PAC LEARNING

- WHAT IS THE "TRUE ERROR" OR GENERALIZATION ERROR OF A CLASSIFIER?

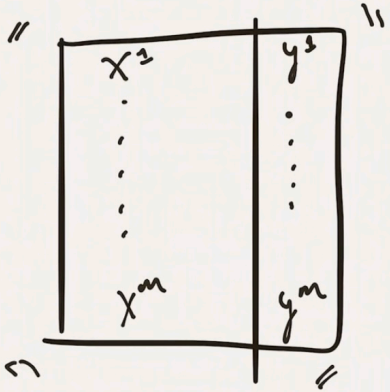
- DECISION TREES: FIX T

D ON NEW EXAMPLES

$$\Pr_{(x,y)\sim D}[T(x) \neq y]$$ OF T

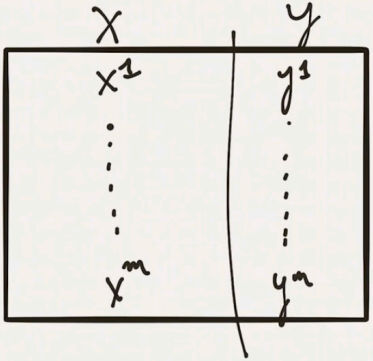
(x, y) — x: CHALLENGE, y: LABEL

TRUE ERROR
GENERALIZATION ERROR



S TRAINING SET

LEARNER IS GIVEN S



EACH

$x^i \in \{0,1\}^n$

$y^i \in \{0,1\}$

Let's BUILD A DECISION TREE

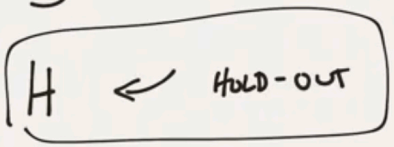
x^i ARE DISTINCT

YOU CAN BUILD A DECISION TREE (SIZE $\geq |S|$)
THAT IS CONSISTENT WITH ALL THE POINTS
IN S.

QUESTION: HOW WELL DOES THIS TREE GENERALIZE
WHAT IS THE TRUE ERROR OF THIS
TREE?

HOW CAN WE ESTIMATE THE TRUE ERROR OF A DECISION TREE?

"HOLD-OUT" OR A "VALIDATION SET"



1. USE S TO BUILD A DECISION TREE

2. ESTIMATE TREE'S TRUE ERROR VIA ITS ERROR ON H.

CROSS-VALIDATION

ANOTHER APPROACH:

TRADE-OFF TRAINING ERROR WITH "MODEL COMPLEXITY"

DEFINE ANOTHER POTENTIAL FUNCTION Φ

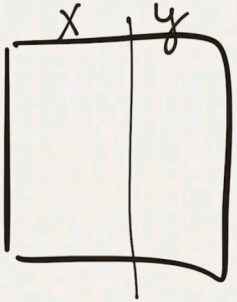
Φ: TREES $\rightsquigarrow \mathbb{R}$ GIVEN A TRAINING SET S

$$\Phi(T) = \text{TRAINING ERROR ON } S + \alpha \cdot \frac{\text{SIZE}(T)}{|S|}$$

MINIMIZE Φ

α: HYPERPARAMETER

ANOTHER APPROACH: MDL "MINIMUM DESCRIPTION LENGTH PRINCIPLE"



SOME NUMBER OF BITS NEEDED TO ENCODE S.

$m \cdot (n+1)$

BUILD A TREE T

LET'S SAY T IS CORRECT ON 90% OF S AND INCORRECT ON 10%.

WE CAN ENCODE S USING #BITS(T) + #BITS TO ENCODE THAT 10%. WE GOT WRONG.

PAC MODEL OF LEARNING

THERE IS A DISTRIBUTION $\mathcal{D}$ on $\{0,1\}^n$ $(\mathbb{R}^n)$

FUNCTION CLASS $\mathcal{C}$ = {DECISION TREES OF SIZE s}

LEARNER (RUNS IN POLYNOMIAL-TIME)

FIX $c \in \mathcal{C}$ — c IS THE UNKNOWN DEC. TREE WE WANT TO LEARN

RECEIVES

(x^1, y^1) $x \sim \mathcal{D}$ $y = c(x)$

(x^2, y^2) $y^i = c(x^i)$

$\vdots$

(x^m, y^m)

GOAL: OUTPUT $h \in \mathcal{C}$

$\Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)] \leq \epsilon = .01$

LEARNER SHOULD BE EFFICIENT (n, s)

"OVER THE DRAWS FROM $\mathcal{D}$"

WITH PROBABILITY AT LEAST $1-\delta$,
THE LEARNER SHOULD OUTPUT A HYPOTHESIS

h s.t. $\Pr_{x \sim \mathcal{D}} [h(x) \neq c(x)] \leq \epsilon$.

RUN-TIME = polynomial $\left(\frac{1}{\epsilon}, \frac{1}{\delta}, n, s\right)$.

IMAGINE LEARNER REQUESTS $x^1, \ldots, x^m$

$x^1 = \ldots = x^m$

PAC → CORRECT

↳ PROBABLY

→ APPROXIMATELY

L. VALIANT [1984]

WHEN CAN WE PAC LEARN A FUNCTION CLASS?
WHAT FUNCTION CLASSES CAN WE PAC LEARN?

GIVE LEARNER AN ALGORITHM A

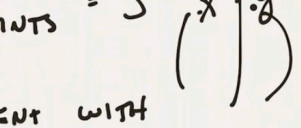
A : TRAINING SETS $\rightsquigarrow$ DECISION TREES

$A(S)$ OUTPUT A TREE T THAT IS CONSISTENT WITH S. SIZE OF T IS GOING TO BE AT MOST s.

A ALWAYS OUTPUTS A CONSISTENT HYPOTHESIS FROM C GIVEN ANY TRAINING SET (ASSUMING THERE IS ONE)

Q: GIVEN ALGORITHM A, HOW CAN WE PAC LEARN C?

- DRAW SUFFICIENTLY MANY TRAINING POINTS = "S"



- USE A TO FIND $c \in C$ CONSISTENT WITH S.
- OUTPUT $\underline{c}$.

Q: HOW LARGE SHOULD S BE?

MARBLE GAME

TWO JARS

JAR 1

ALL BLUE MARBLES

JAR 2

90%. RED MARBLES
10%. BLUE MARBLES.

GOAL: FIGURE OUT IF YOU'VE BEEN GIVEN JAR 1 OR JAR 2

YOU RECEIVE A RANDOM ELEMENT OF JAR ANY-TIME YOU WANT.

- PICK RANDOM MARBLE FROM JAR
  - CASE 1: MARBLE IS RED JAR 2
  - CASE 2: MARBLE IS BLUE.
  - PICK AT MOST 100 MARBLES.

WHAT IS THE PROBABILITY OF FAILURE?

PROB OF FAILURE IS $(.1)^{100}$ ← FAILURE PROB "δ- PARAMETER"

Let's RETURN TO PAC LEARNING

- DRAW MANY SAMPLES S
- RUN A
- OUTPUT CLASSIFIER c THAT IS CONSISTENT WITH S GIVEN FROM A.

WHAT IS THE PROBABILITY THIS PROCEDURE FAILS? $\leq \delta$

BAD EVENT:

OUTPUT c THAT IS CONSISTENT WITH S

BUT HAS TRUE ERROR $> \epsilon$.

$\Pr$ [BAD EVENT] ?

ENUMERATED ALL FUNCTIONS IN $\mathcal{C} = \{c_1, \ldots, c_N\}$

FIX c_1 ASSUME c_1 HAS TRUE ERROR $> \epsilon$

WHAT IS $\Pr_S$ [c_1 IS CONSISTENT WITH S] ? $\leq (1-\epsilon)^{|S|}$

FIX c_2 " c_2 HAS " " $> \epsilon$

$\Pr_S$ [c_2 IS CONSISTENT WITH S] ? $\leq (1-\epsilon)^{|S|}$

FOR EVERY C_i (WITH ERROR $> \epsilon$)

$$\Pr_S[C_i \text{ IS CONSISTENT ON } S] \leq (1-\epsilon)^{|S|}$$

BAD

Q: RANDOMLY FORM S, WHAT IS THE PROB THERE EXISTS A FUNCTION $C \in \mathcal{C}$ WHOSE ERROR $> \epsilon$ AND IS CONSISTENT WITH S?

UNION BOUND A, B $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$

$$\Pr[\text{BAD}] \leq |\mathcal{C}| \cdot (1-\epsilon)^{|S|} \leq \delta$$

SOLVE FOR $|S|$: $(1-x) \approx e^{-x}$

$$|\mathcal{C}| \cdot (1-\epsilon)^{|S|} \leq \delta$$

$$|\mathcal{C}| \cdot e^{-\epsilon|S|} \leq \delta \qquad (1+x \leq e^{x})$$

$$e^{-\epsilon|S|} \leq \frac{\delta}{|\mathcal{C}|}$$

$$-\epsilon|S| \leq \log\left(\frac{\delta}{|\mathcal{C}|}\right)$$

$$|S| > \frac{\log\left(\frac{|\mathcal{C}|}{\delta}\right)}{\epsilon}$$

IF YOU CHOOSE # TRAINING POINTS LARGER THAN $\frac{\log\left(\frac{|\mathcal{C}|}{\delta}\right)}{\epsilon}$

THEN WITH PROB $\geq 1-\delta$, FUNCTION OUTPUT C IS $1-\epsilon$ ACCURATE.

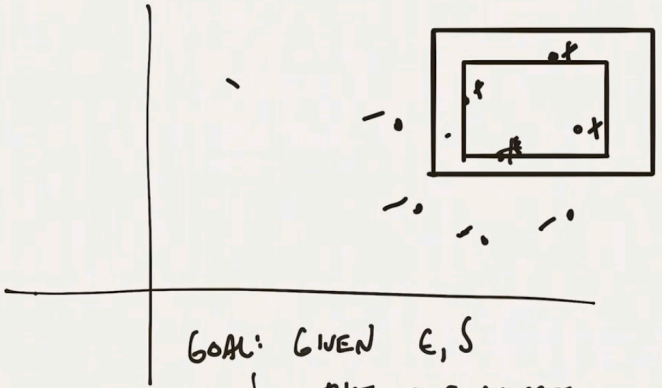
SUGGESTS A

"CONSISTENT HYPOTHESIS"

APPROACH TO LEARNING.

PAC-LEARNING AXIS-PARALLEL RECTANGLES.

WE ARE WORKING IN 2-DIMENSIONS



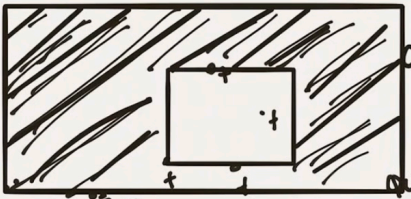
. LABELED + IF THE POINT IS INSIDE C, AXIS-P RECT.

— IF THE POINT IS OUTSIDE C, AXIS-P RECT.

GOAL: GIVEN ϵ, δ

OUTPUT h THAT IS ϵ-ACCURATE

WITH PROB $\geq 1-\delta$



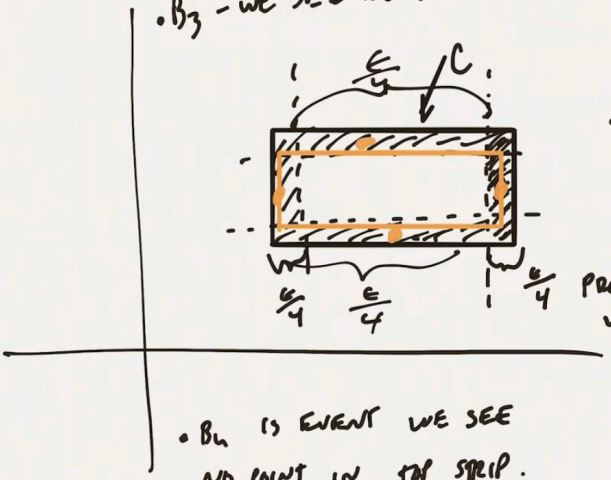
CLAIM: TIGHTEST FITTING RECTANGLE WORKS!

QUESTION: HOW LARGE TO CHOOSE IS?

BAD EVENT: TIGHTEST FITTING RECTANGLE IS SMALL i.e. LOTS OF PROD MASS EXISTS OUTSIDE OF h.

HOW DO WE BOUND THE PROBABILITY THIS HAPPENS?

ANALYZE h (TIGHTEST FITTING) AND SAY SOMETHING ABOUT h VS ALL RECTANGLES THAT ARE LARGE AND CONTAIN h.

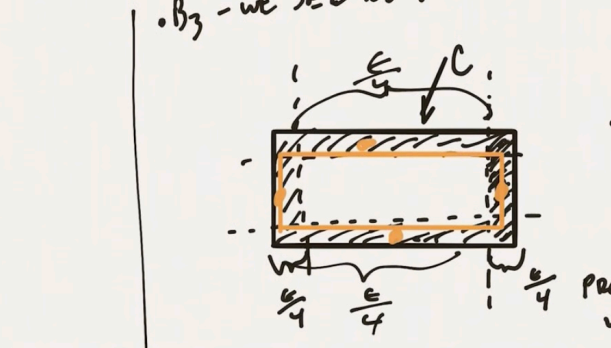


1 BAD EVENT

- B_1 IS THE EVENT WE SEE NO POINT IN RIGHT STRIP

- B_2 IS THE EVENT WE SEE NO POINT IN BOTTOM STRIP.

CLAIM. IF NEITHER $B_1, B_2 B_3, B_4$ OCCUR THEN h, TIGHTEST FITTING RECT IS ϵ-ACCURATE.



1 BAD EVENT

- B_1 IS THE EVENT WE SEE NO POINT IN RIGHT STRIP

- B_2 IS THE EVENT WE SEE NO POINT IN BOTTOM STRIP.

CLAIM. IF I CHOOSE m RANDOM SAMPLES $\Pr[B_1]$

$$\Pr[B_1] \leq \left(1 - \frac{\epsilon}{4}\right)^m$$

$$\Pr[B_1 \cup B_2 \cup B_3 \cup B_4] \leq 4 \cdot \left(1 - \frac{\epsilon}{4}\right)^m \leq \delta$$

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

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

$$e^{-\frac{\epsilon m}{4}} \leq \frac{\delta}{4}$$

$$-\frac{\epsilon m}{4} \leq \log\left(\frac{\delta}{4}\right)$$

$$\Rightarrow m \geq \frac{4 \cdot \log\left(\frac{\delta}{4}\right)}{\epsilon}$$

$1 + x \approx e^x$

$1 - x \approx e^{-x}$

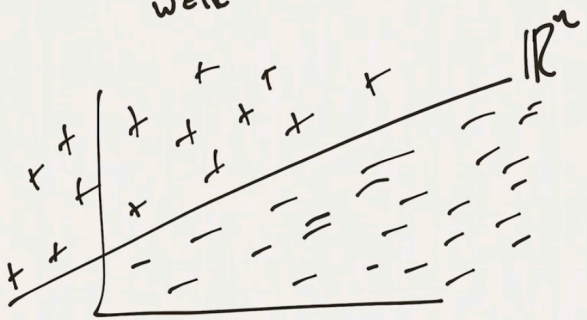
h) TIGHTEST RECT, WILL BE ϵ-ACCURATE w prob $\geq 1 - \delta$

- $C = \{\text{HALFSPACES}\}$

$$f_{(x)} = \text{SIGN}\left(\underline{w} \cdot x - \underline{\theta}\right) \quad \text{(SIGN} : \mathbb{R} \to \mathbb{R}\text{)}$$

f is BOOLEAN 0 or $\underline{1}$

$w \in \mathbb{R}^n$ $\quad$ $x \in \mathbb{R}^n$



PAC LEARNING ALGORITHMS FOR HALFSPACES.

ONE APPROACH FOR LEARNING HALFSPACES

$w \in \mathbb{R}^n$ IS UNKNOWN

$\theta \in \mathbb{R}$ UNKNOWN

$$f = \text{SIGN}\left(\sum_{i=1}^{n} w_i x_i - \theta\right)$$

GIVEN DRAWS FROM $\mathcal{D}$ $(x, f(x))$

$x \uparrow$

$(0101_0, \text{POS}) \rightarrow w_2 + w_4 > \theta$ $\quad w_i \in \mathbb{Z}$ IN SOME BOUNDED RANGE.

$(0110, \text{NEG}) \rightarrow w_2 + w_3 \leq \theta$

EACH LABELED EXAMPLE $\rightarrow$ LINEAR INEQUALITY.

SYSTEM OF LINEAR INEQUALITIES.

CAN WE FIND A CONSISTENT HYPOTHESIS?

GENERAL-PURPOSE TOOL CALLED LINEAR PROGRAMMING

CROSS-VALIDATION

- HOLD-OUT APPROACH FOR TESTING/APPROXIMATING THE TRUE ERROR OF A CLASSIFIER

- LET'S ASSUME CLASSIFICATION; SO HYPOTHESIS h IS GOING TO OUTPUT $\{0,1\}$ $\{-1,+1\}$ VALUES.

"HOLD-OUT"
1. LEAVE SOME PART OF TRAINING SET OUT DURING TRAIN TIME.
2. TEST CLASSIFIER ON THIS HELD OUT SET.

1ST INEQUALITY — MARKOV'S INEQUALITY

Let X be R.V. THAT TAKES ON ONLY POS VALUES.

$$Pr\left[X \geq K \cdot \mathbb{E}[X]\right] \leq \frac{1}{K}$$

CHEBYSHEV'S INEQUALITY

$$VAR[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \qquad \mathbb{E}[X] = \mu$$

$$\sqrt{VAR[X]} = \text{STANDARD DEVIATION}(X) = \sigma$$

$$Pr\left[|X - \mu| > t \cdot \sigma\right] \leq \frac{1}{t^2}$$

CHERNOFF BOUND

$X_1, X_2, \ldots, X_n \qquad \mathbb{E}[X_i] = p \qquad X_i \in \{0,1\}$

$S = \sum_{i=1}^{n} X_i \qquad \mu = \mathbb{E}[S] = p \cdot n$

$\mathbb{E}[X_1 + \cdots + X_n] = p \cdot n$

$$\Pr[S > \mu + \delta_n] \leq e^{-2n\delta^2}$$

$$\Pr[S < \mu - \delta_n] \leq e^{-2n\delta^2}$$

$$\Rightarrow \Pr[|S - \mu| > \delta_n] \leq 2 \cdot e^{-2n\delta^2}$$

APPLY THE CHERNOFF BOUND TO THE CASE OF ESTIMATING THE TRUE ERROR OF A CLASSIFIER

HOLD-OUT SET S WE'LL SAY $|S|=n$

FIX h (GENERATED USING SOME WD TRAINING SET)

RECALL $\mathcal{D}$, S IS A SAMPLE DRAWN FROM $\mathcal{D}$ IND OF TRAINING SET

$$z = \Pr_{x \sim \mathcal{D}} [h(x) \neq c(x)]$$

c: UNKNOWN FUNCTION TRYING TO LEARN

Let X_i be R.V. EQUALS 1 if h is INCORRECT ON THE i^{th} element of S.

0 if h is correct on the i^{th} element of S.

$X_1, \ldots, X_n$ $\quad X_i = \begin{cases} 1 & \text{if } h \text{ is incorrect on } i\text{th} \\ 0 & \text{otherwise.} \end{cases}$

$S = \sum_{i=1}^{n} X_i$ $\quad \mathbb{E}[S] = n \cdot p \leftarrow$ true error of h.

$p = \mathbb{E}[X_i] = \mathbb{E}[X_1] = \mathbb{E}[X_n]$

$$\Pr\left[|S - n \cdot p| > \delta n\right] \leq 2 e^{-2n\delta^2}$$

(RECALL p IS TRUE ERROR OF CLASSIFIER h).

$\delta = .1$

$\Pr\left[|S - n \cdot p| > .1n\right] \leq 2 e^{-\frac{2n}{100}}$

How LARGE to choose n BEFORE THIS QUANTITY BECOMES SMALL?

α

$e^{-\frac{2n}{100}} < \frac{\alpha}{2}$

$\frac{-2n}{100} < \log\left(\frac{\alpha}{2}\right) \Rightarrow n > 50 \cdot \log\left(\frac{\alpha}{2}\right)$

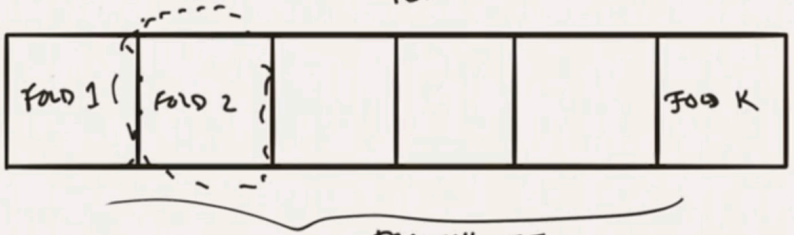
IF $|S - n \cdot p| \leq .1n$ $\Rightarrow$ ERROR RATE ON S IS WITHIN .1 OF TRUE ERROR RATE.

HOLD-OUT SET IS SOMEWHAT EXPENSIVE

- DATA IS EXPENSIVE
- IF WE WANT TO TRY OUT MULTIPLE METHODS FOR GENERATING CLASSIFIERS, WE QUICKLY LOSE CONFIDENCE IN OUR ESTIMATES.

CROSS-VALIDATION

- TRAIN USING FOLDS 2 .. FOLD K
- TEST ON FOLD 1



Let's GO BACK TO DECISION TREES

TRAINING SET S

- SHOULD I BUILD A DECISION TREE OF DEPTH 10
  " " " " " 15

WE DECIDE USING CROSS-VALIDATION

Let's GO BACK TO DECISION TREES

TRAINING SET S

- SHOULD I BUILD A DECISION TREE OF DEPTH 10
  " " " " " " 15

WE DECIDE USING CROSS-VALIDATION