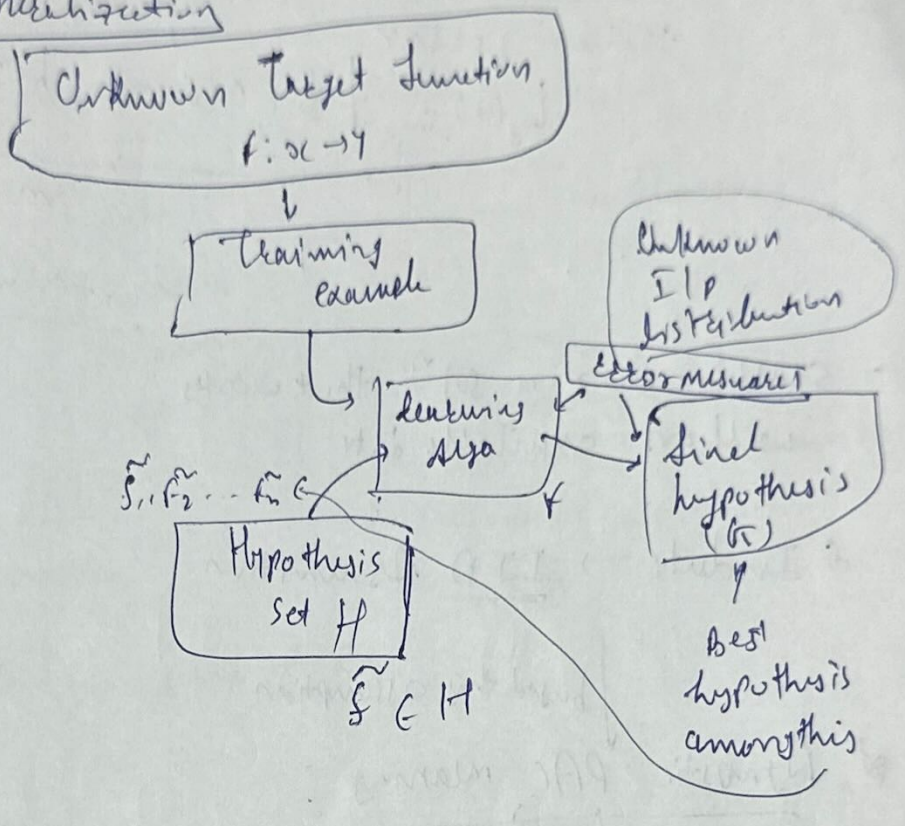
Theory of generalization



A statistical learning framework :-

→ Domain - x , Lable - y

- training data

- Data Generation Model :-

Prob. dist^n over X is D.

lable fun^n f: x → y

- measures of Success.

$L_{D,f}(h) = \underset{x \sim D}{P}[h(x) \neq f(x)]$

D, f → dist^n ; h → hypothesis

f(x) means 'y'

this is error / miss classification

* ERM Framework (Empirical ~~risk~~ Risk minimization)

training error

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

↳ total no. of misclassify points / total instances

- search for a soln that works well on available data

* Induty → IID Assumption

↓ based this assumption

* Agnostic PAC meaning :

↳ Probably Approximately correct

→ H is PAC learnable if

two conditⁿ should be follow

- $m_H : (0,1)^2 \to N$
- $\epsilon, \delta \in (0,1)$, D over $X \times Y$

(ϵ ← error, δ ← failure Prob.)

→ $E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} [h(x_n) \neq f(x_n)]$

training error expectation

$E_{out}(h) = P[h(x) \neq f(x)]$ (out of sample error)

kind of testing error expectation ; (unknown to us)

→ for fixed hypo. h

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq \delta$$

Prob. of failure

(8) Hoeffding's Inequality :

$$P\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - u\right| > \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{(a-b)^2}\right)$$

$\frac{1}{m}\sum_{i=1}^{m}\theta_i$ → Sample mean

u → mean of IID R.V. original mean

ϵ → error (no. of samples)

this quantity

we want to this quantity should be small.

$= \delta$

θ_i → samples (R.V.)

m → no. of samples

a, b → bounds for prob of R.V.

$$\therefore \delta \geq 2\exp\left(\frac{-2m\epsilon^2}{(a-b)^2}\right)$$

$$\therefore \frac{1}{2}\log\delta \geq \frac{-2m\epsilon^2}{(a-b)^2}$$

$$\therefore m\epsilon^2 \geq \frac{1}{2}\log 1/\delta \cdot (a-b)^2$$

$$\therefore m \geq \frac{1}{2\epsilon^2}\log\frac{1}{\delta}\cdot(a-b)^2$$

we want δ should be less so. '$\geq$'

$$\therefore m \geq \frac{1}{2\epsilon^2}\log\frac{1}{\delta}\cdot(a-b)^2$$

$$\therefore m \geq O\left(\frac{1}{2\epsilon^2}\log\frac{1}{\delta}\right)$$

↳ means ϵ, δ are const. so. we need increase successful prob. than m should be less increase.

↳ book => learning from data by abu Ismail.

$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N} 1(h(x_i) \neq f(x_i))$

↓ training error ($\in$ RM)

$E_{out}(h) = P[h(x) \neq f(x)]$

* Marble experiments:

~~A bin with red & green marble~~

~~Pick~~ 1

$\nu = \mu$

$\mu = m\mu$

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

$E_{in}(g) - E_{out}(g) > \epsilon$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

generalization bound

$\therefore \delta \geq 2Me^{-2m\epsilon^2}$

$\therefore \frac{\delta}{2M} \geq e^{-2m\epsilon^2}$

($\because m$ = no of hypothesis)

($\because N$ = no of samples)

$\therefore \ln \frac{\delta}{2M} \geq -2m\epsilon^2$

$\frac{1}{2m} \ln \frac{\delta}{2M} \leq \epsilon^2$

$\sqrt{\frac{1}{2N} \ln \frac{\delta}{2M}} \leq \epsilon$

$\ln \frac{2M}{\delta} \leq 2m\epsilon^2$

$$\epsilon \geq \sqrt{\frac{1}{2m} \ln \frac{2M}{\delta}}$$

* what if $M = \infty$?! atMost always the case ?!!!

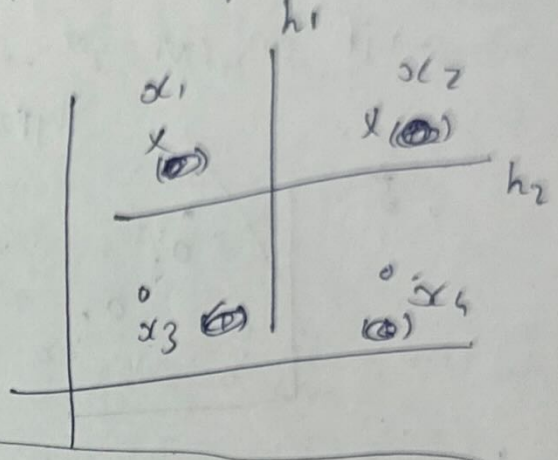
→ So we need to represent the infinite No. to finite no.

* Def A : for $x_1, x_2 \dots x_N \in N$ the dichotomies generated by H on these points is defined by

$$H(x_1 \dots x_N) = \{ h(x_1), \dots h(x_N) \mid h \in H \}$$

* Def : if $H = \{h_1, h_2\}$

$h_1(x_1, x_2, x_3, x_4)$
$= (0, 0, 1, 1)$
$= (1, 0, 1, 0)$
$h_2 : (0, 0, 1, 1)$



$H = \{(1, 0, 1, 0), (0, 0, 1, 1)\}$

this is finite set because the dichotomies max^n we can create dichotomies is 2^N in this case only

* def B : the growth function for a hypothesis set H is

(this fun^n is real no.) $m_H(N) = \max_{x_1, \dots x_N} |H(x_1, x_2 \dots x_N)|$

R^+

$m_H(N) \leq 2^N$

Hypothesis, size of data

⊙ if $m_H(N) = 2^N$, H shatters the points $x_1 \dots x_N$

Ex 3 points $x_1\ x_2\ x_3$

$H_1 = \{h_1\ h_2 h_3\}$

$H_2 = \{h_1 h_2 h_3\ h_4\}$

$H_3 = \{h_1 h_2 h_3 h_4\ h_5$ ↑ $h_6, h_7, h_8\}$

cannot shatter

size it cannot all classify the

it shatters the b/c the it can classify the all points.

y_1	y_2	y_3
0	0	0
.	.	.
.	.	.
.	.	.
1	1	1

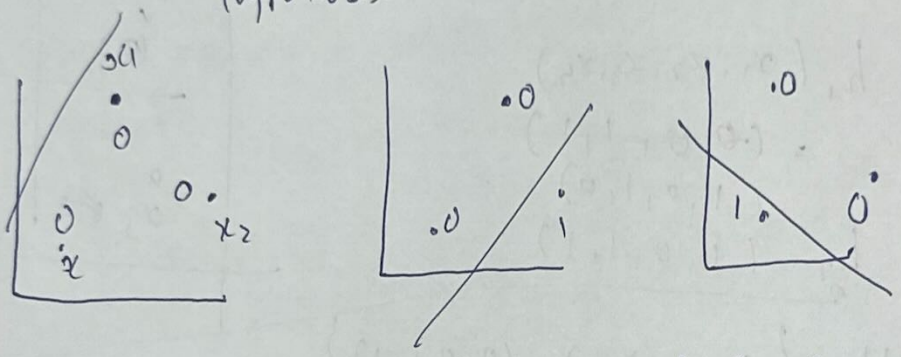
8

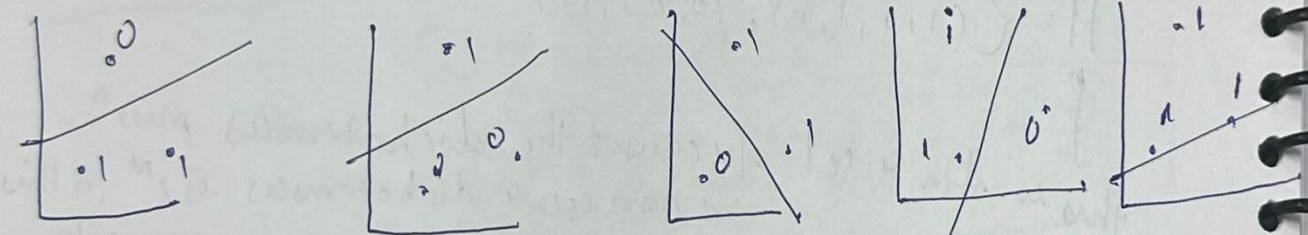
(*) VC dimension: VC-dimension of a hypo. set H denoted as $d_{VC}(H)$ or d_{VC} is largest value of N for which growth fun $m_H(N) = 2^N$.

(-) if $m_H(N) = 2^N$ for $\forall N$, $\rightarrow d_{VC}(H) = \infty$

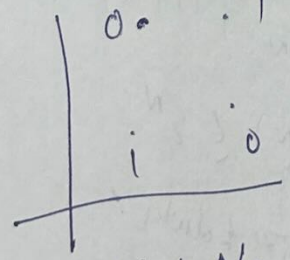
→ if $m_H(N) = 2^N$ → then N is VC dim of that Hypothesis.

Ex





↳ we can classify all point so. $\forall_c d_{VC} = \infty$ at least 3.

→ 

$\underline{X}\ d_{VC} = 4$ not possible

→ for 2D point VC dimension is 3.

(*) If VC-dimension (H) = n then H shatters every set of size n & there exists a set of size (n+1) which can not shatter. (Radon's theorems)

↳ Point are d & d dim. then we can shatters d+1 dimension.

→ Sine waves → VC dim (H) = ∞

- NN → comb VC dim (H) = 0

* Important Results (Sauer-Shelah lemma theorem)

Theorem 1: If $m_H(k) < 2^k$ for some value k then

$$m_H(k) \leq \sum_{i=1}^{k-1} \binom{N}{i}$$

k → can't shatter b'coz of strictly less than

$$\text{Ex: } k=4,\ N=3$$

$$m_H(3) \leq \binom{3}{3} + \binom{3}{1} + \binom{3}{2} + \binom{3}{3}$$

* by definition VC dimension

$$m_H(N) \leq \sum_{i=0}^{d_{vc}} \binom{N}{i}$$

$$\text{Ex: } 5 \text{ Points} - 2\text{dimen. } d_{vc} = 2+1 = 3$$

$$m_H(5) \leq \binom{5}{0} + \binom{5}{1} + \binom{5}{2} + \binom{5}{3}$$

→ If $d_{vc}(H) = n$ ∃ a set of size n which H can shatter and H cannot shatter any set of size (n+1)

* VC-generalization: VC-generalization bound for any tolerance $\delta > 0$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{\delta}{N} \ln \frac{4 m_H(2N)}{\delta}} \to 2^{d_{vc}}$$

w.p. $1-\delta$ with probability.

↓ Note this is finite then we can apply Union bound.

→ if d_{vc} is increased then more Complex the model, More Complex hypothesis space than our model is Overfit. test error is very high.