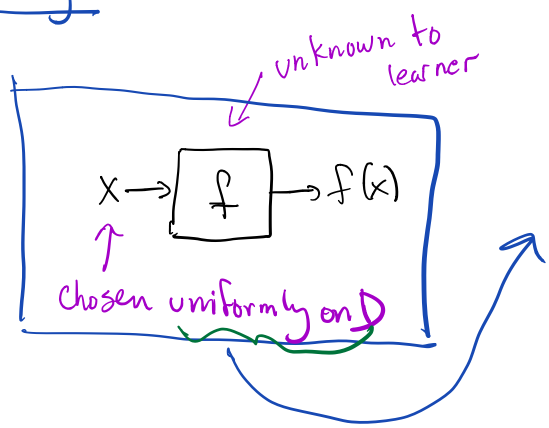
Today's lecture:

- The PAC learning model
  - motivation
  - definition
- Occam's razor
- Learning conjunctions
- (if time: begin learning via Fourier representation)

Learning how to formalize?



labelled examples

$x_1, f(x_1)$
$x_2, f(x_2)$
$\vdots$

m random labelled examples

Example oracle $Ex(f)$

Goal: • ~~output f~~ too hard?

output h s.t. f is ε-close to f

e.g. $\Pr_{x \in D}[f(x) = h(x)] \geq 1 - \varepsilon$

which distribution? today assume uniform!

<u>def</u> given hypothesis h, error of h with respect to f is $\text{error}(h) = \Pr_{x \in_u D}[f(x) \neq h(x)]$

f is ε-close to h w.r.t. uniform on D

<u>Observe</u> if f arbitrary then nontrivial learning is impossible

What if f is in a <u>class</u> of fctns. $\mathcal{C}$

<u>def</u> $\overbrace{\text{uniform distribution}}^{\text{arbitrary dist } q}$ learning algorithm for concept class $\mathcal{C}$ is algorithm $\mathcal{A}$ st.

- $\mathcal{A}$ is given $\varepsilon, \delta > 0$ access to $Ex(f)$ (according to q) for $f \in \mathcal{C}$
- $\mathcal{A}$ output h st. with prob $\geq 1-\delta$ $\underbrace{\text{error}(h) \text{ w.r.t. } f \text{ is } \leq \varepsilon}_{h \text{ is } \varepsilon\text{-close to } f}$ (according to q)

Parameters of interest

- m # samples used by $\mathcal{A}$ "sample complexity"
- ε accuracy parameter
- δ confidence parameter
- runtime   hope for $\text{poly}\left(\log(\text{domain size}), \frac{1}{\varepsilon}, \frac{1}{\delta}\right)$
- description of h? $|\mathcal{C}|$
  - similar to description of all $f \in \mathcal{C}$? (proper learning)
  - at least should be "compact" $O(\log|\mathcal{C}|)$ + efficient to evaluate

Remarks

- dependence on δ needn't be more than $O\left(\log\left(\frac{1}{\delta}\right)\right)$
- uniform dist is a special case

Occam's Razor

learning is easy!
wrt sample complexity
<u>not</u> runtime

brute force algorithm

- draw $M = \frac{1}{\varepsilon}\left(\ln|\mathcal{C}| + \ln\frac{1}{\delta}\right)$ samples
- search over all $h \in \mathcal{C}$ until find one that labels all examples correctly. Output h. (choose arbitrarily if >1)

behavior:

examples come from $f \in \mathcal{C}$
good to output f
bad to output h s.t. h & f not ε-close

h is "bad" if $\text{error}(h)$ wrt $f \geq \varepsilon$

$$\Pr[\text{bad } h \text{ consistent with examples}] \leq (1-\varepsilon)^M$$

$$\Pr[\underline{\text{any}} \text{ bad } h \text{ consistent with examples}]$$

$$\leq |\mathcal{C}| \cdot (1-\varepsilon)^M \qquad \text{union bnd}$$

$$\leq \cancel{|\mathcal{C}|} \cdot \underbrace{(1-\varepsilon)^{\frac{1}{\varepsilon}}}_{e^{-1}} {}^{(\cancel{\ln |\mathcal{C}|} + \ln \frac{1}{\delta})}$$

$$\leq \delta$$

$\Rightarrow$ unlikely to output $\underline{\text{any}}$ bad h ∎

Proof applies to learning under any distribution

Once we have a good hypothesis h:

1) can predict values of f on new $\underbrace{\text{random}}_{\text{according to } \mathcal{D}}$ inputs $\Pr_{x \in \mathcal{D}}[f(x) = h(x)] \geq 1 - \varepsilon$

2) can compress description of samples

$(x_1, f(x_1))\ (x_2, f(x_2))\ \ldots\ (x_m, f(x_m))$ — # bits: $m(\log|D| + \log|R|)$

$\Downarrow$

$x_1 \ldots x_m$, description of h — $m \cdot \log|D| + \log|\mathcal{C}|$

learning $\Rightarrow$ prediction + "compression"

Occam's Razor: simplest explanation is best

An efficient learning algorithm

$\mathcal{C}$ = conjunctions over $\{0,1\}^n$

i.e. $f(x) = x_i x_j \bar{x}_k$, $x = (x_1 \ldots x_n)$

Observe: how to distinguish

$$\left.\begin{array}{l} f(x) = x_1 \ldots x_n \\ \text{from} \\ f(x) = 0 \end{array}\right\} \text{ need } \sim 2^n \text{ samples}$$

$\Rightarrow$ can't hope for poly time & 0-error

Brute force algorithm: (i.e. alg in Occam's razor)

try each $f \in \mathcal{C}$ $\quad |\mathcal{C}| \gg 2^n$

union bnd $\Rightarrow$ need $\Omega\left(\frac{1}{\varepsilon} \ln 2^n + \ln \frac{1}{\delta}\right)$ samples

Poly time algorithm

Simplifying assumption:

assume $\Pr_{x \in \{0,1\}^n}[f(x)=1] > \varepsilon$

$\Rightarrow$ in a sample of size m, $\geq \varepsilon m$ many "positive" ($=1$) examples (in expectation)

1234	+-	f(x)
0110	+	1
1111	−	0
0101	+	1
1100	−	0

Algorithm:

Take M examples, K of which are "positive" $f(x)=1$

let $V = \{$ vars set same way in each positive example $\}$

$V = \{1, 2\}$

output $h(x) = \bigwedge_{i \in V} x_i^b$

$h(x) = \bar{x}_1 x_2$

Behavior: $f(x) = \bar{x}_1$

for $i \in$ conjunction:

must be set same way in each positive example $\Rightarrow$ in V

for $i \notin$ conjunction:

$\Pr[i \in V] \leq \Pr[i$ set same in each of K positive examples$]$

$\leq \frac{1}{2^k} + \frac{1}{2^k} = \frac{1}{2^{k-1}}$

$\Pr[$any i not in conjunction survives$]$

$\leq \frac{n}{2^{k-1}}$

$\leq \delta$ if pick $k = \log \frac{n}{\delta}$

$\Rightarrow$ $\Omega(\log \frac{n}{\delta})$ positive examples

& $\Omega(\frac{1}{\varepsilon} \log \frac{n}{\delta})$ total examples suffice.

More general algorithm:

using poly($1/\varepsilon$) samples $\left\{ \begin{array}{l} \text{estimate } Pr[f(x)=1] \text{ to additive error } \pm\frac{\varepsilon}{4} \\ \text{if estimate } < \varepsilon/2 \text{, output } h=0 \end{array} \right.$

$\Rightarrow$ $Pr[f(x)=1] \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} < \varepsilon$

good answer

O.W. continue

$\Rightarrow$ $Pr[f(x)=1] \geq \frac{\varepsilon}{2} - \frac{\varepsilon}{4} \geq \frac{\varepsilon}{4}$

$\Rightarrow$ see positive example every $\frac{4}{\varepsilon}$ samples

$\Rightarrow$ above algorithm is efficient

$\blacksquare$