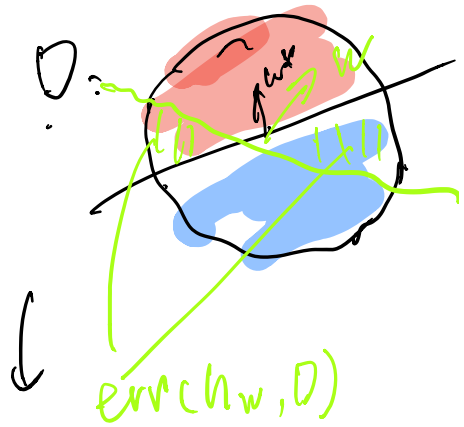
PAC learning : toy example

$\mathcal{H}$

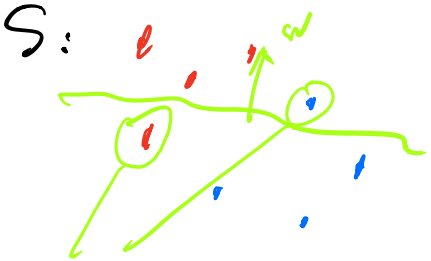
$\mathcal{X} = \mathbb{R}^2$

$\mathcal{Y} = \{\pm 1\}$

$h_w(x) = \text{sign}(w \cdot x)$



$err(h_w, D)$



$err(h_w, S) = \frac{2}{8}$

$= 0.25$

Theorem: Let $\mathcal{H}$ finite. If the consistency alg. is given m iid training examples from D realizable w.r.t. $\mathcal{H}$, then with probability $1-\delta$ ("probably") its output $\hat{h}$ is such that

$$err(\hat{h}, D) \leq \varepsilon_R(m, \delta, \ln|\mathcal{H}|) := \frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m}$$

("approximately correct")

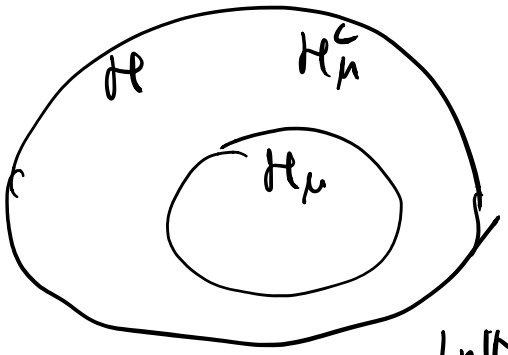
realizable

In other words. It PAC learns $\mathcal{H}$ with sample complexity

$$f(\varepsilon, \delta) = \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

1. error bound decreases polynomially with m.
2. " " depends logarithmically with $|\mathcal{H}|$
3. " " depends logarithmically with $\frac{1}{\delta}$

$\delta = 10^{-5}$



$$\mathcal{H}_\mu = \{ h \in \mathcal{H} : \underline{err(h, D)} > \mu = \frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m} \}$$

Proof: Define event

$$E = \{ \text{for all } h \in \mathcal{H}_\mu : err(h, S) > 0 \}$$

want to show $P(E) \geq 1 - \delta$. (*)

Consistency alg. returns $\hat{h}$ with $err(\hat{h}, S) = 0$.

if E happens, $\hat{h} \in \mathcal{H}_\mu^c \Rightarrow err(\hat{h}, D) \leq \mu$.

proof of (*):

$$E = \bigcap_{h \in \mathcal{H}_\mu} \{ err(h, S) > 0 \}$$

$$\overline{E} = \bigcup_{h \in \mathcal{H}_\mu} \{ err(h, S) = 0 \}$$

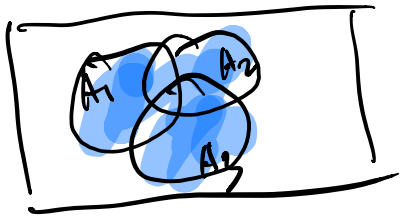
$\downarrow$

A_h

Union bound:

for any events $A_1 \dots A_n$.

$$P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} P(A_i)$$



$$P(\bar{E}) \le \sum_{h \in H_\mu} P(\text{err}(h,S) = 0)$$

given h, $\text{err}(h,D) \ge \mu$.

training example: (x_i, y_i)

$$P(\text{err}(h,S) = 0)$$

$$= P(\text{for all } i,\ h(x_i) = y_i) = \prod_{i=1}^{m} P(h(x_i) = y_i)$$

$P(h(x_i) \neq y_i)$

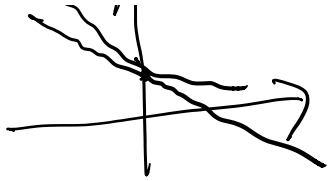
$$= \prod_{i=1}^{m} (1 - \text{err}(h,D))$$

$$\le \prod_{i=1}^{m} (1-\mu) = (1-\mu)^m.$$

$$P(\bar{E}) \le \sum_{h \in H_\mu} (1-\mu)^m$$

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

$\le |H|$



$$\leq \sum_{h \in H_\mu} e^{-m\cdot\mu}$$

$$= |H_\mu| \cdot e^{-m\mu}$$

$$\leq |H| \cdot e^{-m\mu} \quad \leftarrow \quad \mu = \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

$$= |H| \cdot e^{-(\ln(|H|/\delta))}$$

$$= |H| \cdot \frac{\delta}{|H|} = \delta.$$

$$\Rightarrow P(E) = 1 - P(\bar{E}) \geq 1 - \delta.$$

The Agnostic PAC learning model

Def: we call an algorithm A agnostic PAC learns hypothesis class H with sample complexity function $f : (0,1) \times (0,1) \to \mathbb{N}$, if for all distributions D, $\varepsilon > 0$, $\delta > 0$, when A

receives $m \geq f(\varepsilon, \delta)$ i.i.d. training examples from D as input. It outputs a classifier $\hat{h}$, such that with prob. $\geq 1-\delta$,

$$err(\hat{h}, D) \leq \min_{h' \in \mathcal{H}} err(h', D) + \varepsilon.$$

(A, $\mathcal{H}$) A PAC learns $\mathcal{H}$ ⇓ ⇑ ✓ ?

A agnostic PAC learns $\mathcal{H}$;

Algorithm for agnostic PAC learning

$|\mathcal{H}| < \infty$.

Empirical risk minimization:

Return $\hat{h}$ such that it has the smallest training error among $\mathcal{H}$: $\approx err(h, D)$

$$\hat{h} = \arg\min_{h \in \mathcal{H}} err(h, S)$$

Concentration of measure:

Given i.i.d. random variables $X_1 \cdots X_n$.

empirical mean concentrates around $\mathbb{E}[X]$

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

with large probability. (1 - fig)

Given classifier $h \in \mathcal{H}$,

$$err(h, S) \approx err(h, D)$$

$$Z_i = \mathbb{I}(h(X_i) \neq y_i) \sim \text{Bernoulli}(err(h, D))$$

$$err(h, S) = \frac{1}{m}\sum_{i=1}^{m} Z_i$$

Hoeffding

Hoeffding's Inequality:

Suppose that $Z_1 \cdots Z_n$ iid r.v. such that for each i,

$Z_i \in [a, b]$, $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, $\mu = \mathbb{E}[Z_i]$.

Then, for all $\varepsilon > 0$.

$$P\left(|\bar{Z} - \mu| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$

$$\left(\text{"}\delta\right)$$

$\mu - \varepsilon \quad \mu \quad \mu + \varepsilon$

Equivalently. $\forall \delta > 0$,

$$P\left(|\bar{Z} - \mu| > (b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \leq \delta.$$

$\delta = 10^{-5}$

$(b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}$

Comparison with Chebyshev's ineq:

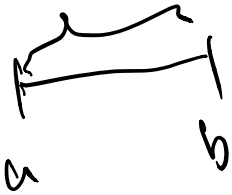
$\bar{Z}$, $\mathbb{E}[\bar{Z}] = \mu$

$$\mathbb{E}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right]$$

$$\mathrm{Var}(\bar{Z}) = \frac{1}{n} \cdot \mathrm{Var}(Z_1)$$

$$\leq \frac{1}{n} \cdot (b-a)^2$$

$$\mathbb{E}\{Z_1 - \mathbb{E}Z_1\}^2 \leq (b-a)^2$$



Chebyshev:

$$P(|\bar{Z} - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{Z})}{\varepsilon^2}$$

$$\leq \frac{(b-a)^2}{n\varepsilon^2} = \delta.$$

Equivalently. $\forall \delta > 0$.

$$P\left(|\bar{Z} - \mu| > (b-a)\cdot\sqrt{\frac{\frac{1}{\delta}}{n}} \right) \leq \delta .$$

$\delta = 10^{-5}$. dependence on δ is much worse than Hoeffding.

proof of Hoeffding's Ineq:

method of moment generating functions

$$X \qquad \phi_X(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$$

Def: random variable X is said to be σ^2-subgaussian, if $\forall \lambda \in \mathbb{R}$, $\mu = \mathbb{E}[X]$

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}$$

(If $X \sim N(\mu, \sigma^2)$, $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\frac{\sigma^2\lambda^2}{2}}$)

Next time: $[a,b]$

- bounded r.v's are subgaussian
- sum of independent subgaussian are subgaussians
- subgaussian $\Rightarrow$ light probability tail