

Theory of Generalization

FACT

- with distribution in training data & testing data
- ⇒ low training error
- ⇕
- low testing error

Def

△ Training error:

$$- E_{tr}(h) = \frac{1}{N} \sum_{n=1}^N e(h(x_n), f(x_n))$$

where x_1, \dots, x_N sampled from D

- h is determined by x_1, \dots, x_N

△ Testing error:

$$- E_{te}(h) = \frac{1}{N} \sum_{n=1}^N e(h(x_n), f(x_n))$$

where x_1, \dots, x_N sampled from D

- h is independent from x_1, \dots, x_N

△ Generalization error

- G. error = Test error (on D) (expected performance)

$$- E(h) = E_{x \sim D} [e(h(x), f(x))] = E_{te}(h)$$

△ Summary

if $E(h) = 0$

then $E(h) \approx E_{tr}(h) \rightarrow$ How?

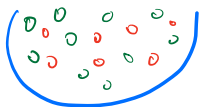
or

$E_{tr}(h) \approx 0 \rightarrow$ Training

Q: How do we make sure

$$E(h) \approx E_{tr}(h)$$

FACT Hoeffding's inequality



$$\Delta P[\text{pick red ball}] = \mu$$

$$P[\text{pick green ball}] = 1 - \mu$$

→ we DO NOT know μ

△ by pick ball's independently we get fraction of ν

△ $\nu \rightarrow \mu$?

perhaps

△ Hoeffding's inequality

$$P[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

anote: ν & μ 的差距, 比 ϵ 大的 P, 很小
多少? → 比 $2e^{-2\epsilon^2 N}$ 还小

△ statement $\mu = \nu$ is

probably approximately correct

(PAC!)

FACT

$$\Delta P[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- valid for N

- $\epsilon > 0$

- independent from μ
(real probability)



△ in learning:

- given a function h

- we randomly draw x_1, \dots, x_N

independent

- generalization error

$$E(h) = E_{x \sim D} [h(x) \neq f(x)] \Leftrightarrow \mu$$

unknown

sample data error

$$E_{tr}(h) = \frac{1}{N} \sum_{n=1}^N [h(x_n) \neq y_n] \Leftrightarrow \nu$$

known

$$P[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

FACT

△ for each h , h is a hypothesis

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

△ for all h , H is a hypothesis set

$$P[|E_{tr}(h_1) - E(h_1)| > \epsilon],$$

$$P[|E_{tr}(h_2) - E(h_2)| > \epsilon],$$

⋮

$$P[|E_{tr}(h_{|H|}) - E(h_{|H|})| > \epsilon]$$

$$\leq P[\sup_{h \in H} |E_{tr}(h) - E(h)| > \epsilon]$$

$$\leq \sum_{m=1}^{|H|} P[|E_{tr}(h_m) - E(h_m)| > \epsilon] \leq 2|H| e^{-2\epsilon^2 N}$$

$$\text{from } P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

△ summary

$$P[|E_{tr}(h) - E(h)| > \epsilon] \leq P[\sup_{h \in H} |E_{tr}(h) - E(h)| > \epsilon] \leq 2|H| e^{-2\epsilon^2 N}$$

NOTE More on Hoeffding's inequality

h_1	D_1	D_2	...	D_N	$P[BAD D \text{ for } h_1] \leq \dots$
h_2	Bad	Bad			$P[BAD D \text{ for } h_2] \leq \dots$
\vdots					
h_M	Bad			Bad	$P[BAD D \text{ for } h_M] \leq \dots$

informed hypothesis (假设现在我知道的模型 i.e., 不是学出来的) \rightarrow 对我手上的资料 D_1, D_2, \dots, D_N 的 h on D 可能导致 "Bad" 亦即 $E_{in}(h) \neq E_{out}(h)$ \rightarrow 之所以要 Hoeffding's ineq. 就是为 \sum 量化 $P[BAD]$ 机率多高?

\rightarrow 答案是很低: bounded by $2e^{-2\epsilon^2 N}$ \rightarrow 不过刚刚是 1 个 h 啊... 我 algo 都是在 $1/\epsilon$ 找 h ... upper bound 是啥?

$$\begin{aligned} \therefore P_0[BAD D] &= P[BAD D \text{ for } h_1 \text{ or } BAD D \text{ for } h_2 \dots \text{ or } BAD D \text{ for } h_M] \\ &\leq P[BAD D \text{ for } h_1] + P[BAD D \text{ for } h_2] + \dots + P[BAD D \text{ for } h_M] \\ &\quad \text{(union bound)} \\ &\leq 2M e^{-2\epsilon^2 N} = 2|H| e^{-2\epsilon^2 N} \end{aligned}$$

- finite-bin version of Hoeffding
- & hope... $E_{in}(g) = E_{out}(g)$ is PAC.
- \rightarrow A will pick h_m w/ min. $E_{in}(h_m)$ as g

FACT establish a finite quantity replace $|H|$

let $|H|$ replaced by m_H

s.t.

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

FACT $|H|$ is over-estimated for BAD events

- BAD events $B_m: |E_{in}(h_m) - E_{out}(h_m)| > \epsilon$
- over-lapping for similar hypothesis $h_1 \approx h_2$
- as ① $E_{out}(h_1) \approx E_{out}(h_2)$

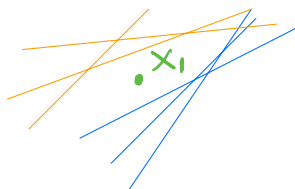
② for most D $E_{in}(h_1) \approx E_{out}(h_2)$

- should be instead of

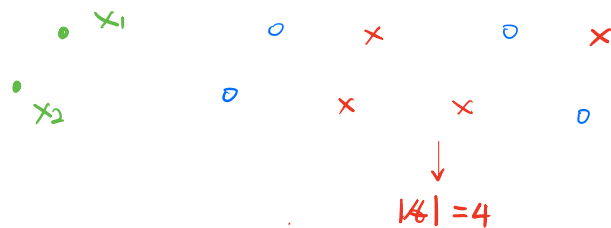


- so: can we group similar kinds?

eg. H in R^2 $|H| \rightarrow \infty$



$\rightarrow |H| = 2$



$N=3$ $|H|=8$ but if on same line, different
 $N=4$ $|H|=14$ but if on same line, different

FACT observation: effective $|H| \leq 2N$

perhaps can replace $|H|$ by effective $|H|$?
 need more rigorous proof

FACT Dichotomies: mini-hypotheses

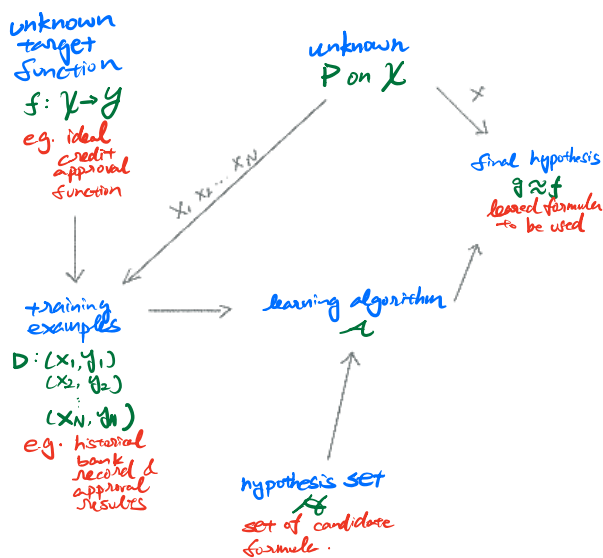
- limited hypothesis: $H(x_1, x_2, \dots, x_N)$
- $|H(x_1, x_2, \dots, x_N)|$: depend on inputs (x_1, x_2, \dots, x_N)
- growth function: remove dependence by taking max of all possible (x_1, x_2, \dots, x_N)

$$m_H(N) = \max_{x_1, x_2, \dots, x_N \in X} |H(x_1, x_2, \dots, x_N)|$$

Δ finite, upper-bounded by 2^N

Q: How to calculate growth function

Summary: statistical learning flow



& hope $E_{out}(g) \approx E_{in}(g) \approx 0$
 test train

- for batch & supervised learning, $g \approx f \Leftrightarrow E_{out}(g) \approx 0$ achieved through $E_{out}(g) \approx E_{in}(g)$ & $E_{in}(g) \approx 0$
- ① can we make sure $E_{out}(g) \approx E_{in}(g)$
- ② can we make $E_{in}(g)$ small enough

FACT $|H| = \infty$

Q: Question: How do we deal with it?

- small $|H|$ $P[BAD] \leq 2|H| e^{-2\epsilon^2 N}$ $E_{in}(g) \rightarrow 0$
 small! great!
 but $|H|$ too little $E_{in}(g) \uparrow$
- large $|H|$ $E_{in}(g) \rightarrow 0$
 small error! great!
 but $|H|$ too large $P[BAD] \uparrow$

FACT Shattered

Δ if $m_H(N) = 2^N \iff$ exists N inputs that can be shattered

Δ eg. convex set

FACT Summary of 4 growth functions

- positive rays $N+1$
- positive intervals $\binom{N+1}{2} + 1$
- convex sets 2^N
- 2D perceptrons $< 2^N$

polynomial good!

exponential bad!

FACT Break point \leadsto k 開始, 無法被 shattered

Δ if no k inputs can be shattered by H
call k a break point for H

$\Delta m_H(k) < 2^k$

$\Delta k+1, k+2, k+3 \dots$ are all break points

Δ study minimum break point

eg. linear case break point $k=4$

note: 4 个 無法被 shattered

FACT conjecture:

Δ no break point: $m_H(N) = 2^N$

Δ break point k : $m_H(N) = O(N^{k-1})$

proof?

FACT $m_H(N) \leq$ maximum possible $m_H(N)$ given k
 $\leq \text{poly}(N)$

FACT Bounding function