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?

When/where learning is possible

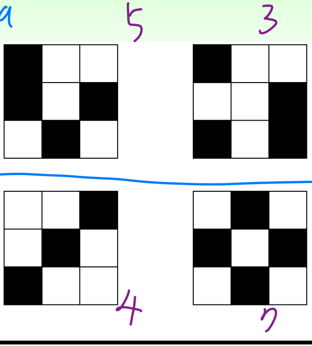
Lecture 3: Feasibility of Learning

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

A Learning Puzzle

training data

Binary classification



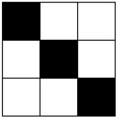
11

$y_n = -1$

$y_n = +1$

16

Batch supervised.



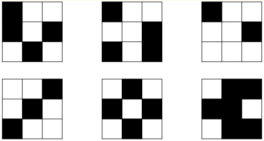
$g(\mathbf{x}) = \,?$ -1

my brain

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

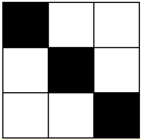
Two Controversial Answers

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



$y_n = -1$

$y_n = +1$



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

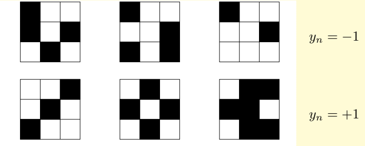
truth $f(\mathbf{x}) = +1$ because . . .

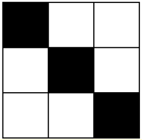
truth $f(\mathbf{x}) = -1$ because . . .

which reason is **correct**?

Two Controversial Answers

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





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

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

- symmetry ⇔ +1 (線對稱)
- (black or white count = 3) or (black count = 4 and middle-top black) ⇔ +1

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

- left-top black ⇔ -1 (只看左上角)
- middle column contains at most 1 black and right-top white ⇔ -1

規則隨人說

⇒所以需要假設

lots of hypothesis.

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

teacher call always call you wrong.

A Brain-Storming Problem

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

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?

offset and sum

151026

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

$7+2+5+151012 = \underline{151026}$ #

correct answer

143547

$$\begin{aligned} g(\mathbf{x}) &= \overset{5\times 3}{x_1 \cdot x_2} \cdot 10000 \\ &+ \overset{5\times 2}{x_1 \cdot x_3} \cdot 100 \\ &+ (\overset{5\times 3}{x_1 \cdot x_2} + \overset{5\times 2}{x_1 \cdot x_3} - \overset{-3}{x_2}) \end{aligned}$$

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	×

3個Bit的training set

其實一共也只有8種可能
這裡已經限制了5種

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

g在所有training data都答对,但g接近f嗎?

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

Infeasibility of Learning

training $\mathcal{D}$

testing

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

f有8种可能,但任选一种都不代表学到了什麼

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

↑ you can be totally wrong.

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

↑ 需要假設

No Free Lunch Theorem for Machine Learning

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

if we allow everything to happen

所以用ML predict random number 不可行!?

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

It Is infeasible for ML to learn.

no algorithm is best
for all learning problems

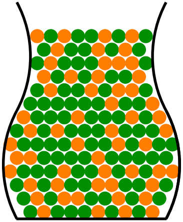
Questions?

Inferring Something Unknown with Assumptions

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

統計学.

bin model

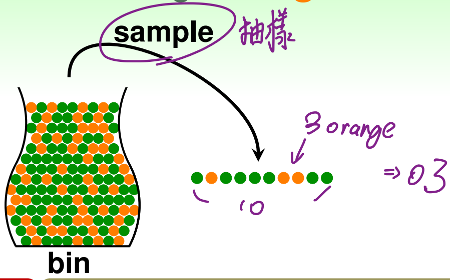


- consider a bin of many many orange and green marbles
- do we **know** the orange portion (probability)? **No!**

can you **infer** the orange probability?

Statistics 101: Inferring Orange Probability

Assumption: Sample independently (uniform)



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? 用sample能回推bin嗎？

Possible versus Probable

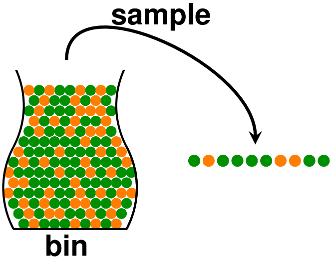
does **in-sample** ν say anything about out-of-sample μ?

No!

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

Yes!

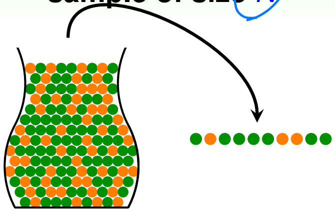
probably yes: in-sample ν likely **close to** unknown μ



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

Hoeffding's Inequality (1/2)

sample of size N



μ = orange
probability in bin

bin

ν = orange
fraction in sample

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

抽樣与母体的差距 ←抽樣的次数

$$\mathbb{P}\left[\left|\underline{\nu-\mu}\right| > \boxed{\epsilon}\right] \le \underbrace{2\exp\left(-2\epsilon^2 \boxed{N}\right)}_{\text{偏差的機率小於這个值}}$$

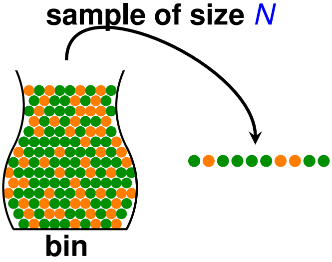
- called **Hoeffding's Inequality**, for marbles, coin, polling, ...

the statement '$\underline{\nu = \mu}$' is
大概是對
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)

$2\exp(-2(0.01)\times 10))$

Questions?

Connection to Learning

i.i.d = Independent and identically distributed

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 $f(\mathbf{x})$

  选定1个固定的h
- marble $\boxed{\mathbf{x}} \in \boxed{\mathcal{X}}$ bin
- h is wrong $\Leftrightarrow$ $h(\mathbf{x}) \neq f(\mathbf{x})$: 漆成orange (一樣)
- h is right $\Leftrightarrow$ $h(\mathbf{x}) = f(\mathbf{x})$: 漆成green
- check h on $\mathcal{D} = \{(\mathbf{x}_n, \underbrace{y_n}_{f(\mathbf{x}_n)})\}$ (不一樣)

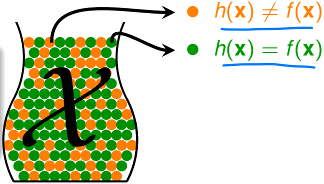
  檢查h在D上的表現

  f(x)的結果

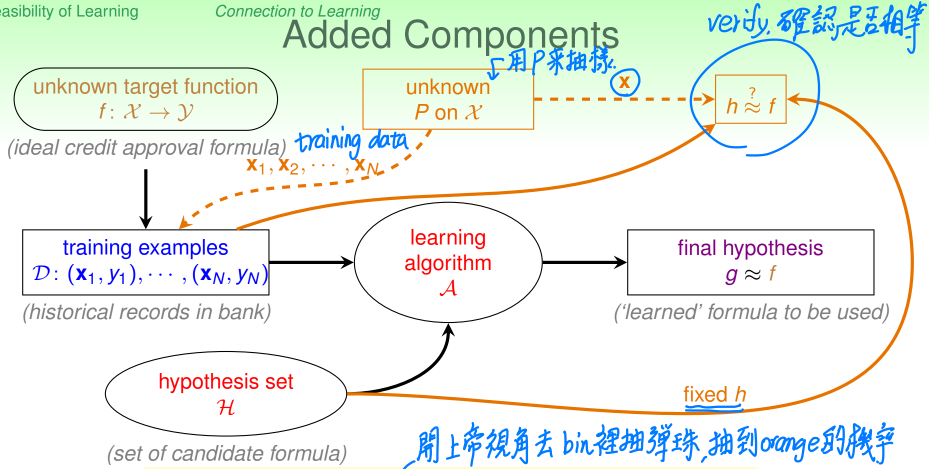
  with i.i.d. $\mathbf{x}_n$

x_n是独立平均的從bin裡抽出來

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



Added Components



關上帝視角去bin裡抽彈珠,抽到orange的機率

for any fixed h, can probably infer

error rate　Exceptation　testing

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

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

*從已經抽樣的training set裡,抽到orange的機率

(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)

\# of sample

$$\mathbb{P}\left[\underbrace{\left|E_{\text{in}}(h) - E_{\text{out}}(h)\right|}_{\text{抽樣內的error 与抽樣外的error 的差距}} > \epsilon\right] \leq 2\exp\left(-2\epsilon^2 \boxed{N}\right)$$

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)**

N 如果夠大就大致上一樣.

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

achieve learning.

Verification of One h

for any fixed h, when data large enough,

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

g是ML選的function.

Can we claim 'good learning' ($g \approx f$)? 想辦法選h接近g

Yes!

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

No!

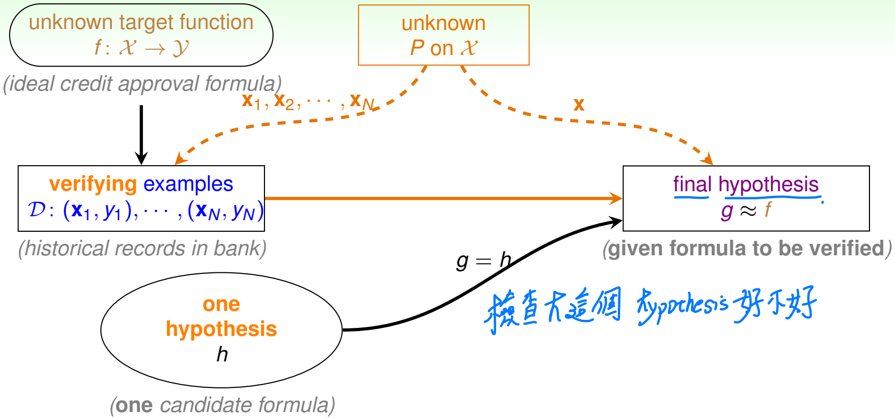
if $\mathcal{A}$ **forced to pick THE** h **as** g
$\Longrightarrow$ $E_{\text{in}}(h)$ **almost always not small**
$\Longrightarrow$ '$g \neq f$' $\boxed{\text{PAC!}}$

用來判斷 g.f 有沒有接近.

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

這邪不是 learning, 因為邪不会做選擇

The 'Verification' Flow (testing)



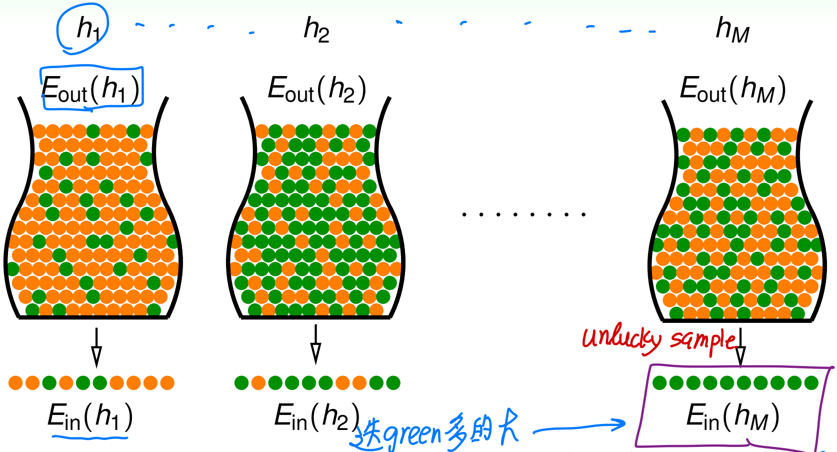
can now use 'historical records' (data) to **verify 'one candidate formula' h**

Questions?

Multiple h

one hypothesis for one bin

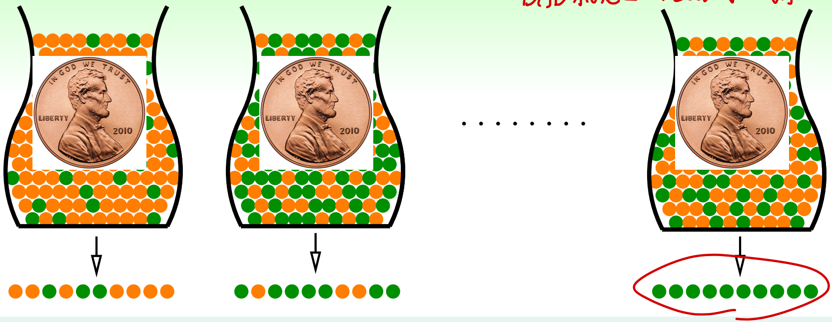
green → correct
orange → incorrect



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

Coin Game

BAD就是Ein, Eout不一樣



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?

高機率不是作弊coin

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

任一個學生丟出連續5個head 機率其實很大

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 E_{out} 和 E_{in} (差很遠)

丟 coin

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)

不會有太多格是不好的

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

*Hoeffding 保證 BAD data 發生機率小

Hoeffding: small

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

BAD Data for Many h

GOOD data for many h

$\iff$ **GOOD** data for verifying any h

Good data can verify every hypothesis

$\iff$ 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

$\iff$ **BAD** data for many h

Super dataset

In real world - we only have one dataset

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	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**	**?**

BAD for some hypothesis

对所有dataset都GOOD

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

壞的 dataset 只要对任一个 hypothesis BAD

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

對任一個 hypothesis 不好的機率

假設都不交集

: not overlapping

← very loose upperbound.

← # of hypothesis

- 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

$E_{in}(h) \simeq E_{out}(h)$

- '$E_{\text{in}}(g) = E_{\text{out}}(g)$' is **PAC**, **regardless of** $\mathcal{A}$

'most reasonable' $\mathcal{A}$ (like PLA): 选最小 Ein 的 h 作為 g.
pick the h_m with **lowest** $E_{\text{in}}(h_m)$ as g

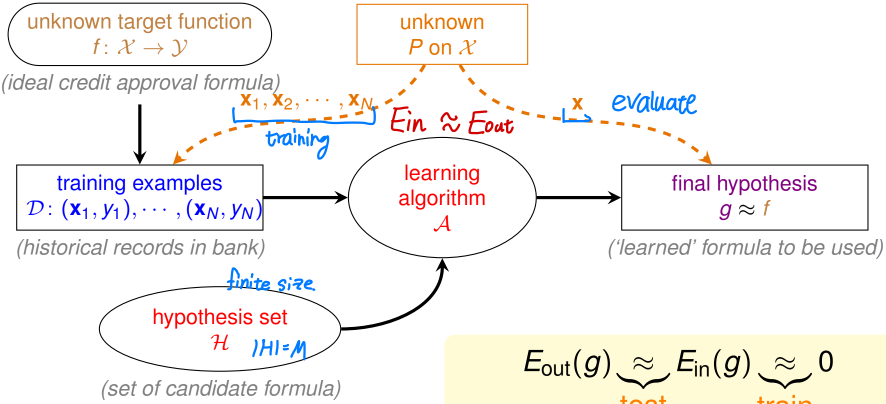
Questions?

The 'Statistical' Learning Flow

\# of Hypothesis   \# of data

if $|\mathcal{H}| = \boxed{M}$ finite, $\boxed{N}$ 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 $\underbrace{E_{\text{in}}(g) \approx 0}_{\text{lecture 1}}$ PLA

if M is finite

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) Can we extend from training → testing data
2. can we make $E_{\text{in}}(g)$ small enough? (train/optimize) learn well on training set!

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

Trade-off on M

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

\# of hypothesis

small M

1. Yes!, 沒有太多hypothesis 跟你screwed over
   $\mathbb{P}[\textbf{BAD}] \le 2 \cdot M \cdot \exp(\ldots)$
2. No!, too few choices

但很有可能学不起来

large M

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

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

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

M很重要，適中比較好

Preview

Known

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

把M换成一个有限的数量: $m_{\mathcal{H}}$

Todo

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

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

a finite quantity

- 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 $|\mathcal{H}| = \infty$?**