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:

- weight: 160 lbs
- weight: 180 lbs
- build: "medium"
- 5'9" - 6'10"
You generate data for aliens:

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

Note: strong assumptions on \( \mathcal{R} \), no assumptions on \( \mathcal{P} \)
Players:

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

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:

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

Note: exploiting our assumption that "ground truth" (R) is a rectangle!
What can we say about $\hat{R}$?

Claim: 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. $R$ & $P$:

$$\epsilon(\hat{R}) \triangleq \Pr \left[ \hat{R}(x) \neq R(x) \right] \quad \text{as functions}$$

$$= P[\hat{R} \Delta R] \quad \text{as sets}$$

Claim: With "high probability", $\epsilon(\hat{R})$ is "small" as long as sample is "large enough".
Analysis

Two inputs/parameters:

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

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

Goal: Show that if Islam is large enough, then w.p. \( \geq 1 - S \), \( \varepsilon(\hat{R}) \leq \varepsilon \).
Remark: Note that

\[ E_{S} \left[ \epsilon(\hat{R}) \right] \leq (1-\delta) \epsilon + 0.1 \]

So why have both \( \epsilon \) and \( \delta \)?

\( S \): Bounds prob. of a wildly unrepresentative sample \( S \)

\( \hat{R} \): Bounds error on representative samples

\( \hat{R} \) is "probably \((\geq 1-\delta)\) approximately \((\leq \epsilon)\)"
Let's define 4 subsets of $R$ (w.r.t. $P$):

\[ P(\{ x \in \text{top} \}) = \frac{3}{4} \]

(Q: What if $P[R] < \frac{3}{4}$? Assume not for now.)
Similarly:

Similarly for right, bottom.

\[ \text{Pr}_{x \in \text{left}} = \frac{3}{4} \]
If sample hits all of top, bottom, left, right:

Then $E(\hat{R}) \leq \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{3}{4} = 3$. 
So let's define a bad sample $S$ as one s.t. $S$ misses any of $e, r, t, b$.

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

- Let $m = |S| = \text{sample size}$
- Remember $S$ i.i.d. wrt $P$

- $\Pr[S \text{ misses top}] = (1 - 3/4)^m$ (indep.)
- Same for $b, l, r$
\[ \text{Pr}[S \text{ misses any of } t, l, c, r] \leq \text{Pr}[S \text{ misses top}] + \text{Pr}[S \text{ misses bottom}] + \text{Pr}[S \text{ misses left}] + \text{Pr}[S \text{ misses right}] \]

(union bound:)
\[ \text{Pr}[A \text{ or } B] \leq \text{Pr}[A] + \text{Pr}[B] \]
\[ \leq 4 \cdot (1 - \frac{3}{4})^m \]
So \( \Pr [s \text{ bad}] \leq \\
\left( 1 - \frac{\epsilon}{4} \right)^m \), s.t. \( m \leq 8 \\
\left( 1 - \frac{\epsilon}{4} \right)^m \leq 8 \\
m \ln \left( 1 - \frac{\epsilon}{4} \right) \leq \ln(8/4) \\
\ln(1 - 2^{-\epsilon}) \\
= \pm 2^{-\epsilon} \) for \\
\epsilon \approx 0 \\
\therefore \ln \left( \frac{3}{4} \right) \approx \ln \left( 1 - \frac{\epsilon}{4} \right) \
\[ m^{3/4} \geq \frac{4}{3} \ln \left( \frac{4}{3} \right) \]

As long as \( S \) is this large, w.p. \( \geq 1-\delta \),
\[ \mathbb{E}(\hat{R}) \leq \varepsilon. \]
Oh wait... what if e.g. \( P[R] < 3/4 \)?

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 \( S \), do algebra
Extensions?

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

Next Up:
A General Model.