to each other.

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