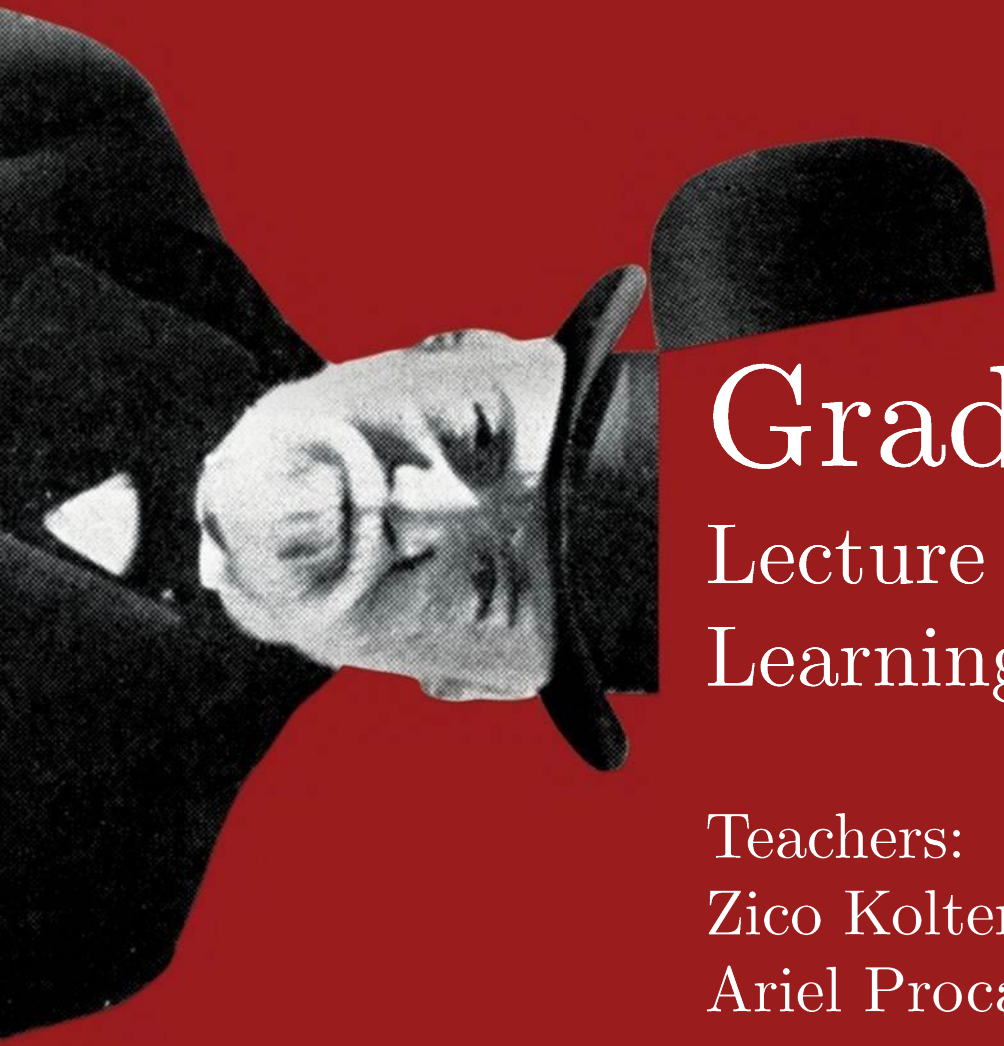


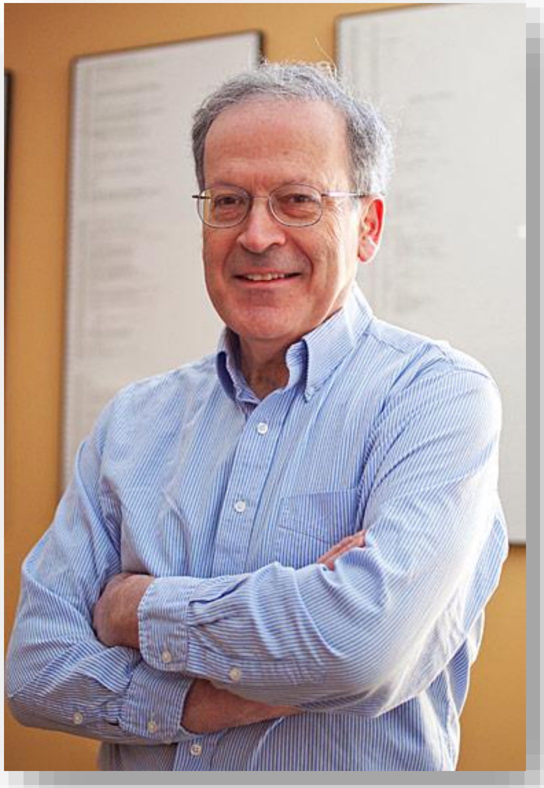
Graduate AI

Lecture 11: Learning Theory

Teachers:
Zico Kolter
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

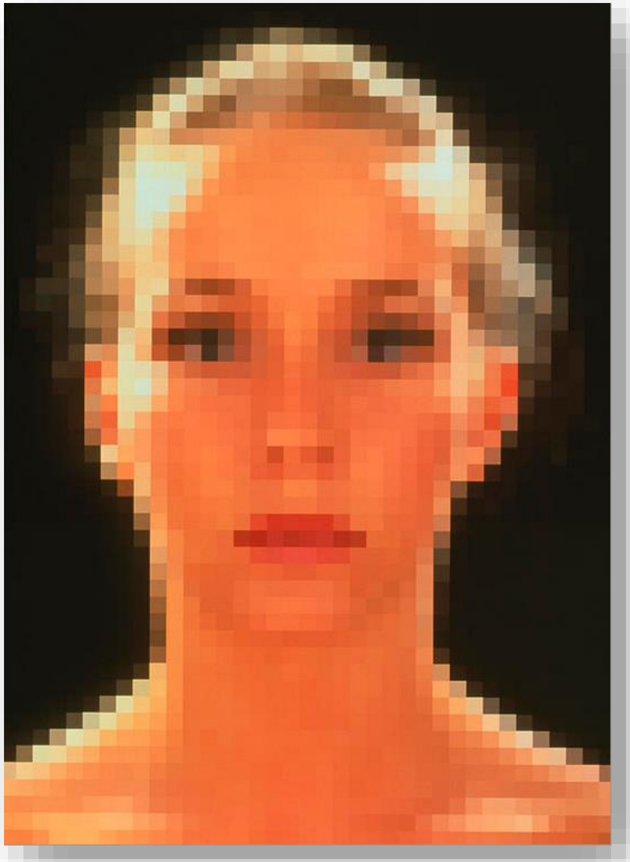


The PAC model

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

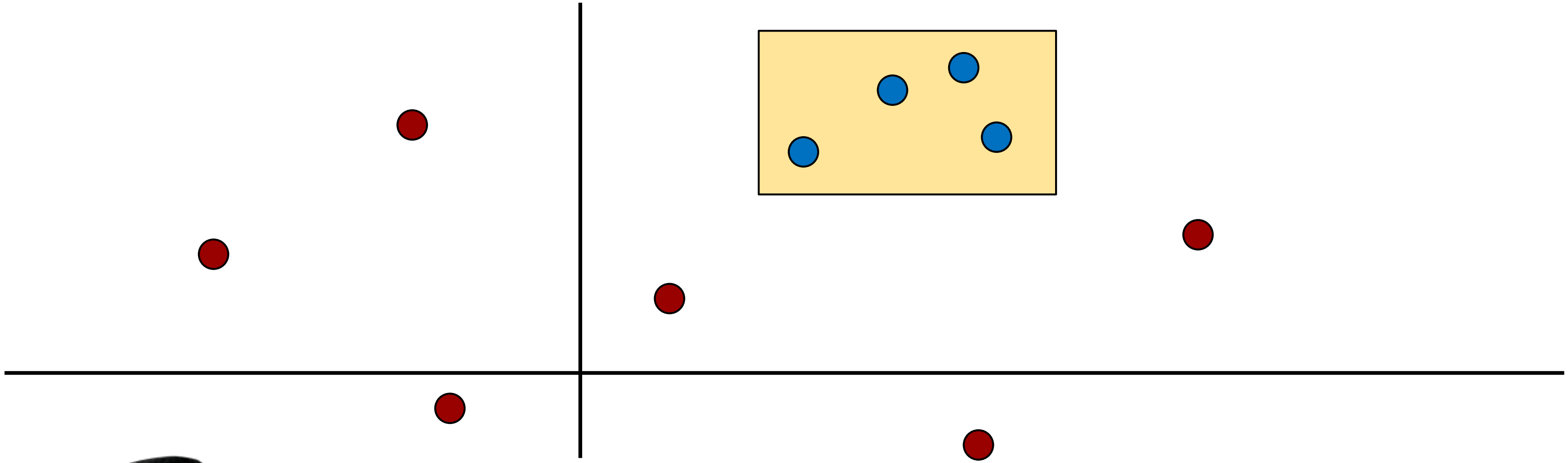
Example: Faces

- $X = \mathbb{R}^{k \times \ell}$
- Each $x \in X$ is a matrix of colors, one per pixel
- $f_t(x) = +$ iff x is a picture of a face
- Training examples: Each is a picture labeled "face" or "not face"



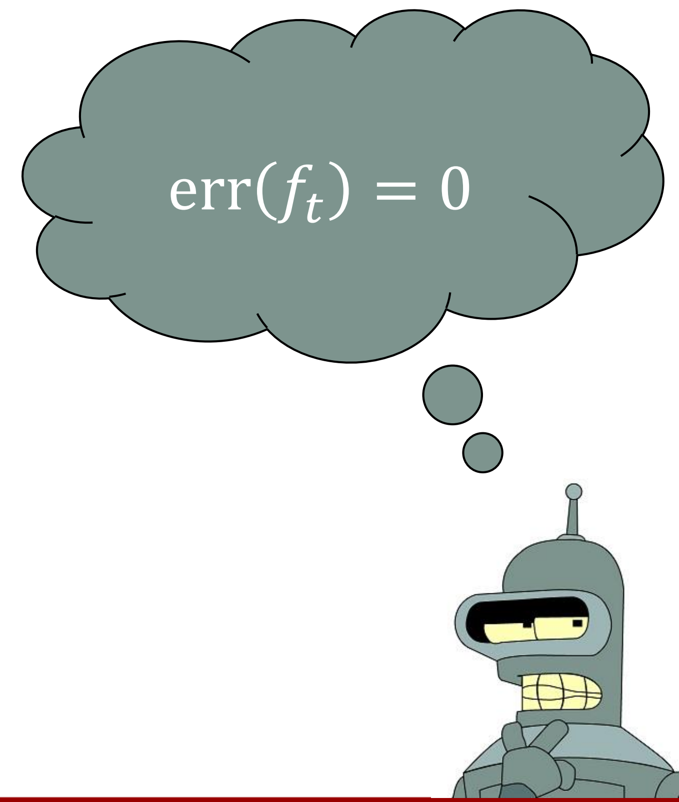
Example: rectangle learning

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



The PAC model

- The error of function f is
  $$\text{err}(f) = \Pr_{x\sim D}[f_t(x) \neq f(x)]$$
- Given accuracy parameter $\epsilon > 0$, would like to find function f with $\text{err}(f) \leq \epsilon$
- Given confidence parameter $\delta > 0$, would like to achieve $\Pr[\text{err}(f) \leq \epsilon] \geq 1 - \delta$

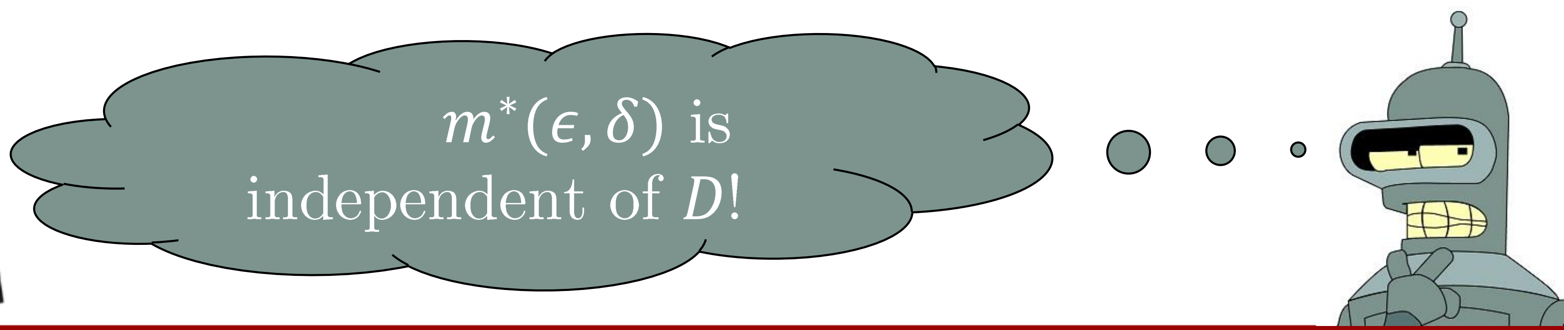


The PAC model

- A learning algorithm L is a function from training examples to F such that: for every $\epsilon, \delta > 0$ there exists $m^*(\epsilon, \delta)$ such that for every $m \geq m^*$ and every D, if m examples Z are drawn from D and $L(Z) = f$ then

$$\Pr[\text{err}(f) \leq \epsilon] \geq 1 - \delta$$

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



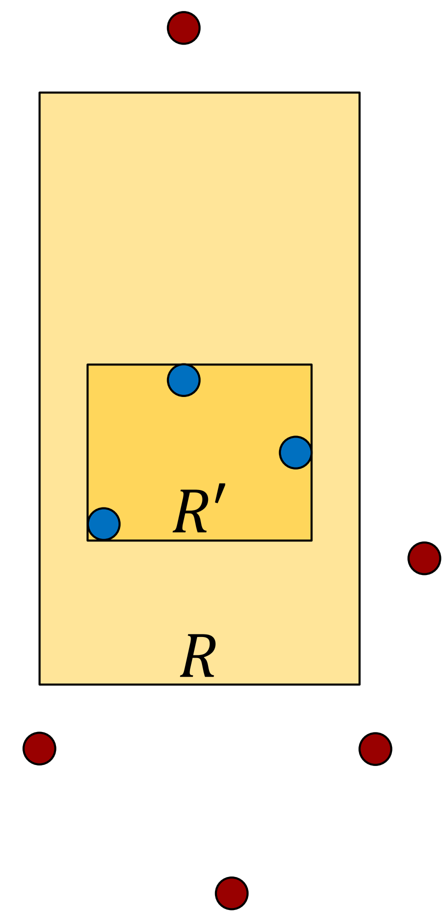
Rectangles are learnable

- $X = \mathbb{R}^2$
- F = axes-aligned rectangles
- Learning algorithm: given training set, return tightest fit for positive examples
- Theorem: axes-aligned rectangles are learnable with sample complexity

$$m^*(\epsilon, \delta) \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}$$

Rectangles are learnable

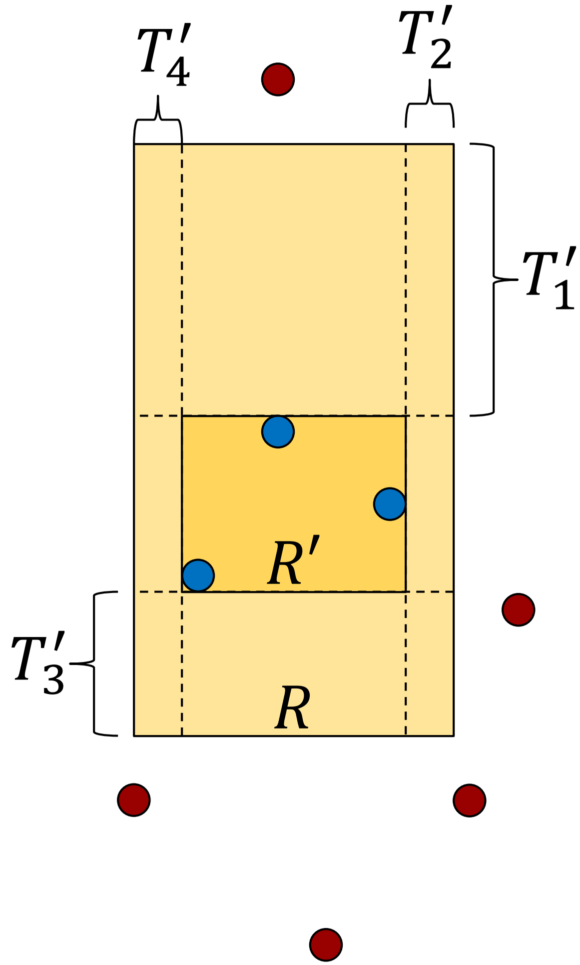
- Proof:
  - Target rectangle R
  - Recall: our learning algorithm returns the tightest-fitting R' around the positive examples
  - For region E, let
    $$w(E) = \Pr_{x \sim D}[x \in E]$$
  - $\operatorname{err}(R') = w(R \setminus R')$ (why?)



Rectangles are learnable

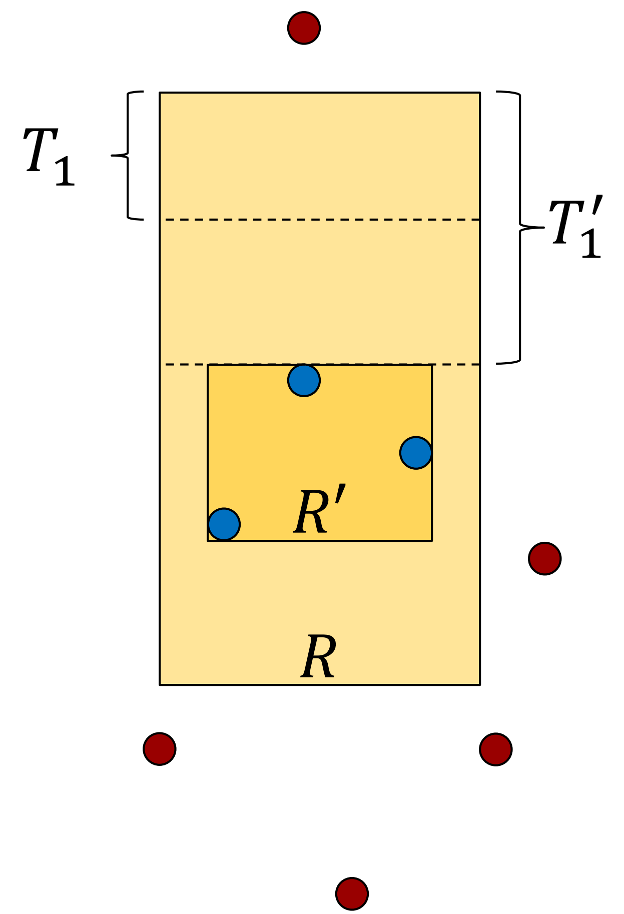
- Proof (cont.):
  - Divide $R \setminus R'$ into four strips T'_1, T'_2, T'_3, T'_4
  - $\mathrm{err}(R') \leq \sum_{i=1}^{4} w(T'_i)$
  - We will estimate

$$\Pr\left[w(T'_i) > \frac{\epsilon}{4}\right]$$



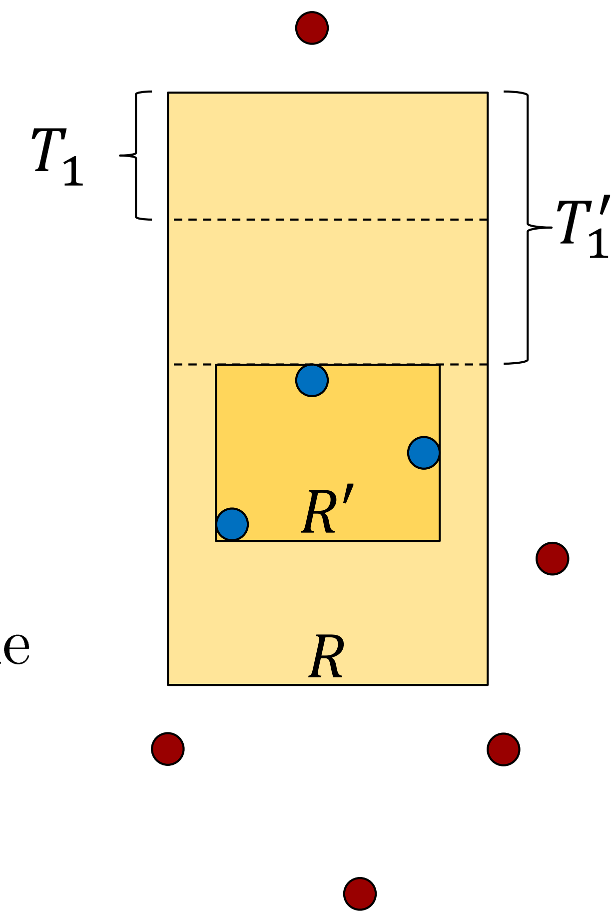
Rectangles are learnable

- Proof (cont.):
  - Focusing wlog on T_1', define a strip T_1 such that $w(T_1) = \frac{\epsilon}{4}$
  - $w(T_1') > \frac{\epsilon}{4} \Leftrightarrow T_1 \subsetneq T_1'$
  - $T_1 \subsetneq T_1' \Leftrightarrow x_1, \ldots, x_m \notin T_1$
  - $w(T_1') > \frac{\epsilon}{4} \Leftrightarrow x_1, \ldots, x_m \notin T_1$
  - $\Pr[x_1, \ldots, x_m \notin T_1] = \left(1 - \frac{\epsilon}{4}\right)^m$



Rectangles are learnable

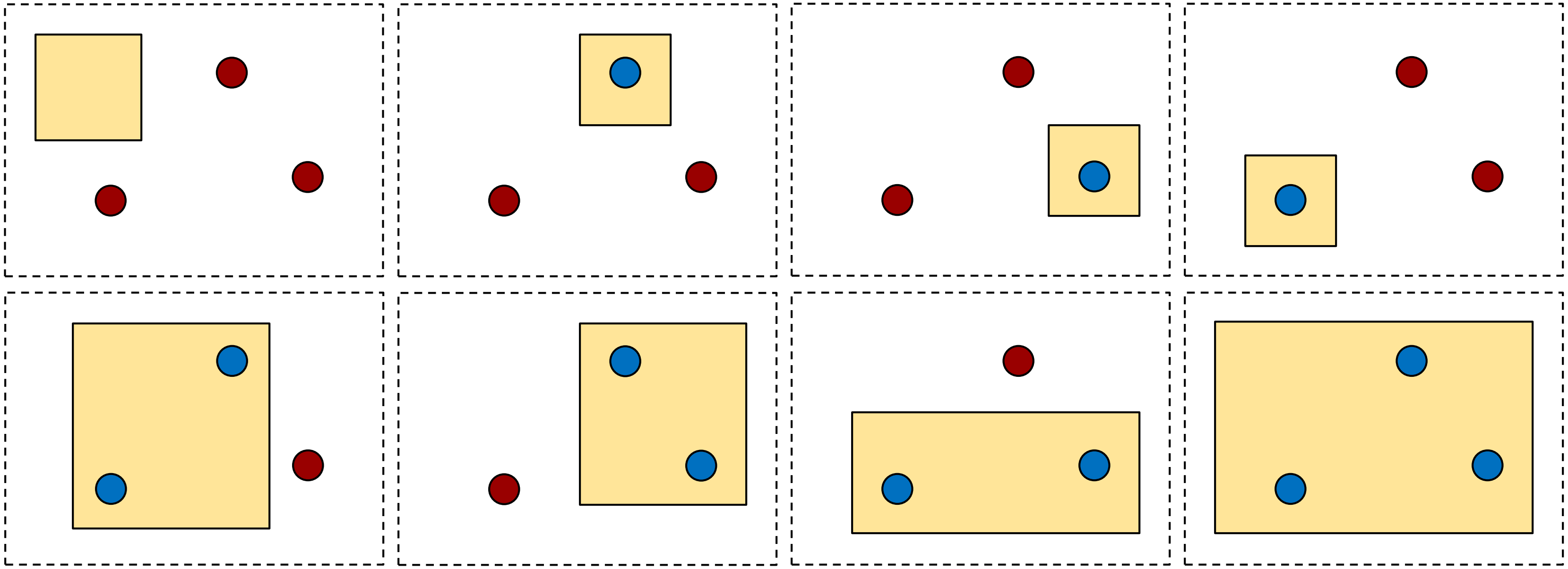
- Proof (cont.):
  - $\Pr[w(R \setminus R') \geq \epsilon] \leq 4\left(1 - \frac{\epsilon}{4}\right)^m$ because at least one T_i' must have $w(T_i') \geq \epsilon/4$
  - So we want $4\left(1 - \frac{\epsilon}{4}\right)^m \leq 4\left(e^{-\frac{\epsilon}{4}}\right)^m \leq \delta$, and with a bit of algebra we get the desired bound ■



VC dimension

- We would like to obtain a more general result
- Let $S = \{x_1, \dots, x_m\}$
- $\Pi_F(S) = \{(f(x_1), \dots, f(x_m)) : f \in F\}$

VC dimension



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

VC DIMENSION

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

VC DIMENSION

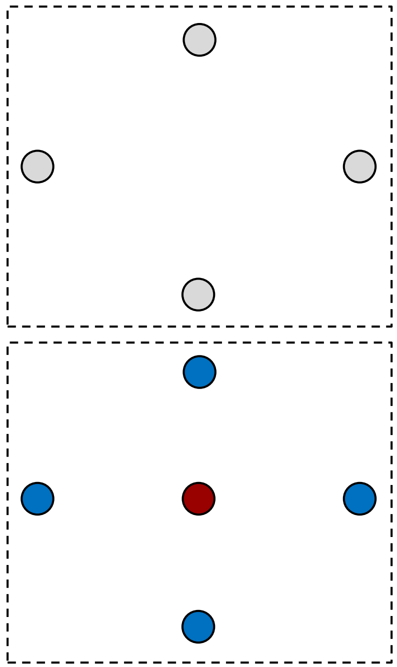
- Poll 2: what is $|\Pi_F(S)|$ for $S =$ 
  1. 5
  2. 6
  3. 7
  4. 8

VC DIMENSION

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

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

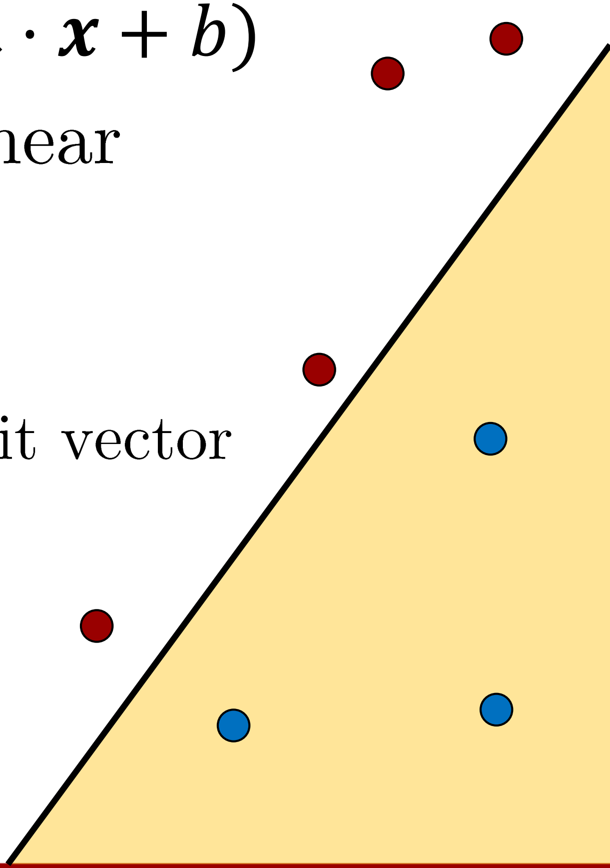


VC DIMENSION

- Poll 3: X = real line, F = intervals, what is VC-dim(F)?
  1. 1
  2. 2
  3. 3
  4. None of the above
- Poll 4: X = real line, F = unions of intervals, what is VC-dim(F)?
  1. 2
  2. 3
  3. 4
  4. None of the above

Example: Linear separators

- $X = \mathbb{R}^d$
- A linear separator is $f(\boldsymbol{x}) = \text{sgn}(\boldsymbol{a} \cdot \boldsymbol{x} + b)$
- Theorem: The VC dimension of linear separators is $d + 1$
- Proof (lower bound):
  - $\boldsymbol{e}^i = (0, \dots, 0,1,0, \dots, 0)$ is the i-th unit vector
  - $S = \{\boldsymbol{0}\} \cup \{\boldsymbol{e}^i : i = 1, \dots, d\}$
  - Given $y^0, \dots, y^d \in \{-1,1\}$, set $\boldsymbol{a} = (y^1, \dots, y^d), b = y^0/2$ ■

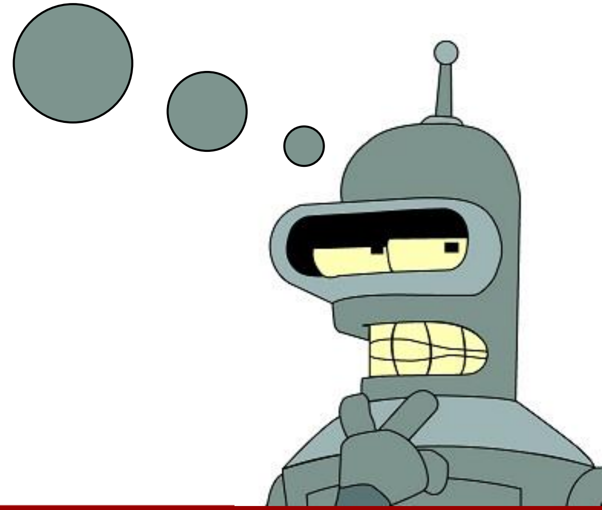


SAMPLE COMPLEXITY

- If for any k there is a sample of size k that can be shattered by F, we say that VC-dim(F)$= \infty$
- Theorem: A function class F with VC-dim(F)$=\infty$ is not PAC learnable
- Theorem: Let F with VC-dim(F)$= d$. Let L be an algorithm that produces an $f \in F$ that is consistent with the given samples S. Then L is a learning algorithm for F with sample complexity

$$m^*(\epsilon, \delta) = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right)$$

Implications for linear classifiers? Overfitting?



SUMMARY

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