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

△ summary

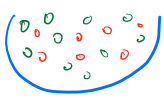
if $E(h) = 0$

then $E(h) \approx E_{tr}(h)$ ⟶ How?

and $E_{tr}(h) \approx 0$ ⟶ Training

Q: How do we make sure $E(h) \approx E_{tr}(h)$

FACT Hoeffding's inequality



△ $P[\text{pick red ball}] = u$
$P[\text{pick green ball}] = 1-u$
⟶ we DO NOT know u

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

△ $v \longrightarrow u$?
perhaps

△ Hoeffding's inequality
$P[|v-u| > \epsilon] \le 2e^{-2\epsilon^2 N}$
annote: v & u 的差距, 比 ϵ 大的 P, 很小多小? ⟶ 比 $2e^{-2\epsilon^2 N}$ 還小

△ statement $u = v$ is probably approximately correct (PAC !)

FACT

△ $P[|v-u| > \epsilon] \le 2e^{-2\epsilon^2 N}$
- valid for N
- $\epsilon > 0$
- independent from u (real probability)

↓

△ in learning :
- given a function h
- we randomly draw $x_1, \dots, x_n$] independent
- generalization error
  $E(h) = E_{x\sim D}[h(x) \neq f(x)] \Leftrightarrow u$ unknown
  sample data error
  $E_{tr}(h) = \frac{1}{N}\sum_{n=1}^{N}[h(x_n) \neq y_n] \Leftrightarrow v$ known

$P[|v-u|>\epsilon] \le 2e^{-2\epsilon^2 N}$

FACT

△ for each h, h is a hypothesis
$P[|E_{tr}(h) - E(h)| > \epsilon] \le 2e^{-2\epsilon^2 N}$

△ for all h, $\mathcal{H}$ is a hypothesis set

$P[|E_{tr}(h_1) - E(h_1)| > \epsilon]$,
$P[|E_{tr}(h_2) - E(h_2)| > \epsilon]$,
⋮
$P[|E_{tr}(h_{|\mathcal{H}|}) - E(h_{|\mathcal{H}|})| > \epsilon]$

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

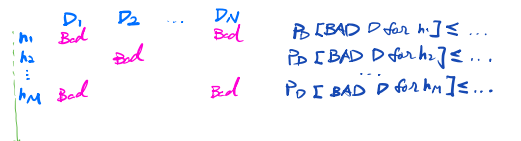
$\le \sum_{m=1}^{|\mathcal{H}|} P[|E_{tr}(h_m) - E(h_m)|] \le 2|\mathcal{H}|e^{-2\epsilon^2 N}$

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

△ summary

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

NOTE More on Hoeffding's inequality



inferred hypothesis (假设现在我讨论的模型 i.e., 不是算出来的) → 对看到我手上的资料 D1, D2 ... DN, inferred的 h on D 可能导致 "Bad" 亦即 Ein(h) ≠ Eout(h) → 之所以要 Hoeffding's ineq. 就是为了量化 P[Bad] 机率多高?

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

∴ $P_D[\text{BAD } D]$

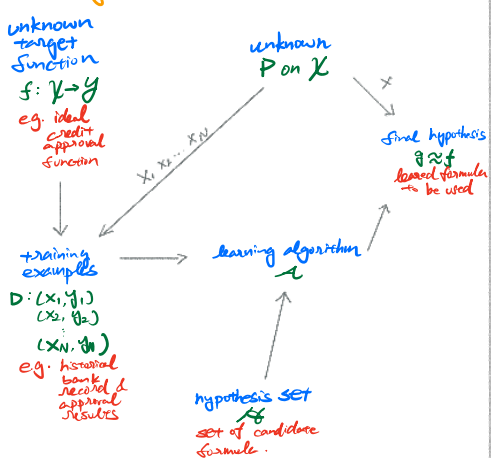
$= P_D[\text{BAD } D \text{ for } h_1 \text{ or BAD } D \text{ for } h_2 \cdots \text{or BAD } D \text{ for } h_M]$

$\leq P_D[\text{BAD } D \text{ for } h_1] + P_D[\text{BAD } D \text{ for } h_2] + \cdots + P_D[\text{BAD } D \text{ for } h_M]$ (union bound)

$\leq 2M e^{-2\epsilon^2 N} = 2|\mathcal{H}| e^{-2\epsilon^2 N}$

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

△ 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 $|\mathcal{H}| = \infty$

△ Question: How do we deal with it?

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

FACT establish a finite quantity replace $|\mathcal{H}|$

for $|\mathcal{H}|$ replaced by $m_{\mathcal{H}}$

s.t.

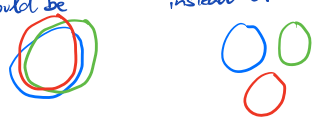
$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 m_{\mathcal{H}} e^{-2\epsilon^2 N}$$

FACT $|\mathcal{H}|$ 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 $E_{out}(h_1) \approx E_{out}(h_2)$

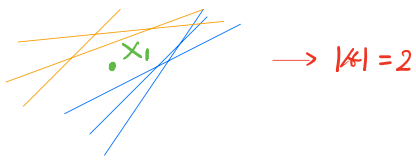
② for most D $E_{in}(h_1) \approx E_{out}(h_2)$

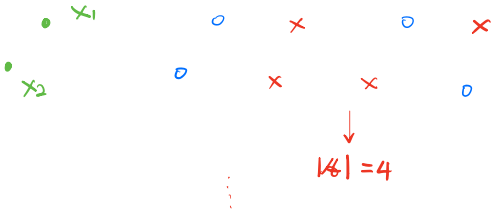
- should be instead of



- so: can we group similar kinds?

e.g. $\mathcal{H}$ in $\mathbb{R}^2$ $|\mathcal{H}| \to \infty$





N = 3	$\lvert\mathcal{H}\rvert = 8$	but if on same line, different
N = 4	$\lvert\mathcal{H}\rvert = 14$	but if on same line, different

FACT observation: effective $|\mathcal{H}| \leq 2^N$

perhaps can replace $|\mathcal{H}|$ by effective $|\mathcal{H}|$?

need more rigorous proof

FACT Dichotomies : mini-hypotheses

△ limited hypothesis : $\mathcal{H}(x_1, x_2 .. x_N)$

△ $|\mathcal{H}(x_1, x_2, \ldots, x_N)|$: depend on inputs $(x_1, x_2 .., x_N)$

△ growth function:

remove dependence by taking max of all possible $(x_1, x_2 \ldots, x_N)$

$$m_{\mathcal{H}}(N) = \max_{x_1, x_2 \ldots x_N \in X} |H(x_1, x_2 .. x_N)|$$

△ finite, upper-bounded by 2^N

Q: How to calculate growth function

FACT shattered

△ if $m_{\mathcal{H}}(N) = 2^N$ ⇔ exists N inputs that can be shattered

△ eg. convex set

FACT summary of 4 growth function

- positive rays $N+1$
- positive intervals $C_2^{N+1}+1$
- convex sets 2^N
- 2D perceptrons $< 2^N$

polynomial good !

exponential bad !

FACT Break point ⇝ k 開始，無法被 shattered

△ if no k inputs can be shattered by $\mathcal{H}$, call k a break point for $\mathcal{H}$

△ $m_{\mathcal{H}}(k) < 2^k$

△ $k+1$, $k+2$, $k+3$ ... are all break points

△ study minimum break point

eg linear case break point $k=4$

note: 4 ↑ 無法被 shattered

FACT conjecture:

△ no break point : $m_H(N) = 2^N$

△ break point k : $m_{\mathcal{H}}(N) = O(N^{k-1})$

proof ?

FACT $m_{\mathcal{H}}(N) \le$ maximum possible $m_{\mathcal{H}}(N)$ given k

$\le$ poly (N)

FACT Bounding function