Machine Learning
(機器學習)

Lecture 3: Feasibility of Learning

Hsuan-Tien Lin (林軒田)

htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)

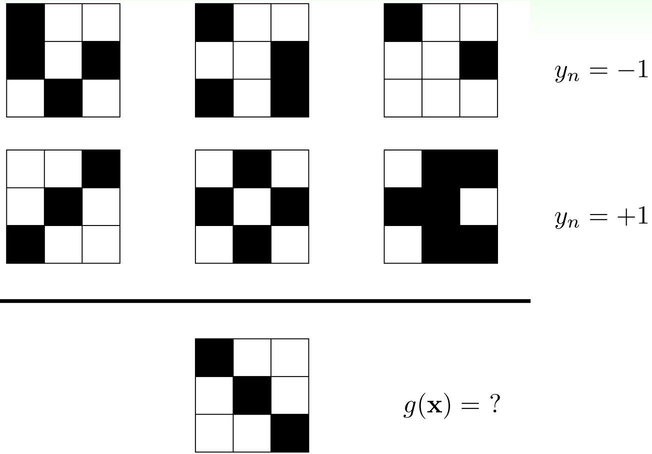
Roadmap

1. **When** Can Machines Learn?

Lecture 3: Feasibility of Learning

- Learning is Impossible?
- Probability to the Rescue
- Connection to Learning
- Connection to Real Learning
- Feasibility of Learning Decomposed

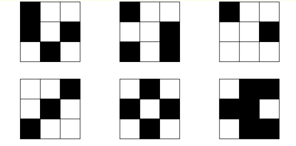
A Learning Puzzle



let's test your 'human learning' with 6 examples :-)

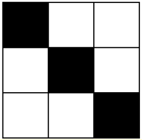
Two Controversial Answers

whatever you say about $g(\mathbf{x})$,



$y_n = -1$

$y_n = +1$



$$g(\mathbf{x}) = \ ?$$

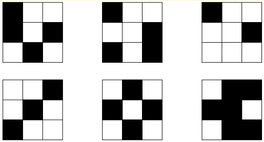
truth $f(\mathbf{x}) = +1$ because ...

truth $f(\mathbf{x}) = -1$ because ...

which reason is **correct**?

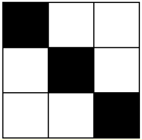
Two Controversial Answers

whatever you say about $g(\mathbf{x})$,



$y_n = -1$

$y_n = +1$



$$g(\mathbf{x}) = ?$$

truth $f(\mathbf{x}) = +1$ because . . .

- symmetry $\Leftrightarrow$ +1
- (black or white count = 3) or (black count = 4 and middle-top black) $\Leftrightarrow$ +1

truth $f(\mathbf{x}) = -1$ because . . .

- left-top black $\Leftrightarrow$ -1
- middle column contains at most 1 black and right-top white $\Leftrightarrow$ -1

all valid reasons, your **adversarial teacher** can always call you '**didn't learn**'. **:-(**

A Brain-Storming Problem

$$(5, 3, 2) \to 151022, \qquad (7, 2, 5) \to \textbf{?}$$

It is like a 'learning problem' with $N = 1$, $\mathbf{x}_1 = (5, 3, 2)$, $y_1 = 151022$.
Learn a hypothesis from the one example to predict on $\mathbf{x} = (7, 2, 5)$.
What is your answer?

151026

$$g(\mathbf{x}) = 151012 + x_1 + x_2 + x_3$$

143547

$$\begin{aligned} g(\mathbf{x}) &= x_1 \cdot x_2 \cdot 10000 \\ &+ x_1 \cdot x_3 \cdot 100 \\ &+ (x_1 \cdot x_2 + x_1 \cdot x_3 - x_2) \end{aligned}$$

which one is the **smarter** answer that only top 2% people can think of?

What is the Next Number?

1,4,1,5

What is the Next Number?

1,4,1,5

1,4,1,5,**0**,-1,1,6
by $y_t = y_{t-4} - y_{t-2}$

1,4,1,5,**1**,6,1,7
by $y_t = y_{t-2} + [\![t \text{ is even}]\!]$

1,4,1,5,**2**,9,3,14
by $y_t = y_{t-4} + y_{t-2}$

any number can be the next!

A ‘Simple’ Binary Classification Problem

$\mathbf{x}_n$	$y_n = f(\mathbf{x}_n)$
0 0 0	○
0 0 1	×
0 1 0	×
0 1 1	○
1 0 0	×

- $\mathcal{X} = \{0, 1\}^3$, $\mathcal{Y} = \{\circ, \times\}$, can enumerate all candidate f as $\mathcal{H}$

pick $g \in \mathcal{H}$ with all $g(\mathbf{x}_n) = y_n$ (like PLA),
does $g \approx f$?

Infeasibility of Learning

	$\mathbf{x}$	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\mathcal{D}$	0 0 0	○	○	○	○	○	○	○	○	○	○
$\mathcal{D}$	0 0 1	×	×	×	×	×	×	×	×	×	×
$\mathcal{D}$	0 1 0	×	×	×	×	×	×	×	×	×	×
$\mathcal{D}$	0 1 1	○	○	○	○	○	○	○	○	○	○
$\mathcal{D}$	1 0 0	×	×	×	×	×	×	×	×	×	×
	1 0 1		?	○	○	○	○	×	×	×	×
	1 1 0		?	○	○	×	×	○	○	×	×
	1 1 1		?	○	×	○	×	○	×	○	×

- $g \approx f$ inside $\mathcal{D}$: sure!
- $g \approx f$ outside $\mathcal{D}$: **No!** (but that's really what we want!)

learning from $\mathcal{D}$ (to infer something outside $\mathcal{D}$)
is doomed if **any 'unknown' f can happen**. **:-(**

No Free Lunch Theorem for Machine Learning

Without any assumptions on the learning problem on hand,
all learning algorithms perform the same.

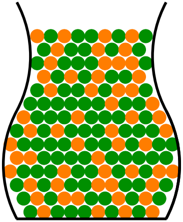
(CC-BY-SA 2.0 by Gaspar Torriero on Flickr)

no algorithm is best
for all learning problems

Questions?

Inferring Something Unknown with Assumptions

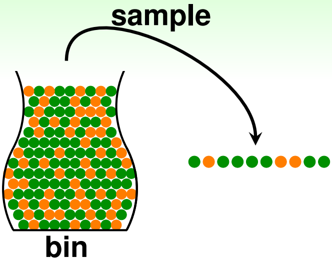
difficult to infer **unknown target f outside $\mathcal{D}$** in learning;
can we infer **something unknown** in **other scenarios**?



- consider a bin of many many orange and green marbles
- do we **know** the orange portion (probability)? **No!**

can you **infer** the orange probability?

Statistics 101: Inferring Orange Probability



bin

assume

orange probability = μ,
green probability = $1 - \mu$,
with μ **unknown**

sample

assume *N* marbles sampled independently:

orange fraction = ν,
green fraction = $1 - \nu$,

now ν **known**

does **in-sample** ν say anything about out-of-sample μ?

Possible versus Probable

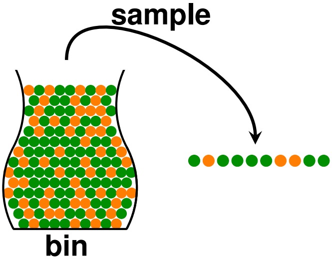
does **in-sample** ν say anything about out-of-sample μ?

No!

possibly not: sample can be mostly green while bin is mostly orange

Yes!

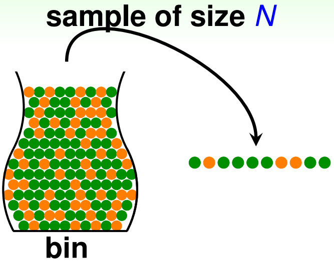
probably yes: in-sample ν likely **close to** unknown μ



formally, **what does ν say about μ?**

Hoeffding's Inequality (1/2)

$\mu =$ orange
probability in bin



$\nu =$ orange
fraction in sample

- in big sample (N large), ν is probably close to μ (within ϵ)

$$\mathbb{P}\left[\left|\nu - \mu\right| > \epsilon\right] \leq 2\exp\left(-2\epsilon^2 N\right)$$

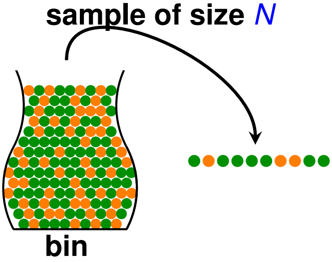
- called **Hoeffding's Inequality**, for marbles, coin, polling, . . .

the statement '$\nu = \mu$' is
probably approximately correct (PAC)

Hoeffding's Inequality (2/2)

$$\mathbb{P}\left[\left|\nu - \mu\right| > \epsilon\right] \le 2\exp\left(-2\epsilon^2 N\right)$$

- valid for all N and ϵ
- does not depend on μ, **no need to 'know' μ**
- larger sample size N or looser gap ϵ $\Longrightarrow$ higher probability for '$\nu \approx \mu$'



if **large N**, can **probably** infer unknown μ by known ν (under iid sampling assumption)

Questions?

Connection to Learning

bin

- unknown orange prob. μ
- marble $\bullet \in$ bin
- orange $\bullet$
- green $\bullet$
- size-N sample from bin

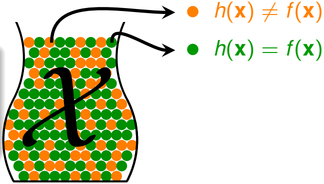
  of i.i.d. marbles

learning

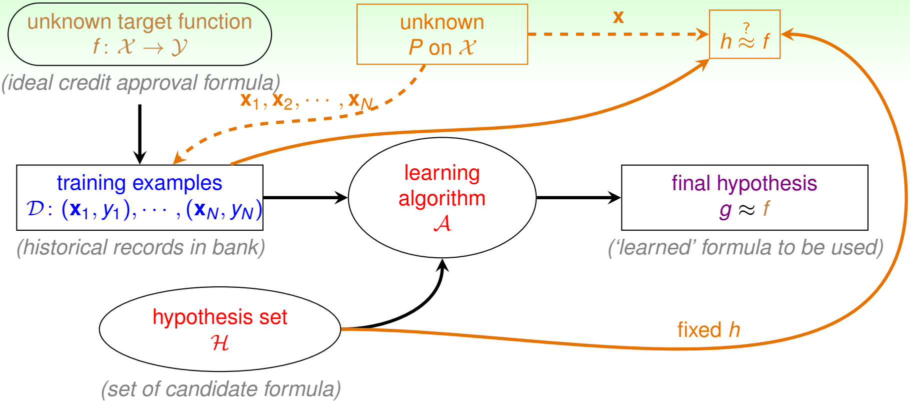
- fixed hypothesis $h(\mathbf{x}) \stackrel{?}{=}$ target $f(\mathbf{x})$
- $\mathbf{x} \in \mathcal{X}$
- h is wrong $\Leftrightarrow h(\mathbf{x}) \neq f(\mathbf{x})$
- h is right $\Leftrightarrow h(\mathbf{x}) = f(\mathbf{x})$
- check h on $\mathcal{D} = \{(\mathbf{x}_n, \underbrace{y_n}_{f(\mathbf{x}_n)})\}$

  with i.i.d. $\mathbf{x}_n$

if **large** N **& i.i.d.** $\mathbf{x}_n$, can **probably** infer
unknown $[\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$ probability
by known $[\![h(\mathbf{x}_n) \neq y_n]\!]$ fraction



Added Components



for any fixed h, can probably infer

$$\textbf{unknown } E_{\text{out}}(\mathbf{h}) = \underset{\mathbf{x} \sim P}{\mathcal{E}} [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$$

$$\text{by } \textbf{known } E_{\text{in}}(\mathbf{h}) = \frac{1}{N} \sum_{n=1}^{N} [\![h(\mathbf{x}_n) \neq y_n]\!]$$

(under iid sampling assumption)

The Formal Guarantee

for any fixed h, in 'big' data (N large),

$\boxed{\text{in-sample error } E_{\text{in}}(h)}$ is probably close to

$\boxed{\text{out-of-sample error } E_{\text{out}}(h)}$ (within ϵ)

$$\mathbb{P}\left[\left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \le 2\exp\left(-2\epsilon^2 N\right)$$

same as the 'bin' analogy ...

- valid for all N and ϵ
- does not depend on $E_{\text{out}}(h)$, **no need to 'know'** $E_{\text{out}}(h)$
  —f and P can stay unknown
- '$E_{\text{in}}(h) = E_{\text{out}}(h)$' is **probably approximately correct (PAC)**

if '$E_{\text{in}}(h) \approx E_{\text{out}}(h)$' and '$E_{\text{in}}(h)$ **small**'
$\Longrightarrow E_{\text{out}}(h)$ small $\Longrightarrow h \approx f$ with respect to P

Verification of One h

for any fixed h, when data large enough,

$$E_{\text{in}}(h) \approx E_{\text{out}}(h)$$

Can we claim 'good learning' ($g \approx f$)?

Yes!

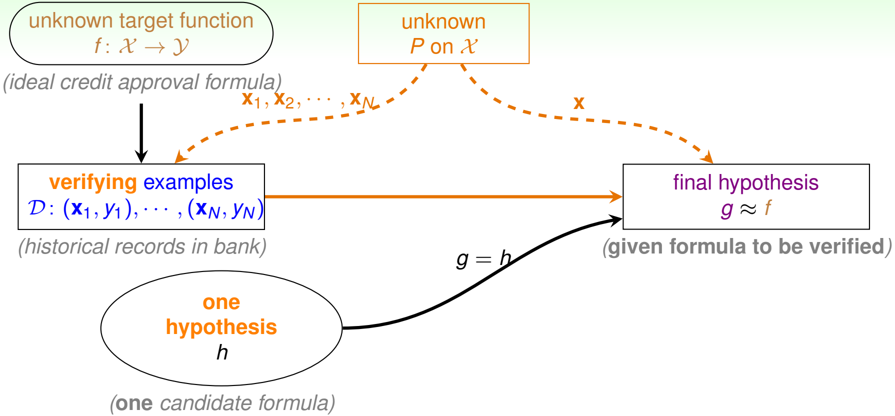
if $E_{\text{in}}(h)$ **small for the fixed** h
and $\mathcal{A}$ **pick the** h **as** g
$\Longrightarrow$ '$g = f$' PAC

No!

if $\mathcal{A}$ **forced to pick THE** h **as** g
$\Longrightarrow$ $E_{\text{in}}(h)$ **almost always not small**
$\Longrightarrow$ '$g \neq f$' PAC!

real learning:
$\mathcal{A}$ shall **make choices** $\in \mathcal{H}$ (like PLA)
rather than **being forced to pick one** h**. :-(**

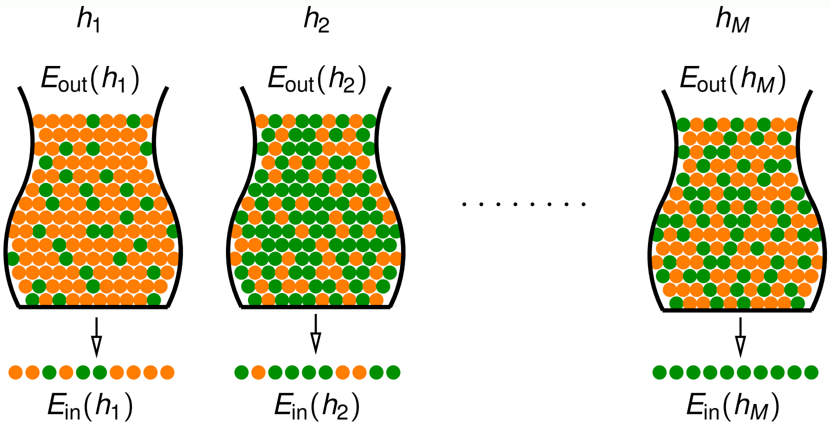
The ‘Verification’ Flow



can now use ‘historical records’ (data) to
verify ‘one candidate formula’ h

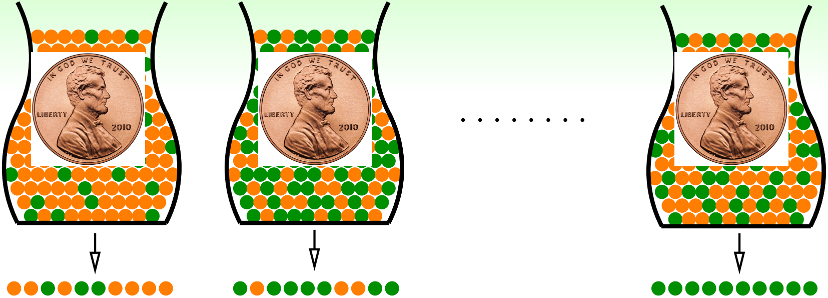
Questions?

Multiple h



real learning (say like PLA):
BINGO when getting ●●●●●●●●●●?

Coin Game



Q: if everyone in size-400 NTU ML class flips a coin 5 times, and **one of the students gets 5 heads for her coin 'g'**. Is 'g' really magical?

A: No. Even if all coins are fair, the probability that **one of the coins** results in **5 heads** is $1 - \left(\frac{31}{32}\right)^{400} > 99\%$.

BAD sample: E_{in} and E_{out} far away
—can get worse when involving 'choice'

BAD Sample and BAD Data

BAD Sample

e.g., $E_{\text{out}} = \frac{1}{2}$, but getting all heads ($E_{\text{in}} = 0$)!

BAD Data for One h

$E_{\text{out}}(h)$ and $E_{\text{in}}(h)$ far away:
e.g., E_{out} big (far from f), but E_{in} small (correct on most examples)

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	...	Hoeffding
h	**BAD**					**BAD**		$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h] $\leq \ldots$

Hoeffding: small

$$\mathbb{P}_{\mathcal{D}}\left[\textbf{BAD } \mathcal{D}\right] = \sum_{\text{all possible} \mathcal{D}} \mathbb{P}(\mathcal{D}) \cdot [\![\textbf{BAD } \mathcal{D}]\!]$$

BAD Data for Many h

GOOD data for many h
$\Longleftrightarrow$ **GOOD** data for verifying any h
$\Longleftrightarrow$ there exists **no BAD** h such that $E_{\text{out}}(h)$ and $E_{\text{in}}(h)$ far away
there exists some h **such that** $E_{\text{out}}(h)$ and $E_{\text{in}}(h)$ far away
$\Longleftrightarrow$ **BAD** data for many h

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	Hoeffding
h_1	**BAD**					**BAD**	$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_1] $\leq \ldots$
h_2		**BAD**					$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_2] $\leq \ldots$
h_3	**BAD**	**BAD**				**BAD**	$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_3] $\leq \ldots$
...							
h_M	**BAD**					**BAD**	$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_M] $\leq \ldots$
all	**BAD**	**BAD**		**GOOD**		**BAD**	**?**

do *not* know if $\mathcal{D}$ is **BAD** or not;
wish $\mathbb{P}_{\mathcal{D}}$[**BAD** $\mathcal{D}$] small & pray for "**GOOD luck**"

Bound of BAD Data

$$
\begin{aligned}
& \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D}] \\
= \quad & \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_1 \textbf{ or } \textbf{BAD } \mathcal{D} \text{ for } h_2 \textbf{ or } \ldots \textbf{ or } \textbf{BAD } \mathcal{D} \text{ for } h_M] \\
\leq \quad & \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_1] + \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_2] + \ldots + \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_M] \\
& \text{(union bound)} \\
\leq \quad & 2\exp\left(-2\epsilon^2 N\right) + 2\exp\left(-2\epsilon^2 N\right) + \ldots + 2\exp\left(-2\epsilon^2 N\right) \\
= \quad & 2M\exp\left(-2\epsilon^2 N\right)
\end{aligned}
$$

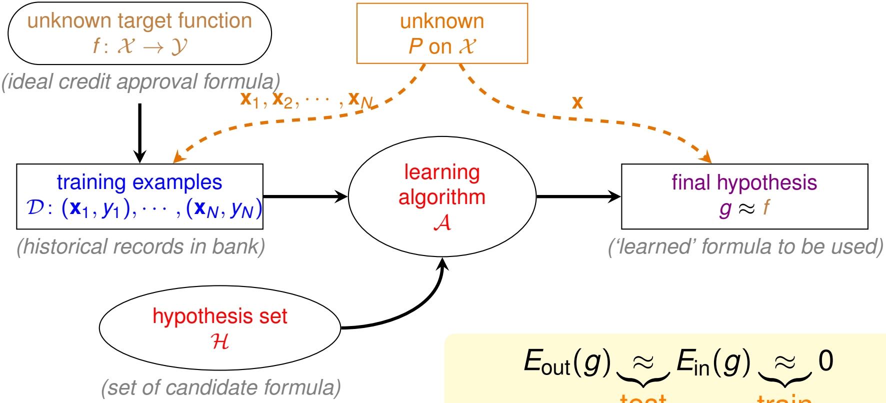
- finite-bin version of Hoeffding, valid for all M, N and ϵ
- does not depend on any $E_{\text{out}}(h_m)$, **no need to 'know'** $E_{\text{out}}(h_m)$ —f and P can stay unknown
- '$E_{\text{in}}(g) = E_{\text{out}}(g)$' is **PAC**, **regardless of** $\mathcal{A}$

'most reasonable' $\mathcal{A}$ (like PLA):
pick the h_m with **lowest** $E_{\text{in}}(h_m)$ as g

Questions?

The ‘Statistical’ Learning Flow

if $|\mathcal{H}| = M$ finite, N large enough,
for whatever g picked by $\mathcal{A}$, $E_{\text{out}}(g) \approx E_{\text{in}}(g)$
if $\mathcal{A}$ finds one g with $E_{\text{in}}(g) \approx 0$,
PAC guarantee for $E_{\text{out}}(g) \approx 0 \Longrightarrow$ **learning possible :-)**



$$E_{\text{out}}(g) \underbrace{\approx}_{\text{test}} E_{\text{in}}(g) \underbrace{\approx}_{\text{train}} 0$$

Two Central Questions

for $\underbrace{\text{batch \& supervised binary classification}}_{\text{lecture 2}}$, $\underbrace{g \approx f}_{\text{lecture 1}} \iff E_{\text{out}}(g) \approx 0$

achieved through $\underbrace{E_{\text{out}}(g) \approx E_{\text{in}}(g)}_{\text{lecture 3}}$ and $\underbrace{E_{\text{in}}(g) \approx 0}_{\text{lecture 1}}$

learning split to two central questions:

1. can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$? (test/generalize)
2. can we make $E_{\text{in}}(g)$ small enough? (train/optimize)

what role does $\underbrace{M}_{|\mathcal{H}|}$ play for the two questions?

Trade-off on M

1. can we make sure that $E_{out}(g)$ is close enough to $E_{in}(g)$?
2. can we make $E_{in}(g)$ small enough?

small M

1. Yes!,
   $\mathbb{P}[\textbf{BAD}] \leq 2 \cdot M \cdot \exp(\ldots)$
2. No!, too few choices

large M

1. No!,
   $\mathbb{P}[\textbf{BAD}] \leq 2 \cdot M \cdot \exp(\ldots)$
2. Yes!, many choices

using the right M (or $\mathcal{H}$) is important

$M = \infty$ **doomed?**

Preview

Known

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \le 2 \cdot M \cdot \exp\left(-2\epsilon^2 N\right)$$

Todo

- establish **a finite quantity** that replaces M

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \overset{?}{\le} 2 \cdot m_{\mathcal{H}} \cdot \exp\left(-2\epsilon^2 N\right)$$

- justify the feasibility of learning for infinite M
- study $m_{\mathcal{H}}$ to understand its trade-off for 'right' $\mathcal{H}$, just like M

mysterious PLA to be fully resolved
"soon" :-)

Questions?

Summary

1. **When** Can Machines Learn?

Lecture 2: The Learning Problems

Lecture 3: Feasibility of Learning

- Learning is Impossible?
  absolutely no free lunch outside $\mathcal{D}$
- Probability to the Rescue
  probably approximately correct outside $\mathcal{D}$
- Connection to Learning
  verification possible if $E_{\text{in}}(h)$ small for fixed h
- Connection to Real Learning
  learning possible if $|\mathcal{H}|$ finite and $E_{\text{in}}(g)$ small
- Feasibility of Learning Decomposed
  two questions: $E_{\text{out}}(g) \approx E_{\text{in}}(g)$, and $E_{\text{in}}(g) \approx 0$

2. Why Can Machines Learn?
   - **next: what if $|\mathcal{H}| = \infty$?**