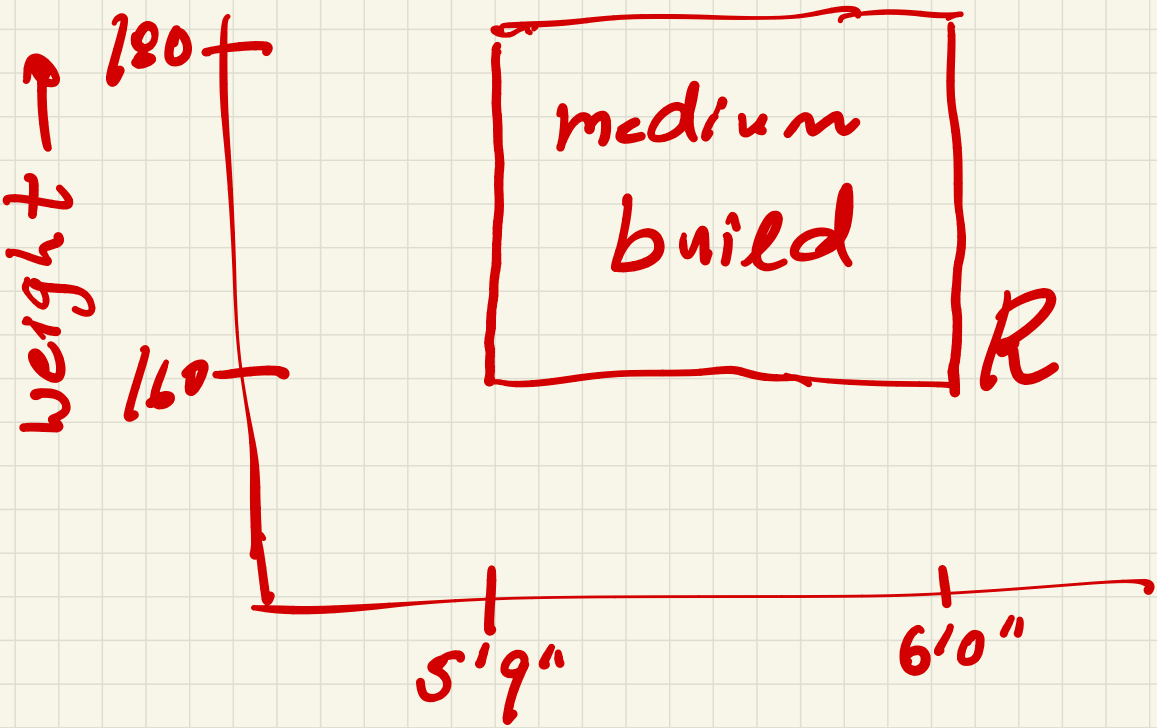
A "Toy" ML Problem: Rectangles in $\mathbb{R}^2$

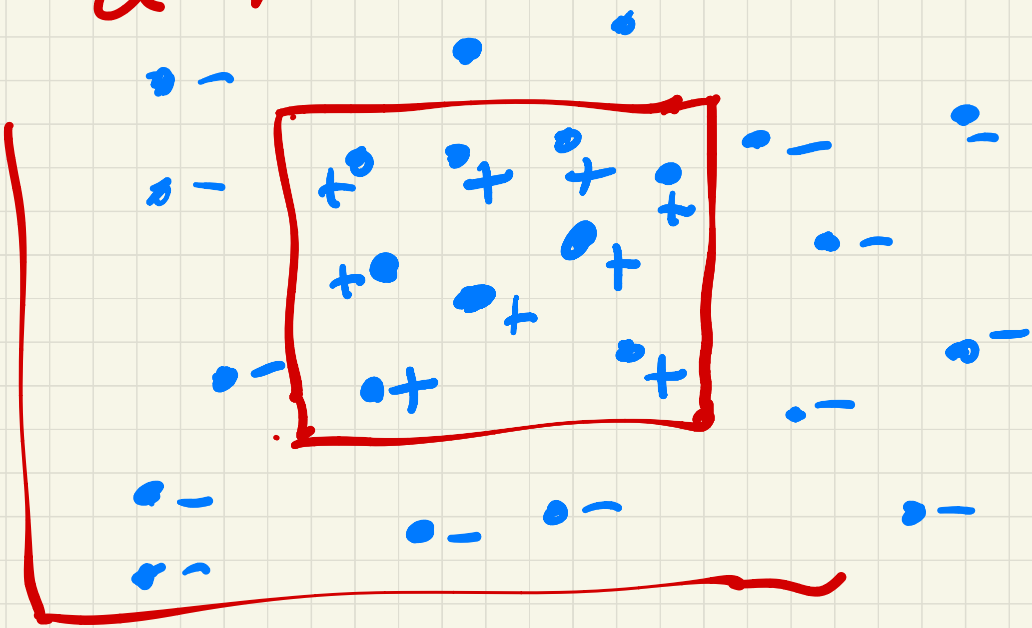
- Aliens arrive from outer space
- You'd like to teach them the concept of "medium build" for adult males (assume binary)
- You can label but not describe

Let's formalize this:

- You (teacher): rectangle R in x-y plane:



You generate data for aliens:



Assume: points drawn i.i.d from $\mathcal{P}$ over $\mathbb{R}^2$

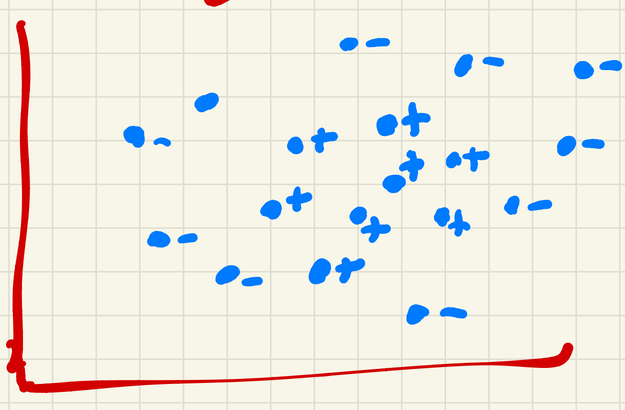
Note: strong assumptions on R, no assumptions on $\mathcal{P}$

Players:

- Input domain: $\mathbb{R}^2$
- Model class: rectangles (binary functions)
- "Target" rectangle R
- Input distribution P

$\Downarrow$

Data:

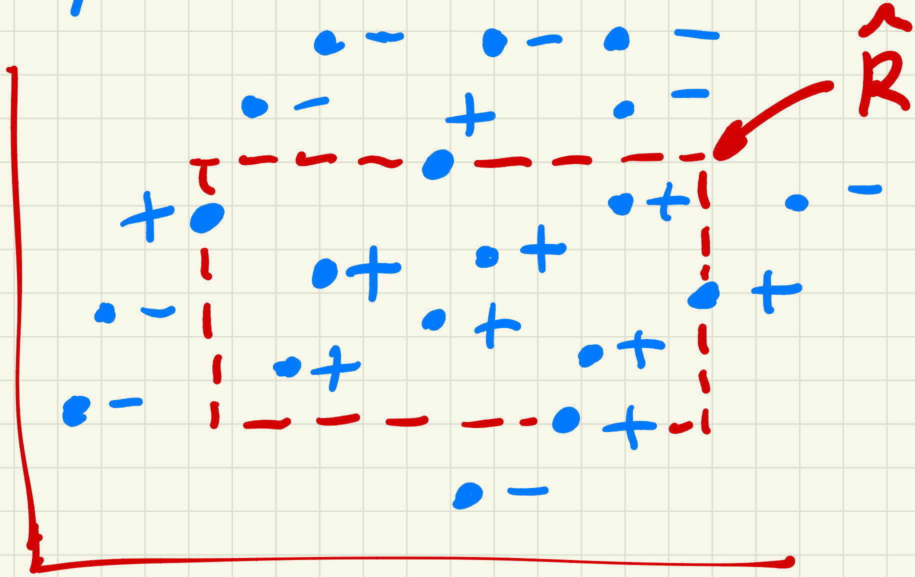


Alien (learner) goal:
from data, learn
a "good" hypothesis
rectangle $\hat{R}$.

- What should "good" mean?
- What algorithm should alien use?

A Proposed Algo:

$\hat{R}$ = "tightest fit" to positive (+) examples.



- Note: exploiting our assumption that "ground truth" (R) is a <u>rectangle</u>!

What can we say about $\hat{R}$?

<u>Claim:</u> Viewed as sets,

$$\hat{R} \subseteq R.$$

What about something stronger/more interesting?

Remember points are drawn i.i.d. from P.

Let's define the error of $\hat{R}$ w.r.t. R & P:

$$\mathcal{E}(\hat{R}) \triangleq \Pr_{x \sim P}\left[\hat{R}(x) \neq R(x)\right] \quad \text{(as } \underline{\text{functions}}\text{)}$$

$$= P\left[\hat{R} \Delta R\right] \quad \text{(as } \underline{\text{sets}}\text{)}$$

<u>Claim:</u> With "high probability" $\mathcal{E}(\hat{R})$ is "small" as long as sample is "large enough".

Analysis

Two inputs/parameters:

- small $\delta > 0$:
  "with high prob" =
  with prob $\geq 1-\delta$ w.r.t.
  draw of sufficiently
  large sample S

- small $\varepsilon > 0$:
  "$\varepsilon(\hat{R})$ small" =
  $\varepsilon(\hat{R}) \leq \varepsilon$

<u>Goal</u>: Show that if $|S| = m$
is large enough, then
w.p. $\geq 1-\delta$, $\varepsilon(\hat{R}) \leq \varepsilon$.

Remark: Note that

$$E_S[\varepsilon(\hat{R})] \leq (1-\delta)\varepsilon + \delta \cdot 1$$

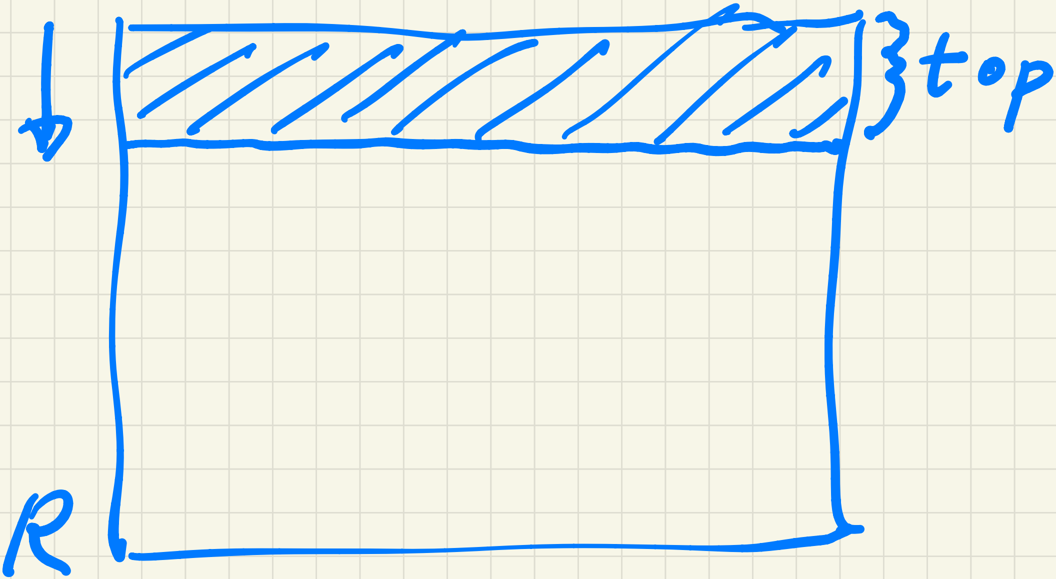
So why have **both** ε & δ?

δ: Bounds prob. of a wildly unrepresentative sample S

ε: Bounds error on representative samples

$\hat{R}$ is "Probably ($\geq 1-\delta$) Approximately ($\leq \varepsilon$) Correct"

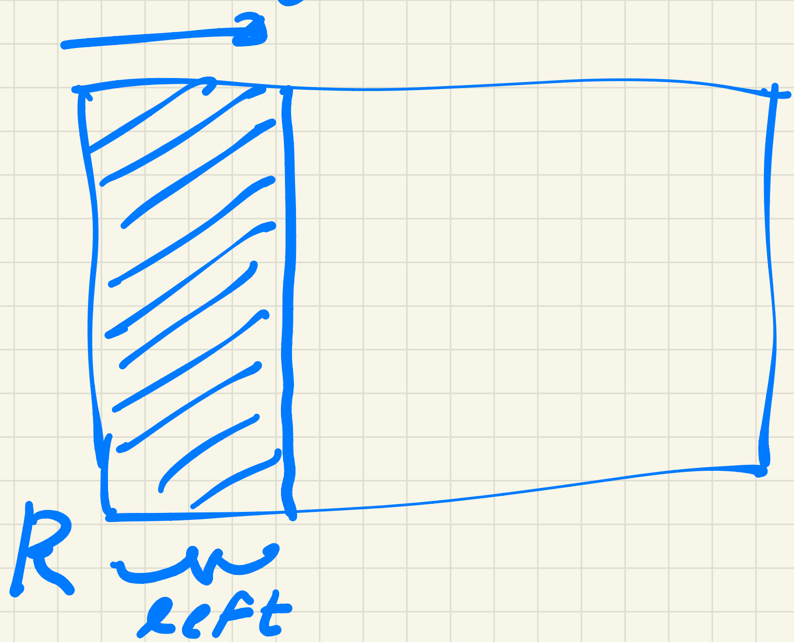
Let's define 4 subsets of R (w.r.t. P):



$$\Pr_{x \sim P}[x \in top] = \varepsilon/4$$

(Q: What if $P[R] < \varepsilon/4$?
Assume not for now.)

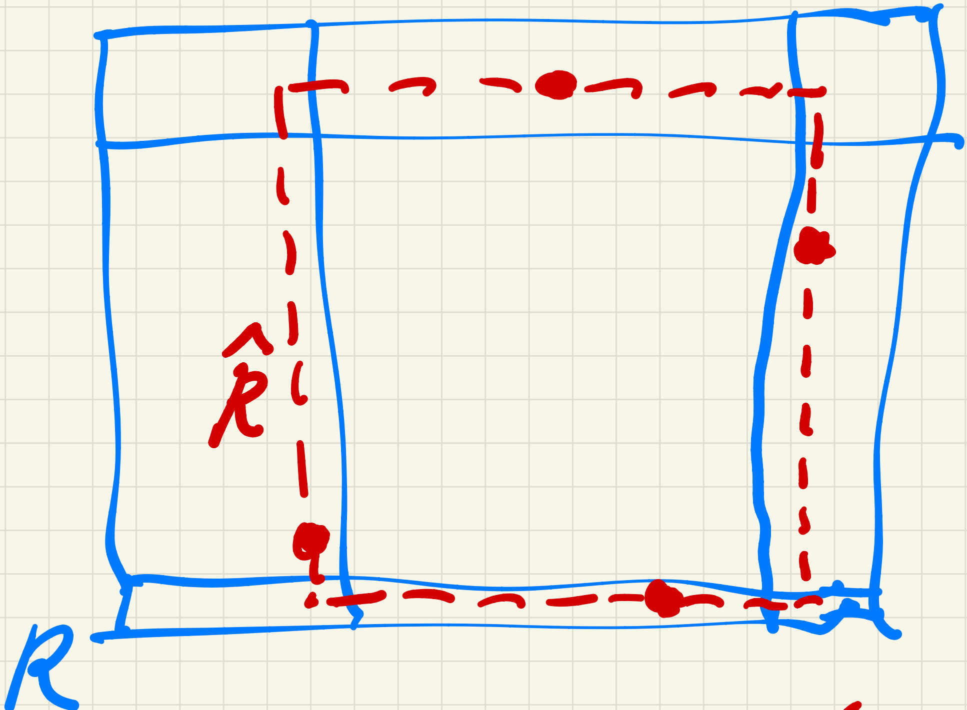
Similarly:



$$\Pr_{x \sim P}[x \in \text{left}] = \varepsilon/4$$

Similarly for right, bottom.

<u>If</u> sample hits all of top, bottom, left, right:



Then $\varepsilon(\hat{R}) \le \varepsilon/4 + \varepsilon/4 + \varepsilon/4 + \varepsilon/4$

$= \varepsilon.$

So let's define a bad sample S as one s.t. S misses any of l, r, t, b.

Goal: bound $Pr[S \text{ is bad}]$ by δ.

- Let $m = |S|$ = sample size
- Remember S i.i.d. wrt P
- $Pr[S \text{ misses top}]$ $= (1 - \varepsilon/4)^m$ (indep.)
- Same for b, l, r

$\therefore$ Pr[S misses **any** of t, b, l, r]

$\leq$ Pr[S misses top] +
Pr[S misses bottom] +
Pr[S misses left] +
Pr[S misses right]

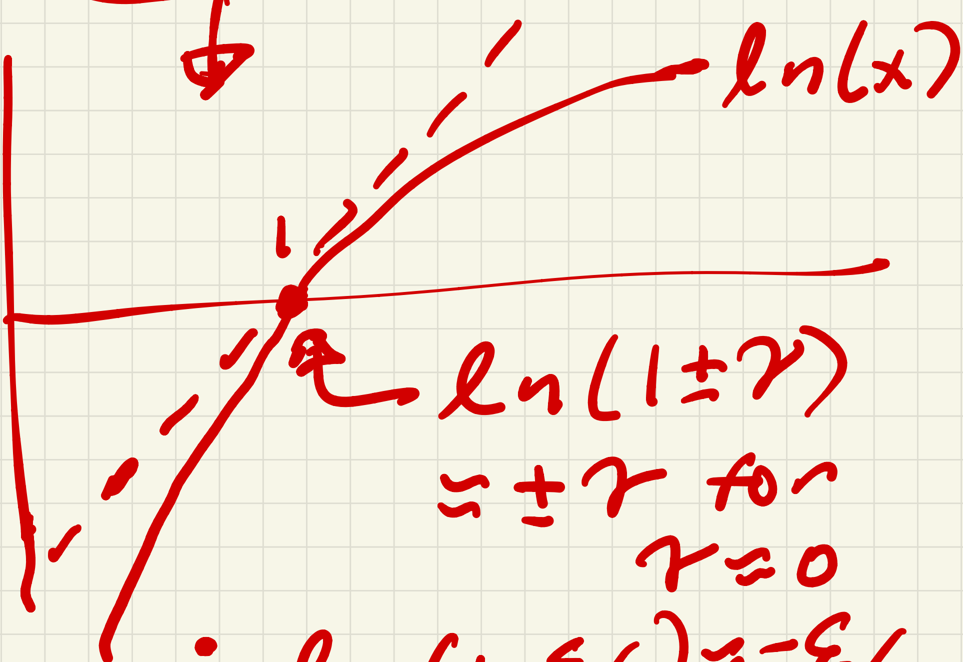
(union bound:
Pr[A or B] $\leq$
Pr[A] + Pr[B])

$$\leq \quad 4 \cdot (1 - \varepsilon/4)^m$$

- So $\Pr[S \text{ bad}] \le 4(1-\varepsilon/4)^m$, set $\le \delta$:

$$4(1-\varepsilon/4)^m \le \delta$$

$$m \ln(1-\varepsilon/4) \le \ln(\delta/4)$$



$\therefore \ln(1-\varepsilon/4) \approx -\varepsilon/4$

$$-m\varepsilon/4 \leq \ln(\delta/4)$$

$$m\varepsilon/4 \geq \ln(4/\delta)$$

$$\boxed{m \geq 4/\varepsilon \ln(4/\delta)}$$

As long as δ is this large, w.p. $\geq 1-\delta$,

$$\varepsilon(\hat{R}) \leq \varepsilon.$$

Oh wait... what if e.g. $P[R] < \varepsilon/4$?

So have a fast algo with small sample complexity and a rigorous analysis.

Proof overview:

- specify algo
- define "bad" events for algo
- bound prob. of each bad event
- take union bound
- set less than δ, do algebra

Extensions?

- Rectangles in $\mathbb{R}^d$?
- Parallelograms in $\mathbb{R}^2$?
- Circles? Triangles?
- Union of 2 rectangles?

$\vdots$

Next Up:

A General Model.