Probably Approximately Correct Learning (PAC Learning) →

→ A good learner will learn with high probability and close apporoximation to the target concept.

→ With High probability, the selected hypothesis will have lower the error ("Approximately correct") with the parameters ε and δ.

ε (epsilon) δ (delta)

→ PAC learning requires:

- → small parameters ε and δ
- → with probability atleast $(1-\delta)$, a system learn the concept concept at most ε.

• ε is upper bound on error, in accuracy, i.e. the hypothesis with error less than ε.

Accuracy $= 1-\varepsilon$

- δ gives the probability of failure in achieving this accuracy δ, $(0 < \delta \leq 1)$, the hypothesis generated is approximately correct at least $(1-\delta)$ of the time.

  Confidence $= 1-\delta$

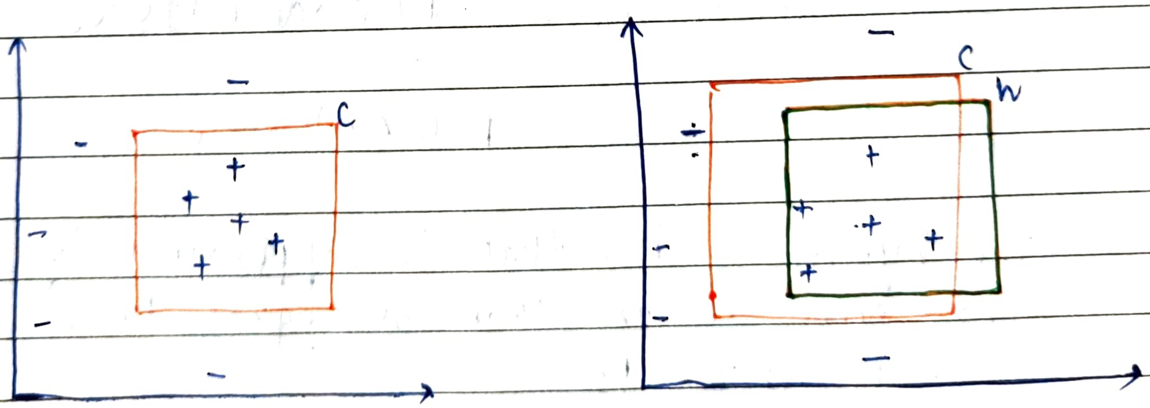
PAC learning Example →

→ N number of car having Price & Engine Power (p, e) as training set → find the car is family car or not.

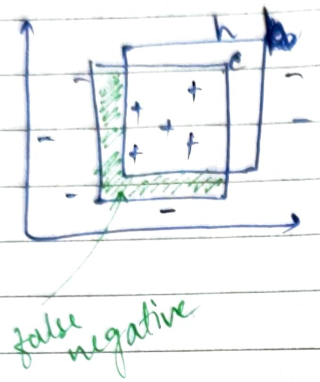
→ C → Target function

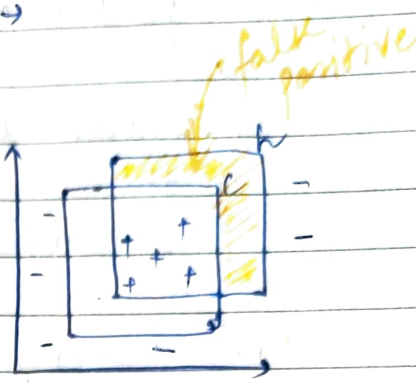
→ Instances within the rectangle represents → family car
  outside → not family car.

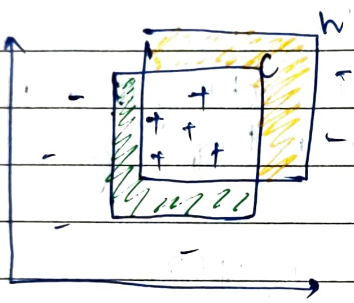
→ Hypothesis h → closely approximate C, and there may be error region.



false negative & false positive →







Error Region →

$c \,\Delta\, h$

Δ = Region of difference b/w c & h

Prob. of Error Region should be very small.

Error Region : $P(c \,\Delta\, h) <= \varepsilon$

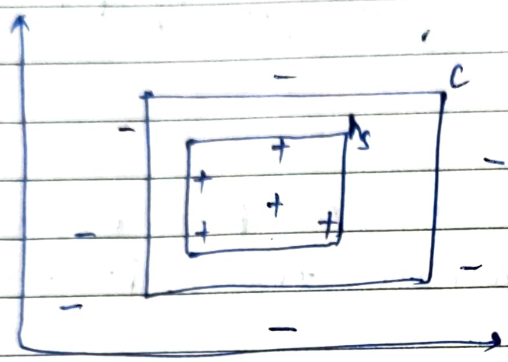
The hypothesis h, that is approximately correct, and error is less than or equal to ε, where $0 \le \varepsilon \le 1/2$

So, $P(c \,\Delta\, h) \le \varepsilon$

low generalization error with high probability -
i.e.

$$P(P(C \Delta h) <= \epsilon) \geq 1-\delta$$

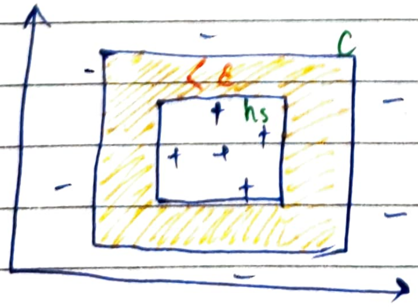
PAC learnability for axis-aligned rectangle:-



h_s is the tightest possible rectangle around a set of positive examples.

h_s is subset of C, hence
Error Region = $C - h_s$

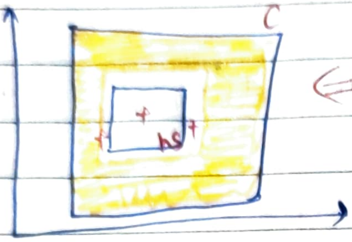
Approximately correct →



If an hypothesis lies between h_s and C (shaded Region) then it is approximately correct.

→ If the generated hypothesis does not touch any of these region ⇒ next page

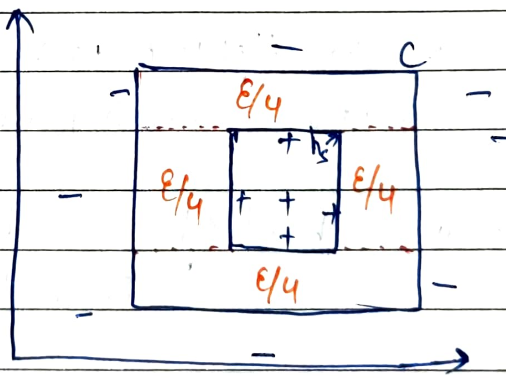
⇒ then Error Region is greater than ε and not approximately correct, because the error region got increased.



⇐ Error region got increased here, error region > ε so NOT approximately correct.

** ⇒ If there is atleast one positive example at each side of rectangle than it will be correct h_s.

Error Region :-



- Error Region = sum of four rectangular strips < ε
- Each strip is atmost ε/4.

- Prob. of positive example falling in anyone of the strip ⇒ (error region = ε/4)

- Prob. that a randomly drawn positive example misses a strip $= 1 - \epsilon/4$
- $P(m \text{ instances miss a strip}) = (1-\epsilon/4)^m$

$P(m \text{ instances miss any strip}) < 4\left(1-\frac{\epsilon}{4}\right)^m$

- and we want this to be almost δ.

so, $$4\left(1-\frac{\epsilon}{4}\right)^m \leq \delta$$

Dividing both the sides by 4, taking (natural) log and rearranging terms, we get

$$m \geq \frac{4}{\epsilon} \log(4/\delta).$$

So, provided that we take atleast $(4/\epsilon)\log(4/\delta)$ independent examples from c and use the tightest rectangle as our hypothesis h, with confidence prob. atleast $1-\delta$, a given point will be misclassified with error prob. at most ϵ.

Example:

①

Sr. No	Error(h_1)
1	0.001
2	0.025
3	0.07 ✓
4	0.003
5	0.035
6	0.045
7	0.027
8	0.065 ✓
9	0.012
10	0.036

Hypothesis h_1 generated the errors with respect to price and engine power of given 10 samples.

Given $\varepsilon = 0.05$, $\delta = 0.20$

$P(h_1) >= 1-\delta$

$P(h_1) = \frac{8}{10} = 0.80$ {3rd & 8th value are greater than ε}

$1-\delta = 1-0.20 = 0.80$

$\therefore 0.80 \geq 0.80$ (True)

Hence h_1 is probably approximately correct.

② $\Leftarrow$ Errors generated by hypothesis h_2.

Sr. No	Error(h_2)
1	0.012
2	0.015
3	0.071 ✓
4	0.063 ✓
5	0.022
6	0.045
7	0.011
8	0.029
9	0.066 ✓
10	0.031

Given $\varepsilon = 0.05$, $\delta = 0.20$

$P(h_2) \geq 1-\delta$

$P(h_2) = 7/10 = 0.7$

$1-\delta = 1-0.20 = 0.8$

$0.7 \geq 0.8$ False

$\therefore h_2$ is <u>not</u> probably approximately correct.