CMU 15-251
Learning theory

Teachers:
Anil Ada
Ariel Procaccia (this time)

The PAC model

- PAC = probably approximately correct
- Introduced by Valiant [1984]
- Learner can do well on training set but badly on new samples
- Establish guarantees on accuracy of learner when generalizing from examples

The PAC model

- Input space X
- D distribution over X: unknown but fixed
- Learner receives a set S of m instances $x_1, \ldots, x_m$, independently sampled according to D
- Function class C of functions $h: X \to \{+, -\}$
- Assume target function $c_t \in C$
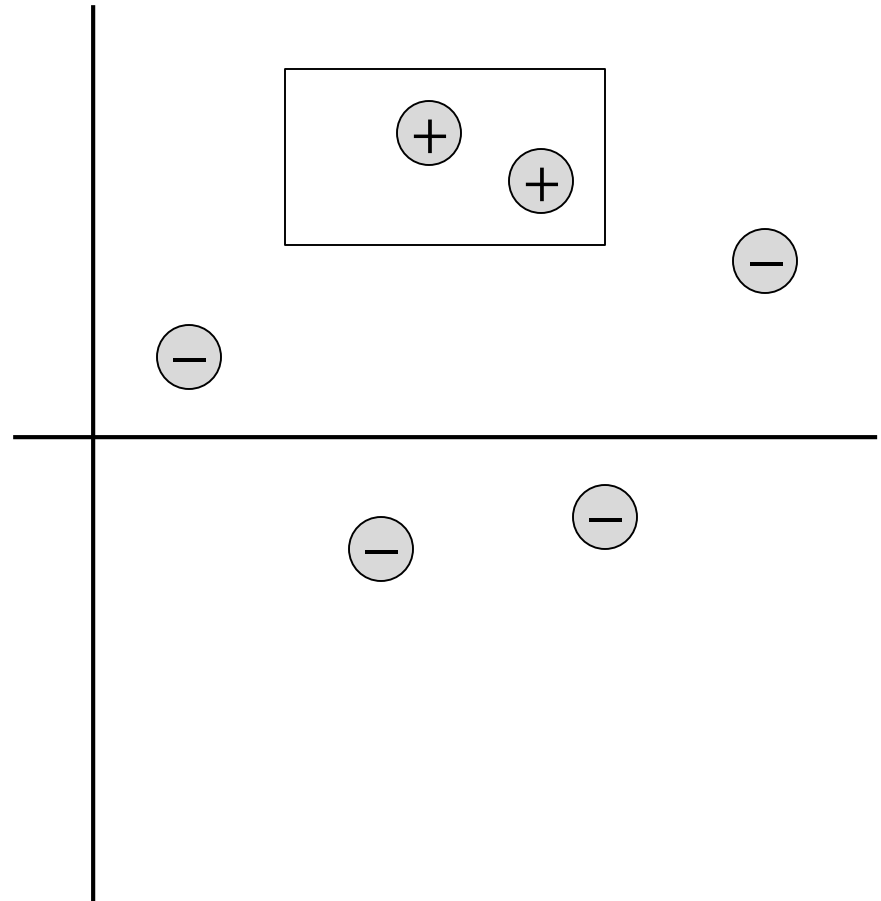- Training examples $Z = \{(x_i, c_t(x_i))\}$

Example: Faces

- $X = \mathbb{R}^n$
- Each $x \in X$ is a vector of colors, one per pixel
- $c_t(x) = +$ iff x is a picture of a face
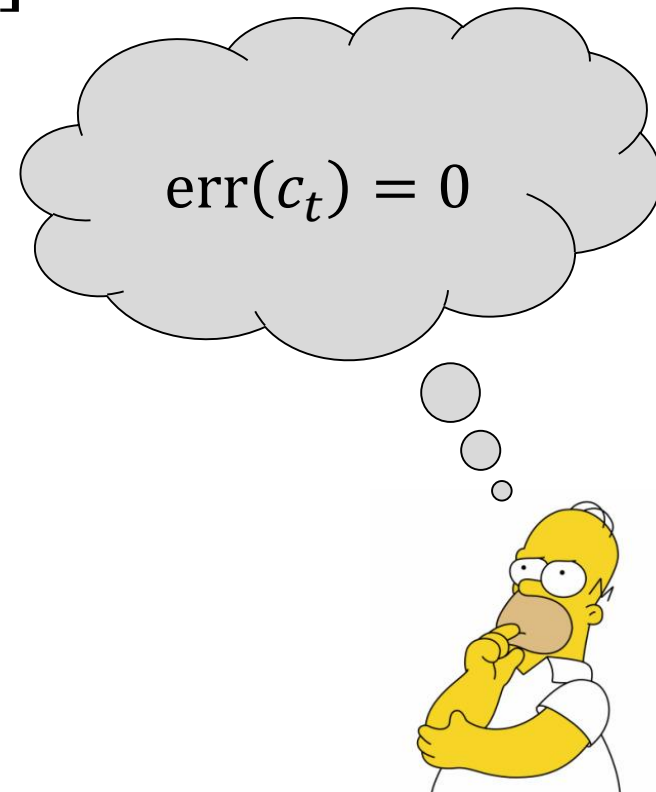- Training examples: Each is a picture labeled "face" or "not face"
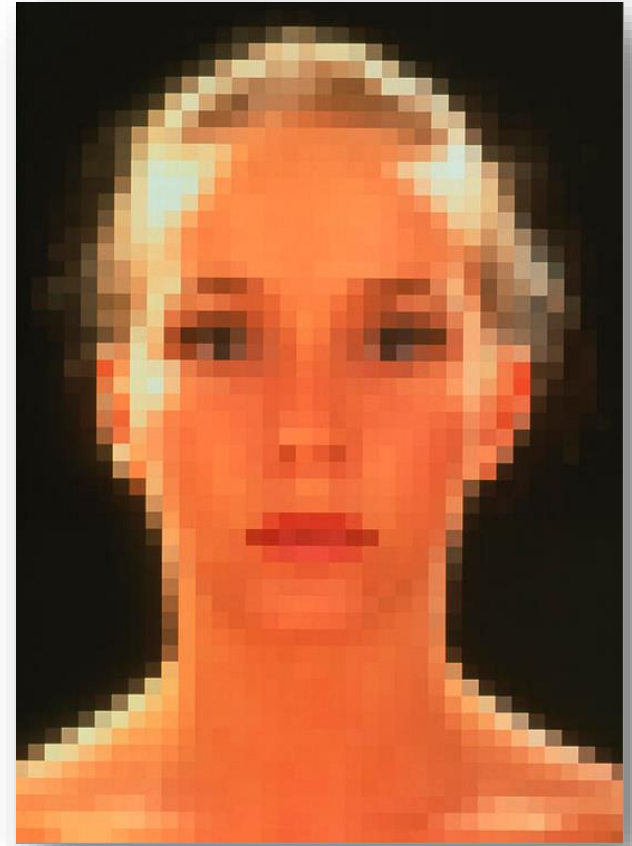
Example: rectangle learning

- $X = \mathbb{R}^2$
- $C =$ axes-aligned rectangles
- $h(x) = +$ iff x is contained in h

The PAC model

- The error of function h is
$$\text{err}(h) = \Pr_{x \sim D}[x : c_t(x) \neq h(x)]$$

- Given accuracy parameter $\epsilon > 0$, would like to find function h with $\text{err}(h) \leq \epsilon$

- Given confidence parameter $\delta > 0$, would like to achieve $\Pr[\text{err}(h) \leq \epsilon] \geq 1 - \delta$

$$\text{err}(c_t) = 0$$

Carnegie Mellon University

The PAC model

- A learning algorithm L is a function from training examples to C such that: for every $\epsilon, \delta > 0$ there exists $m_0(\epsilon, \delta)$ such that for every $m \geq m_0$ and every D, if m examples Z are drawn from D and $L(Z) = h$ then
$$\Pr[\text{err}(h) \leq \epsilon] \geq 1 - \delta$$

- C is learnable if there is a learning algorithm for C

$m_0(\epsilon, \delta)$ is independent of D!
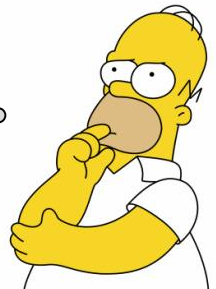
Rectangles are learnable

- Learning algorithm: given training set, return tightest fit for positive examples

- Theorem: axes-aligned rectangles are learnable with $m_0(\epsilon, \delta) \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}$
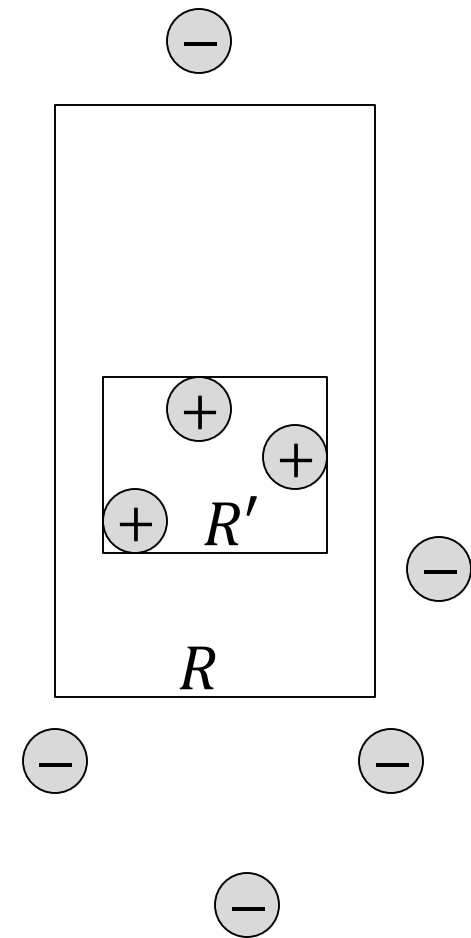
Rectangles are learnable*

- Proof:
  - Target rectangle R
  - Recall: our learning algorithm returns the tightest-fitting R' around the positive examples
  - For region E, let
    $$w(E) = \Pr_{x \sim D}[x \in E]$$
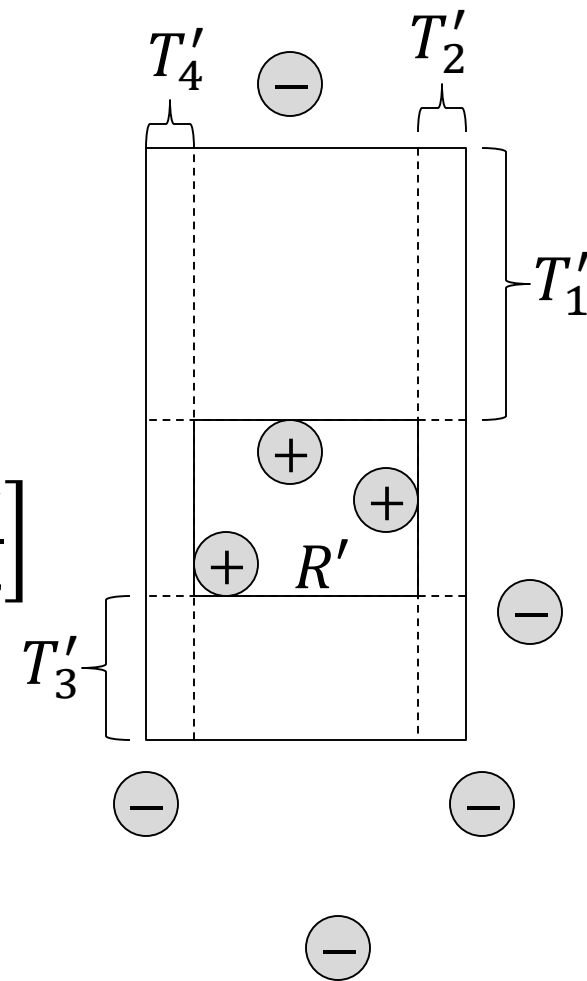 - $\text{err}(R') = w(R \setminus R')$ (why?)

Rectangles are learnable*

- Proof (cont.):
  - Divide $R \setminus R'$ into four strips T_1', T_2', T_3', T_4'
  - $\mathrm{err}(R') \leq \sum_{i=1}^{4} w(T_i')$
  - We will estimate $\Pr\left[w(T_i') \geq \frac{\epsilon}{4} \right]$
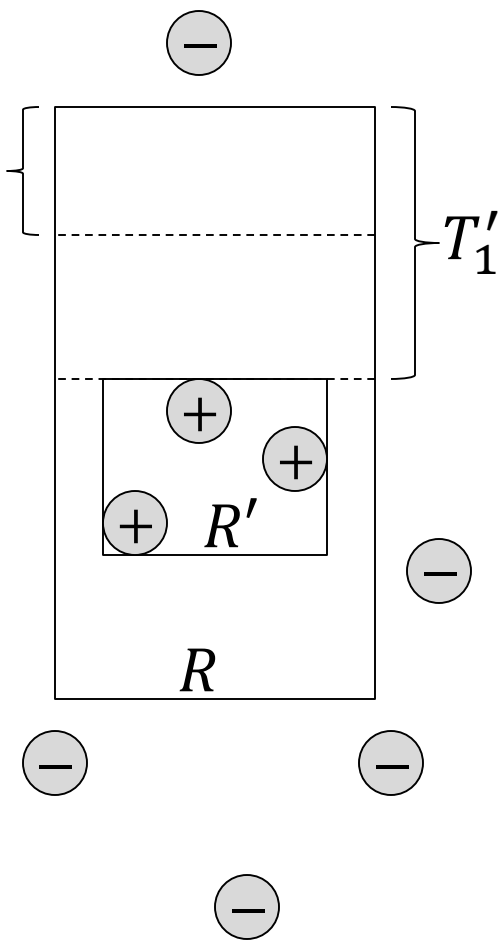
Rectangles are learnable*

- Proof (cont.):
  - Focusing wlog on T_1', define a strip T_1 such that $w(T_1) = \frac{\epsilon}{4}$
  - $w(T_1') \geq \frac{\epsilon}{4} \Leftrightarrow T_1 \subseteq T_1'$
  - $T_1 \subseteq T_1' \Leftrightarrow x_1, \dots, x_m \notin T_1$
  - $w(T_1') \geq \frac{\epsilon}{4} \Leftrightarrow x_1, \dots, x_m \notin T_1$
  - $\Pr[x_1, \dots, x_m \notin T_1] = \left(1 - \frac{\epsilon}{4}\right)^m$
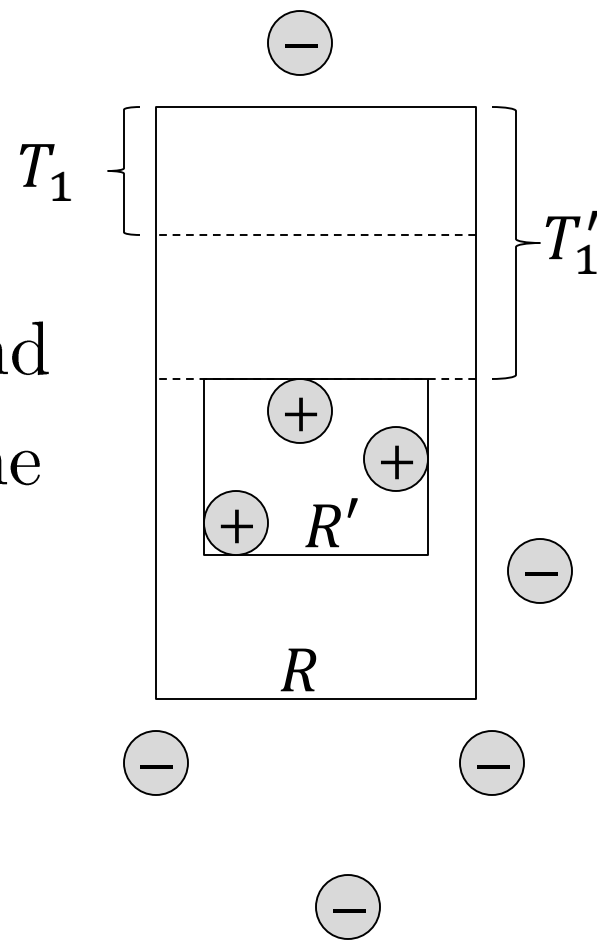
Rectangles are learnable*

- Proof (cont.):
  - $\Pr[w(R \setminus R') \geq \epsilon] \leq 4 \left(1 - \frac{\epsilon}{4}\right)^m$
  - So we want $4 \left(1 - \frac{\epsilon}{4}\right)^m \leq \delta$, and with a bit of algebra we get the desired bound ∎
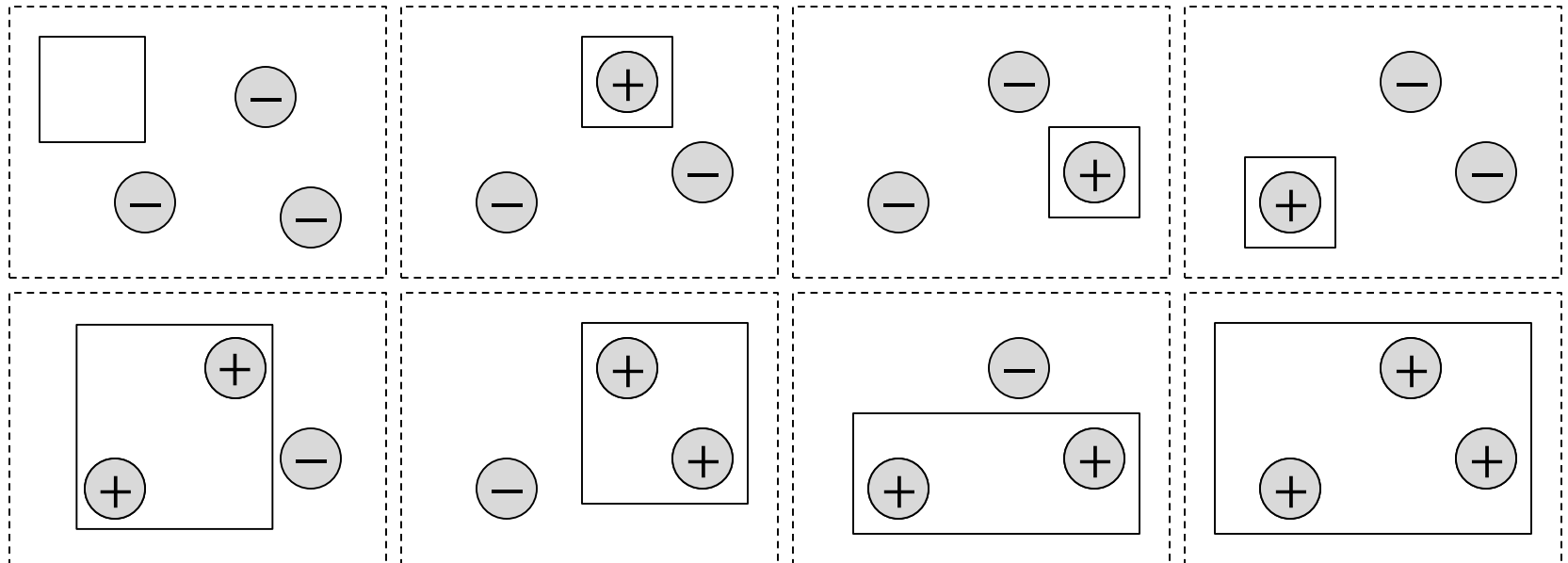
* Just for fun

VC DIMENSION

- We would like to obtain a more general result

- Let $S = \{x_1, \ldots, x_m\}$

- $\Pi_C(S) = \{(h(x_1), \ldots, h(x_m)) \mid h \in C\}$

VC DIMENSION

$$\Pi_C(S) = \{(-,-,-),(-,+,-),(-,-,+),(+,-,-),$$
$$(+,+,-),(-,+,+),(+,-,+),(+,+,+)\}$$

VC DIMENSION

- X = real line
- C = intervals; points inside interval are labeled by $+$, outside by $-$
- Poll 1: what is $|\Pi_C(S)|$ for $S = $ ⸺○⸺○⸺
  1. 1
  2. 2
  3. 3
  4. 4

VC DIMENSION

- Poll 2: what is $|\Pi_C(S)|$ for $S = $ ⎯⎯⎯○⎯⎯○⎯⎯○⎯⎯

  1. 5
  2. 6
  3. 7
  4. 8

VC DIMENSION

- S is shattered by C if $|\Pi_C(S)| = 2^{|S|}$
- The VC dimension of C is the cardinality of the largest set that is shattered by C

How do we prove upper and lower bounds?

Example: rectangles

- There is an example of four points that can be shattered
- For any choice of five points, one is "internal"
- A rectangle cannot label outer points by 1 and inner point by 0
- VC dimension is 4
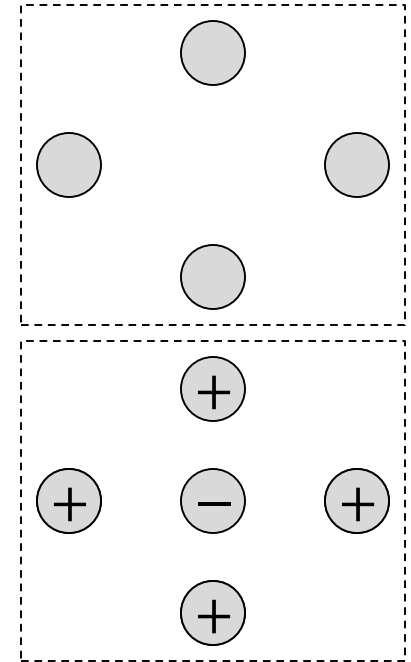
VC DIMENSION

- Poll 3: X = real line, C = intervals, what is VC-dim(C)?

  1. 1
  3. 3

  ②. 2
  4. None of the above

- Poll 4: X = real line, C = unions of intervals, what is VC-dim(C)?

  1. 2
  3. 4

  2. 3
  ④. None of the above

SAMPLE COMPLEXITY

- Theorem: a function class C with VC-dim$(C)= \infty$ is not PAC learnable

- Theorem: Let C with VC-dim$(C)= d$. Let L be an algorithm that produces an $h \in C$ that is consistent with the given samples S. Then L is a learning algorithm for C with $m_0 = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right)$

Summary

- Definitions
    - PAC model
    - Error, accuracy, confidence
    - Learning algorithm
    - $\Pi_C(S)$, shattering
    - VC-dimension
- Turing-award-winning ideas:
    - Learnability can be formalized