Is learning feasible

The ultimate goal is $g \approx f$

what does it mean?

- It means $E_{out}(g) \approx 0$

But we cannot know $E_{out}(g)$ during learning. So how to ensure $E_{out}(g) \approx 0$ or $g \approx f$ during learning.

$E_{out}(g)$ can be achieved through

$$\boxed{E_{out}(g) \approx E_{in}(g) \quad \& \quad E_{in}(g) \approx 0}$$

Thus learning is nothing but

- To make sure $E_{out}(g)$ is as close to $E_{in}(g)$ as possible — Hoeffding's inequality
- To try to make $E_{in}(g)$ as small as possible

Hoeffding's Inequality

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

ϵ is the tolerance.

But we are interested in g not h, So for g it becomes,

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

Why? because we are more concerned about the upper bound g is the best amongst all the hypothesis and in probability notation it means

$$P\left[|E_{out}(g) - E_{in}(g)| > \epsilon \right] \leq P\left[|E_{out}(h_1) - E_{in}(h_1)| > \epsilon \right] \text{ or}$$

$$P\left[|E_{out}(h_2) - E_{in}(h_2)| > \epsilon \right] \text{ or}$$

$$\vdots$$

$$P\left[|E_{out}(h_M) - E_{in}(h_M)| > \epsilon \right]$$

$$\leq \sum_{m=1}^{M} 2e^{-2\epsilon^2 N}$$

$$\boxed{\leq 2Me^{-2\epsilon^2 N}}$$

But M is infinite as we know that there are infinite number of hypothesis.

So what does that mean? It means

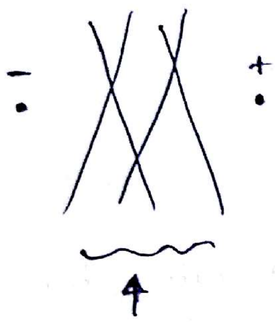
$|E_{out}(g) - E_{in}(g)| \leq \epsilon$ is not possible as M is infinite.

This again means we cannot generate $E_{out}(g) \approx E_{in}(g)$

So what to do now??

Duplicate hypothesis

Two hypothesis are said to be duplicate if they result in the same MSE.



So for a given problem in hand and a hypothesis set we can ensure that their will be lot of duplicate hypothesis.
This gives us some breathe because we can now replace M with something smaller.

Dichotomies

It is called mini hypothesis

A hypothesis, $h : X \rightarrow \{+1, -1\}$

A dichotomy $h : \{x_1, x_2, x_3, \ldots, x_N\} \rightarrow \{+1, -1\}$

Number of hypothesis $|H|$ is infinite

Number of dichotomies $|H(x_1 \ldots x_N)|$ is at most 2^N

& so is a candidate for replacing M

Now we could replace M (infinity) to 2^N (exponential)

Can we do any better ?

Growth Function

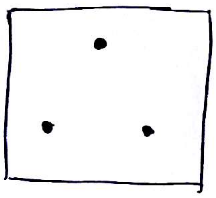
The growth function counts the most dichotomies on any N points

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

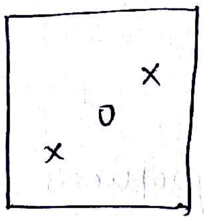
The growth function satisfies

$$m_H(N) \le 2^N$$

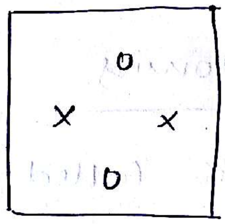
So lets count the number of dichotomies for a binary classifier



$N = 3$



$N = 3$

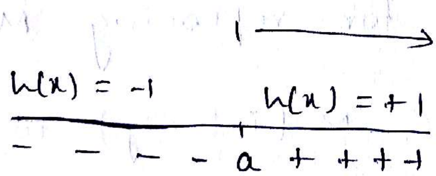


$N = 4$

$m_H(3) = 8$ $\quad$ $m_H(4) = 14$

Let us take other examples

(i) positive rays

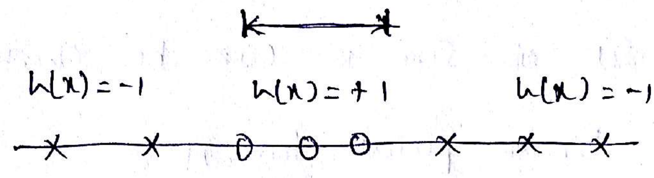


$$h(x) = \text{sign}(x - a)$$

$$m_H(N) = N + 1$$

Scanned by CamScanner

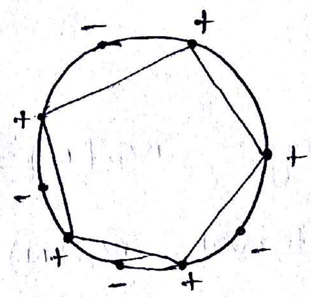
(ii) positive intervals



$$H : \mathbb{R} \rightarrow \{+1, -1\}$$

$$m_H(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

(iii) convex sets



$$H : R^2 \rightarrow \{-1, +1\}$$

$$m_H(N) = 2^N$$

All 'N' points are shattered by convex sets

What do we learn from the above three models?

We learn that the $m_H(N)$ can be polynomial.

This is the transition so far

from $\infty \rightarrow 2^N \rightarrow$ polynomial

But wait, convex set is still 2^N.

So how to know if a model is really polynomial.

Break points

If no data set of size k can be shattered by H, then k is a break point for H

$$\boxed{m_H(k) < 2^k}$$

For binary classifier, $k=4$

Examples

(i) positive rays, $m_H(N) = N+1$, $k=2$

(ii) positive interval, $m_H(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$, $k=3$

(iii) convex sets, $m_H(N) = 2^N$, $k = \infty$

This means,

No break point $\Rightarrow$ $m_H(N) = 2^N$

Any break point $\Rightarrow$ $m_H(N)$ is polynomial in N

Final equation

$$P\left[\,|E_{out}(g) - E_{in}(g)| > \epsilon\,\right] \le 4\, m_H(2N)\, e^{-\frac{1}{8}\epsilon^2 N}$$

It is called Vapnik - chervonenkis Inequality

To summarize,

(i) Bigger N, it is better

(ii) less complex hypothesis set.

So this ends the first part of learning which talks about $E_{out} \approx E_{in}$. Now lets focus on $E_{in} \approx 0$

The VC dimension

- The VC dimension $d_{vc}(H)$ is the largest value of N for which $m_H(N) = 2^N$

"The most points H can shatter"

growth function bound

$$m_H(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}, \quad \text{where } k \text{ is the break point.}$$

So in terms of VC dimension it is

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

$$m_H(N) \leq N^{d_{vc}}$$

As an example lets take $N=4$ & $d_{vc}=3$, $K=4$;

$$m_H(4) = \sum_{i=1}^{3} \binom{4}{i} = \binom{4}{1} + \binom{4}{2} + \binom{4}{3}$$

$$= 4 + 6 + 4 = \underline{\underline{14}}$$

$\therefore$ $m_H(4) = 14$ for binary classifier

Its always the case that $\boxed{d_{vc} = d+1}$ where,

d is the dimension or the parameters $w_0, w_1, \dots w_d$

$\boxed{d_{vc} = \text{dim} + 1}$ can be proved but we are not going into the proof.

VC dimension is also called as the capacity of the model. It is also referred to as "degrees of freedom"

Capacity ⇒ Memory. more the capacity, more the memory and hence it can learn more as it can store more information about the data.

It is the effective number of parameters.

If in doubt and do not know what is $m_H(N)$ for the model that you select, just use $N^{d_{vc}}$

Now coming to the number of data points needed to train the model as a rule of thumb would be

$$\boxed{N \geq 10 d_{vc}}$$

How does vc dimension relate to $E_{in} \approx 0$?

The higher the vc dimension, the lower E_{in} would be.

Generalization

The ability to perform well on previously unobserved inputs is called generalization.

Generalization bound

$$P\left[|E_{out} - E_{in}| > \epsilon \right] \leq \underbrace{4 m_H(2N) e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

$$\epsilon = \underbrace{\sqrt{\frac{8}{N} \ln \frac{4 m_H(2N)}{\delta}}}_{\Omega}$$

with probability $\geq 1-\delta$ $|E_{out} - E_{in}| \leq \epsilon$
$\leq \Omega$

$[E_{out} \leq E_{in} + \Omega]$ ← Generalization bound