PAC LEARNING

PROBABLY APPROXIMATELY CORRECT

CLASSIFICATION:

INPUT

m data points

$(x_1, y_1), \ldots, (x_m, y_m)$

x_i chosen i.i.d. from some distribution

$y_i = c(x_i)$

← UNKNOWN CONCEPT

$c: X \mapsto y$

GOAL

learn c

APPROACH

- consider a finite set of hypotheses $\mathcal{H}$ (decision trees of depth d)
- find h with

$$err_{train}(h) = 0$$

(we say that h is CONSISTENT with data)

WHAT IS THE PROBABILITY THAT

$$err_{true}(h) > \varepsilon \; ?$$

$$\Pr[h(x) \neq c(x)]$$

PROBABILITY OF A MISTAKE

BAD HYPOTHESES: $h_1, \ldots, h_K$
(TRUE ERROR $> \varepsilon$)

STEP I:
What is the probability that a fixed BAD hypothesis gets all data right?
say h_3:

$$\Pr[h_3 \text{ makes an error}] > \varepsilon \qquad \text{[BAD HYPOTHESIS]}$$

$$\Pr[h_3 \text{ gets a single point right}] \leq 1-\varepsilon$$

$$\Pr[h_3 \text{ gets } m \text{ points right}] \leq (1-\varepsilon)^m$$

STEP II:

Prob. that any bad hypothesis gets all m points right

$$\leq K \cdot (1-\varepsilon)^m \qquad \text{[UNION BOUND]}$$

$$\leq |\mathcal{H}| \cdot (1-\varepsilon)^m \leq |\mathcal{H}| \cdot e^{-m\varepsilon} \qquad [\text{by } 1-x \leq e^{-x}]$$

WITH ENOUGH DATA, WE MANAGE TO EXCLUDE ALL BAD HYPOTHESES WITH HIGH PROBABILITY

THEOREM [Haussler 1988]
For a finite hypothesis space $\mathcal{H}$, m i.i.d. training examples, $0 < \varepsilon < 1$, and any h consistent w/ data:

$$\Pr[\text{err}_{\text{true}}(h) > \varepsilon] \leq |\mathcal{H}| e^{-m\varepsilon}$$

SUPPOSE WE WANT THIS PROBABILITY $< \delta$:

- How many examples suffice?

  WANT: $|\mathcal{H}| e^{-m\varepsilon} \leq \delta$

  REQUIRE: $m \geq \dfrac{\ln|\mathcal{H}| + \ln(1/\delta)}{\varepsilon}$

  SAMPLE COMPLEXITY

  NOTE:
  - grows gently with $|\mathcal{H}|$ and amount of confidence $1-\delta$
  - can be large if we want ε to be small

- What is the largest error we get after perfectly fitting m samples?

$$err_{true}(h) \leq \frac{\ln|\mathcal{H}| + \ln(1/\delta)}{m}$$

REWRITE ERROR BOUND:

$$err_{true}(h) \leq \frac{1}{\log_2 e} \cdot \frac{\overbrace{\log_2|\mathcal{H}|}^{\#\text{bits to identify } h} + \log_2(1/\delta)}{m}$$

If $\mathcal{H}$ finite, encode each h by its index: $1, \ldots, |\mathcal{H}|$ – requires $\lceil \log_2|\mathcal{H}| \rceil$ bits for each h.

Other coding schemes possible and the result still holds ($\mathcal{H}$ not necessarily finite):

$$err_{true}(h) \leq \frac{1}{\log_2 e} \cdot \frac{size(h) + \log_2(1/\delta)}{m}$$

OCCAM'S RAZOR: SHORTER HYPOTHESES THAT EXPLAIN THE DATA HAVE BETTER GENERALIZATION

EXAMPLE:

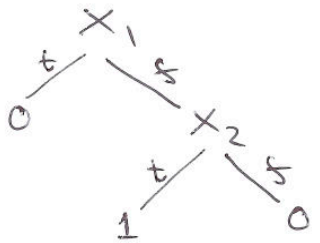
DECISION TREES FOR n BINARY ATTRIBUTES $x_0, \dots, x_{n-1}$

- define encoding recursively:

$$\text{ENCODE}(\text{leaf}) = \begin{cases} 00 & \text{if predicts } 0 \\ 01 & \text{if predicts } 1 \end{cases}$$

$$\text{ENCODE}\left(\begin{array}{c} x_i \\ t \swarrow \quad \searrow f \\ \text{LEFT} \qquad \text{RIGHT} \end{array}\right) = 1 \bullet \langle i \text{ in binary} \rangle \bullet \text{ENCODE(LEFT)} \bullet \text{ENCODE(RIGHT)}$$

SAT: ATTRIBUTES x_0, x_1, x_2, x_3



1 01 00 1 1 00 1 00
x_1 pred. x_2 pred. pred.
0 1 0

k leaves, $k-1$ inner nodes: $\text{size}(h) = 2k + (k-1)\log n$ BITS

IF WE CAN FIT DATA WITH A TREE WHICH HAS k leaves:

$$\text{err}_{\text{true}}(h) \leq \frac{1}{\log_2 e} \cdot \frac{2k + (k-1)\log_2 n + \log_2(1/\delta)}{m}$$

< 1 (BOUND NON-TRIVIAL) if $k < \frac{m}{\ln(4n/\delta)}$

LIMITATIONS OF PAC LEARNING

- consistency requirement:
  what if
  concept $\notin \mathcal{H}$
  or there is inherent noise in data?
- coding can be cumbersome
  e.g.: how to encode half-spaces?

DROP CONSISTENCY REQUIREMENT:

- m i.i.d. data points,
  but y_i not necessarily
  a deterministic function of x_i

APPROACH

- find $h \in \mathcal{H}$ that makes <u>fewest</u>
  mistakes on the data

NOW WE CAN SHOW:

$$Pr[err_{true}(h) > err_{train}(h) + \varepsilon] \leq |\mathcal{H}| \cdot e^{-2m\varepsilon^2}$$

BEFORE: THIS = 0

BEFORE: THIS = $|\mathcal{H}| \cdot e^{-m\varepsilon}$

PROVED USING THE UNION BOUND (OVER "BAD" EVENTS)
AND HOEFFDING'S INEQUALITY

Let $\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\}$

BAD EVENT j: $j = 1, \ldots, |\mathcal{H}|$

$$err_{true}(h_j) > err_{train}(h_j) + \varepsilon$$

$$Pr[err_{true}(h_j) - err_{train}(h_j) > \varepsilon] \leq ???$$

=

DETOUR: CLASSIFICATION AS COIN FLIPS

INCORRECT $\hat{=}$ heads

CORRECT $\hat{=}$ tails

$err_{true}(h_j) \hat{=}$ PROBABILITY OF HEADS, θ

$err_{train}(h_j) \hat{=}$ MAXIMUM LIKELIHOOD ESTIMATE

$$\hat{\theta} = \frac{\#\text{misclassifications}}{m}$$

HOEFFDING INEQUALITY:

$$Pr[\theta - \hat{\theta} > \varepsilon] \leq e^{-2m\varepsilon^2}$$

($\hat{\theta}$: RANDOM VARIABLE)

GENERAL HOEFFDING BOUND:

$Z_1, Z_2, \ldots, Z_m$ indep. samples from a distribution over $[0,1]$ with mean μ

$$Pr[\mu - \underbrace{\frac{1}{m}\sum_{i=1}^{m} Z_i}_{\text{empirical average}} > \varepsilon] \leq e^{-2m\varepsilon^2}$$

SUMMARIZING
INCONSISTENT HYPOTHESIS MODEL:

with prob $1-\delta$, for all $h \in \mathcal{H}$:

$$err_{true}(h) \leq err_{train}(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln(1/\delta)}{2m}}$$

or similarly as before

$$\underbrace{err_{true}(h)}_{\text{want small}} \leq err_{train}(h) + \sqrt{\frac{1}{\log 2} \cdot \frac{size(h) + \log_2(1/\delta)}{2m}}$$

fixed m, δ			
for complex hypotheses	small	large
for simpler hypotheses	large	small
	"bias"	"variance"

DECREASES TO ZERO w/ MORE DATA

NOTE: a worse dependence on # samples (need more samples) than the CONSISTENT CASE