RECAP

- PAC Learning:
  - all target concepts and distributions, accuracy & confidence parameters.
  - allow learning algorithm to output a hypothesis from a different (typically larger) hypothesis class.

    [If we insist on the output hypothesis to be from the concept class being learned, we will call it PROPER LEARNING.]
  - Efficient if the running time is polynomial in $1/\varepsilon$, $1/\delta$, n and size(c).
  - Notion of "Polynomially evaluatable" hypothesis classes

RESULTS:

- (Proper) PAC-learning algorithms for learning CONJUNCTIONS and axis-aligned RECTANGLES (in $\mathbb{R}^2$).
- Hardness of <u>proper</u> PAC-learning of 3-term DNF.
- PAC-learning algorithm (improper) for learning 3-term-DNF by using a (proper) PAC learning algorithm for 3-CNF.

Consistent Learning

Explanatory model:

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ $\quad x_i \in X$ (instance space)

$y_i \in \{0,1\}$.

"Can we find a hypothesis that is consistent with all of the given data?"

if $\underline{x} = \underline{x}_1$ then output y_1.
else if $\underline{x} = x_2$ " " y_2
?
" $\underline{x} = x_m$ " " y_m
else output 0.

works as long as there ~~are~~ is no pair (x_i, y_i) & (x_j, y_j) where $x_i = x_j$ but $y_i \neq y_j$.

$X = \{0,1\}^n$

Exercise: Show that you can find a consistent hypothesis that is a DNF formula with at most m terms.

Occam's Razor: Find a simple an explanation as possible.

"Shortest" hypothesis that is consistent with the given data.

For learning, "short enough" hypothesis will suffice.

"Explanation is shorter that the given data!

Looking for succint explanations.

[Connections to Kolmogorov Complexity; "Minimum Description Length Principle"; Bayesian interpretation]

Consistent Learner: We say that there exists a consistent learner for a concept class C using a hypothesis class H, if $\forall n \geq 1$, $\forall c \in C_n$, $\forall m \geq 1$,

given $(x_1, c(x_1)), (x_2, c(x_2)), \ldots (x_m, c(x_m))$, the a learner outputs $h \in H_n$, such that $h(x_i) = c(x_i)$ for $i = 1, \ldots, m$.

We say that there is an efficient consistent learner if the algorithm runs in time polynomial in n, size(c), m.

(EMPIRICAL RISK MINIMIZATION - using the zero-one loss)

(Occam's Razor Theorem)

Theorem: Let C be a concept class and H a hypothesis class. Let L be a consistent learner for C using h, and furthermore L is efficient. Then $\forall n \geq 1$, $\forall D$ over X_n, $\forall c \in C_n$, if L is given a sample of size m drawn from $EX(c, D)$ such that

$$m \geq \frac{1}{\varepsilon}\left(\log |H_n| + \log 1/\delta\right)$$

then L is guaranteed to output a hypothesis $h \in H_n$, that with probability at least $1-\delta$, satisfies $err(h) \leq \varepsilon$. (If $\log |H_n|$ is polynomial in n and size(c) then C is efficiently PAC learnable using H.)

Proof: Call hypothesis $h \in H_n$ "bad" if $err(h) \geq \varepsilon$.

A_h event that m random examples from $EX(c, D)$ are consistent with h.

if h is bad, then $\mathbb{P}(A_h) \leq (1-\varepsilon)^m$.

$$\mathcal{E} = \bigcup_{h \text{ bad}} A_h$$

$$\mathbb{P}(\mathcal{E}) \underbrace{\leq}_{\text{union bound}} \sum_{h \text{ bad}} \mathbb{P}(A_h) \leq |H_n| \cdot (1-\varepsilon)^m.$$

[want this to be $\leq \delta$]

Want

$$|H_n| \cdot (1-\varepsilon)^m \le \delta$$

$|H_{n,m}|$

$\cancel{f(m) = o\left(\frac{m}{\log m}\right)}$

$\cancel{m^{0.55} \quad \frac{m}{\log m}}$

It suffices to show that

$$|H_n| \cdot e^{-\varepsilon m} \le \delta$$

$(1-x) \le e^{-x}$

$\log |H_{n,m}| \le \text{poly}(n) \cdot m^{\beta}$, $\beta < 1$.

equivalent to $\quad m \ge \frac{1}{\varepsilon}\left(\underbrace{\log |H_n|}_{\log |H_{n,m}|} + \log \frac{1}{\delta}\right)$

If $\mathcal{E}$ does not occur, then the consistent learner must output $h \in H_n$, s.t. $err(h) \le \varepsilon$. $\square$

For CONJUNCTIONS: Provided $m \ge \frac{2n}{\varepsilon} \log\left(\frac{2n}{\delta}\right)$ the algorithm "works". [Proved last week]. $= \frac{2n \log(2n)}{\varepsilon} + \frac{2n}{\varepsilon} \log(1/\delta)$.

$|H_n = |\text{CONJUNCTIONS}| = 3^n$, $\quad \log |H_n| = n \log 3$.

if $m \ge \frac{1}{\varepsilon}\left(n \log 3 + \log \frac{1}{\delta}\right)$ the consistent learner gives a PAC learner.

3-term DNF

3-term DNF of size n: 2^{6n} $\quad \boxed{\log \ldots \ \Theta(n)}$

3-CNF formulae on n variables: $2^{8n^3} \longrightarrow \Theta(n^3)$.

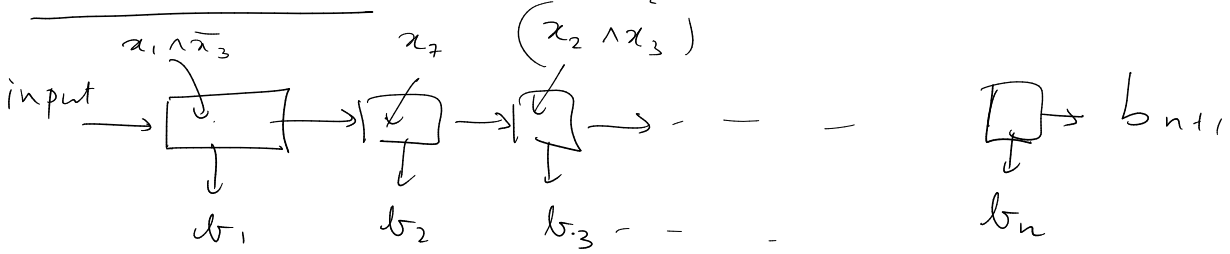
[Without worrying about computational complexity, any 3-term DNF formula consistent with $\frac{1}{\varepsilon}(c \cdot n + \log 1/\delta)$ examples has a PAC-guarantee.

(If we output a 3-CNF formula, Occam's Razor Theorem will tell us that $\frac{1}{\varepsilon}(c \cdot n^3 + \log 1/\delta)$ examples are sufficient)

---

k-Decision Lists:

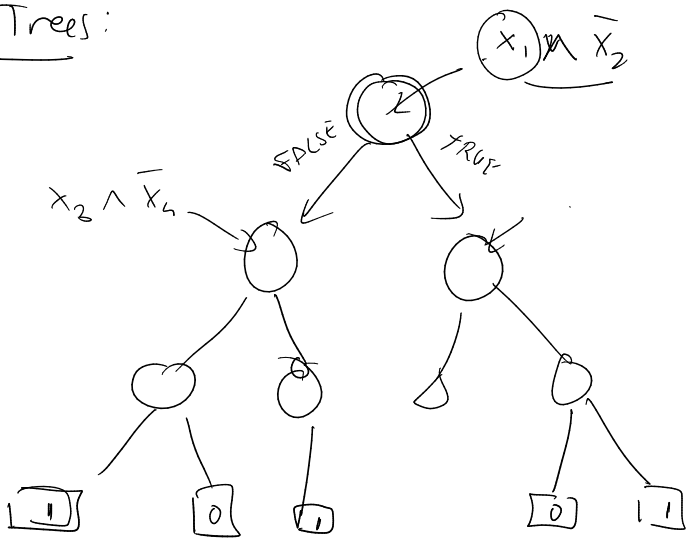
conjunctions on at most k variables.



$b_i \in \{0,1\}$.

(PAC-Learnable)

Decision Trees:



DT :

(Decision trees of poly size not PAC-learnable)