

Learning From Data

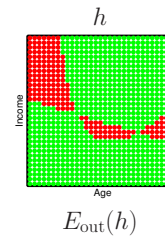
Lecture 4

Real Learning is Feasible

Real Learning vs. Verification
 The Two Step Solution to Learning
 Closer to Reality: Error and Noise

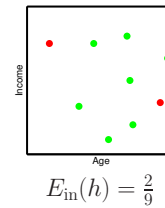
M. Magdon-Ismail
 CSCI 4100/6100

RECAP: Verification



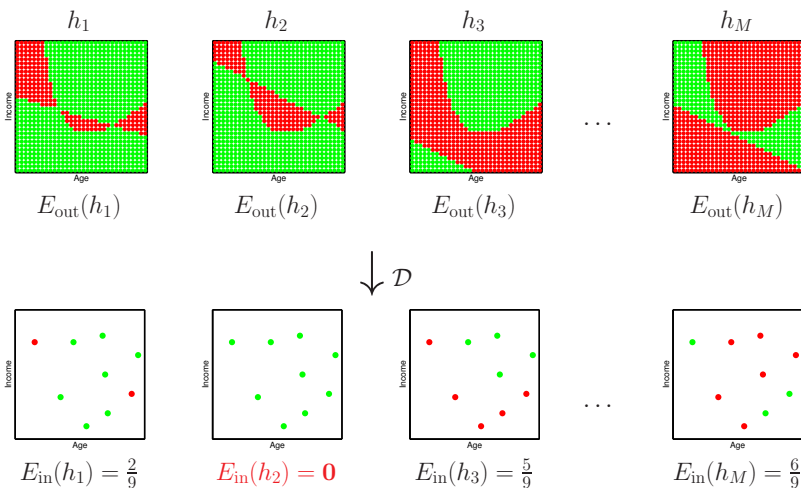
Hoeffding: $E_{out}(h) \approx E_{in}(h)$ (with high probability)

↓ \mathcal{D}



$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2N\epsilon^2}$$

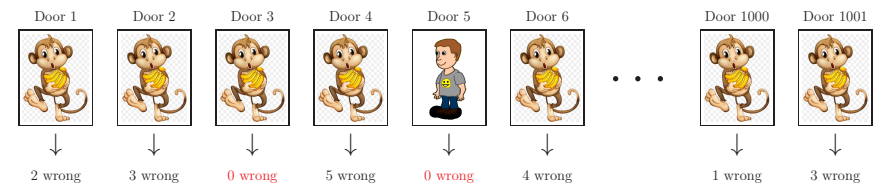
Real Learning – Finite Learning Models



Pick the hypothesis with minimum E_{in} ; will E_{out} be small?

RECAP: 1000 Monkeys Behind Closed Doors

5-question A/B test. Monkeys answer randomly. Child gets all right.



- What are your chances of picking the child?
- What can you do about it? (You can't peek behind the door. ☹)

More Monkeys: E_{in} Can't Reach Out to E_{out} .

Coin tossing example:

- If we toss one coin and get no HEADS, its very surprising. $\mathbb{P} = \frac{1}{2^N}$
We expect it is biased: $\mathbb{P}[\text{heads}] \approx 0$.
- Tossing 70 coins, and *find one* with no heads. Is it surprising? $\mathbb{P} = 1 - \left(1 - \frac{1}{2^N}\right)^{70}$
Do we expect $\mathbb{P}[\text{heads}] \approx 0$ for the selected coin?
Similar to the “birthday problem”: among 30 people, two will likely share the same birthday.
- This is called *selection bias*.
Selection bias is a very serious trap. For example medical screening.

Search Causes Selection Bias

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \text{for any } \epsilon > 0.$$

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| \leq \epsilon] \geq 1 - 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \text{for any } \epsilon > 0.$$

We don't care how g was obtained, *as long as it is from \mathcal{H}*

Some Basic Probability

Events A, B

Implication

If $A \implies B$ ($A \subseteq B$) then $\mathbb{P}[A] \leq \mathbb{P}[B]$.

Union Bound

$\mathbb{P}[A \text{ or } B] = \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$.

Bayes' Rule

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}$$

Proof: Let $M = |\mathcal{H}|$.

The event “ $|E_{in}(g) - E_{out}(g)| > \epsilon$ ” implies
“ $|E_{in}(h_1) - E_{out}(h_1)| > \epsilon$ ” OR ... OR “ $|E_{in}(h_M) - E_{out}(h_M)| > \epsilon$ ”

So, by the implication and union bounds:

$$\begin{aligned} \mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] &\leq \mathbb{P}\left[\bigcup_{m=1}^M |E_{in}(h_m) - E_{out}(h_m)| > \epsilon\right] \\ &\leq \sum_{m=1}^M \mathbb{P}[|E_{in}(h_m) - E_{out}(h_m)| > \epsilon], \\ &\leq 2Me^{-2\epsilon^2 N}. \end{aligned}$$

(The last inequality is because we can apply the Hoeffding bound to each summand)

Interpreting the Hoeffding Bound for Finite $|\mathcal{H}|$

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \text{for any } \epsilon > 0.$$

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| \leq \epsilon] \geq 1 - 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \text{for any } \epsilon > 0.$$

Theorem. With probability at least $1 - \delta$,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}.$$

We don't care how g was obtained, *as long as $g \in \mathcal{H}$*

Proof: Let $\delta = 2|\mathcal{H}|e^{-2\epsilon^2 N}$. Then

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| \leq \epsilon] \geq 1 - \delta.$$

In words, with probability at least $1 - \delta$,

$$|E_{in}(g) - E_{out}(g)| \leq \epsilon.$$

This implies

$$E_{out}(g) \leq E_{in}(g) + \epsilon.$$

From the definition of δ , solve for ϵ :

$$\epsilon = \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}.$$

E_{in} Reaches Outside to E_{out} when $|\mathcal{H}|$ is Small

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}.$$

If $N \gg \ln |\mathcal{H}|$, then $E_{out}(g) \approx E_{in}(g)$.

- Does not depend on \mathcal{X} , $P(\mathbf{x})$, f or how g is found.
- Only requires $P(\mathbf{x})$ to generate the data points independently *and also* the test point.

What about $E_{out} \approx 0$?

The 2 Step Approach to Getting $E_{\text{out}} \approx 0$:

- (1) $E_{\text{out}}(g) \approx E_{\text{in}}(g)$.
 - (2) $E_{\text{in}}(g) \approx 0$.

Together, these ensure $E_{\text{out}} \approx 0$.

How to verify (1) since we do not know E_{out}

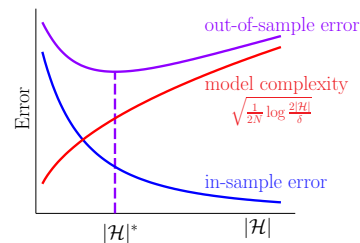
– must *ensure* it theoretically - Hoeffding.

We can ensure (2) (for example PLA)

– modulo that we can guarantee (1)

There is a tradeoff:

- Small $|\mathcal{H}| \implies E_{\text{in}} \approx E_{\text{out}}$
- Large $|\mathcal{H}| \implies E_{\text{in}} \approx 0$ is more likely.



Feasibility of Learning (Finite Models)

- No Free Lunch: can't know anything outside \mathcal{D} , *for sure*.
- Can “learn” with high probability if \mathcal{D} is *i.i.d.* from $P(\mathbf{x})$.
 $E_{\text{out}} \approx E_{\text{in}}$ (E_{in} can reach outside the data set to E_{out}).
- **We want** $E_{\text{out}} \approx 0$.
- The two step solution. We trade $E_{\text{out}} \approx 0$ for 2 goals:
 - (i) $E_{\text{out}} \approx E_{\text{in}}$;
 - (ii) $E_{\text{in}} \approx 0$.

We know E_{in} , not E_{out} , but we can *ensure* (i) if $|\mathcal{H}|$ is small.

This is a big step!

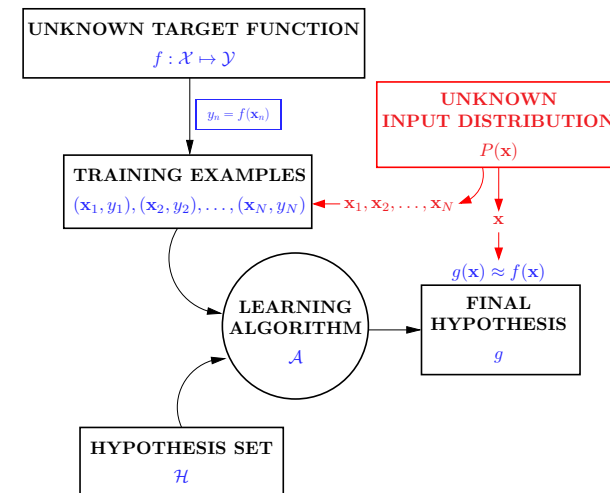
- What about infinite \mathcal{H} - the perceptron?

“Complex” Target Functions are Harder to Learn

What happened to the “difficulty” (complexity) of f ?

- Simple $f \implies$ can use small \mathcal{H} to get $E_{\text{in}} \approx 0$ (need smaller N).
- Complex $f \implies$ need large \mathcal{H} to get $E_{\text{in}} \approx 0$ (need larger N).

Revising the Learning Problem – Adding in Probability

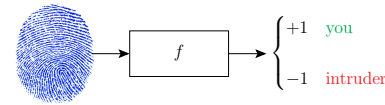


Error and Noise

Error Measure: How to quantify that $h \approx f$.

Noise: $y_n \neq f(\mathbf{x}_n)$.

Finger Print Recognition



Two types of error.

		f	
		+1	-1
h	+1	no error	false accept
	-1	false reject	no error

In any application you need to think about how to penalize each type of error.

		f	
		+1	-1
h	+1	0	1
	-1	10	0

Supermarket

		f	
		+1	-1
h	+1	0	1000
	-1	1	0

CIA

Take Away

Error measure is specified by the user.

If not, choose one that is
 - plausible (conceptually appealing)
 - friendly (practically appealing)

Almost All Error Measures are Pointwise

Compare h and f on individual points \mathbf{x} using a pointwise error $e(h(\mathbf{x}), f(\mathbf{x}))$:

Binary error: $e(h(\mathbf{x}), f(\mathbf{x})) = \mathbb{1}[h(\mathbf{x}) \neq f(\mathbf{x})]$ (classification)

Squared error: $e(h(\mathbf{x}), f(\mathbf{x})) = (h(\mathbf{x}) - f(\mathbf{x}))^2$ (regression)

In-sample error:

$$E_{\text{in}}(h) = \frac{1}{N} \sum_{n=1}^N e(h(\mathbf{x}_n), f(\mathbf{x}_n)).$$

Out-of-sample error:

$$E_{\text{out}}(h) = \mathbb{E}_{\mathbf{x}}[e(h(\mathbf{x}), f(\mathbf{x}))].$$

Noisy Targets

age	32 years
gender	male
salary	40,000
debt	26,000
years in job	1 year
years at home	3 years
...	...

Consider two customers with the **same** credit data.

They can have **different** behaviors.

Approve for credit?

The target 'function' is not a deterministic function but a *stochastic* function.

$$'f(\mathbf{x})' = P(y|\mathbf{x})$$

Learning Setup with Error Measure and Noisy Targets

