Analysis of finite hypothesis class

<u>Setup</u>

$h_S \in ERM_{\mathcal{H}}(S)$, assume $|\mathcal{H}| < \infty$,

assume realizability ($\exists h^* \in \mathcal{H}$, $L_{D,f}(h^*) = 0$)

and have binary classification w/ 0-1 loss

so $L_{D,f}(h) = \mathbb{P}_{x \sim D}(h(x) \neq f(x))$

Goal is PAC learning: $\forall \varepsilon, \delta \in (0,1)$

prove $L_{D,f}(h_S) \leq \varepsilon$ w.p. $\geq 1-\delta$

aka $D^m(\{S : L_{D,f}(h_S) > \varepsilon\}) < \delta$

<u>Analysis</u>:

let $\mathcal{H}_B$ = "bad hypotheses"

$= \{h \in \mathcal{H} : L_{D,f}(h) > \varepsilon\} \subseteq \mathcal{H}$

so $|\mathcal{H}_B| \leq |\mathcal{H}|$

let M = Misleading samples

$= \{S : \exists h \in \mathcal{H}_B \text{ s.t. } \underbrace{h \in ERM_{\mathcal{H}}(S)}_{\text{equiv., } \hat{L}_S(h) = 0 \text{ by realizability}}\}$

Our $h_S \in ERM_{\mathcal{H}}(S)$

so if $S \notin M$ we know h_S is <u>good</u>.

How big is M?

$$M = \bigcup_{h \in \mathcal{H}_B} \{S : \hat{L}_S(h) = 0\}$$

so $D^m(M) \leq \sum_{h \in \mathcal{H}_B} D^m(\{S : \hat{L}_S(h) = 0\})$

via union bound

(aside: union bound: $D(A \cup B) \le D(A) + D(B)$ ↗"or"

Directly useful for bounding failures.

For successes, use $D(A) = 1 - D(A^c)$

and De Morgan's laws: $(\cup A_i)^c = \cap A_i^c$

$(\cap A_i)^c = \cup A_i^c$)

Each term in sum: $D^m(\{S : \hat{L}_S(h) = 0\})$, fix $h \in \mathcal{H}_B$

Since $\hat{L}_S(h) = 0$ means $h(x_i) = \underbrace{f(x_i)}_{\text{true label}} \ \forall i \in [m]$

$$D^m(\ldots) = D^m(\{S : h(x_i) = f(x_i) \ \forall i \in [m]\})$$

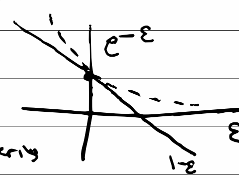
$$= \prod_{i=1}^{m} D(\{x : h(x) = f(x)\}) \quad \text{by independence}$$

$$= D(\{x : h(x) = f(x)\})^m \quad \text{since identically distr.}$$

$$= (1 - D(\{x : h(x) \neq f(x)\}))^m \quad \text{take complement}$$

$$= (1 - \underbrace{L_{D,f}(h)}_{> \varepsilon \text{ since } h \in \mathcal{H}_B})^m \quad \text{since 0-1 loss}$$

$$\le (1-\varepsilon)^m$$



$$\le e^{-\varepsilon \cdot m} \quad \text{since } (1-\varepsilon) \le e^{-\varepsilon}$$

nicer looking

proof via Taylor Series or derivative + convexity

Going back to union bound,

$$D^m(M) \le \sum_{h \in \mathcal{H}_B} e^{-\varepsilon m} = |\mathcal{H}_B| \cdot e^{-\varepsilon m} \le \underbrace{|\mathcal{H}| \cdot e^{-\varepsilon m}}_{\text{"}\delta\text{"}}$$

want $\delta = |\mathcal{H}| \cdot e^{-\varepsilon m}$, or, $-\varepsilon m = \log(\delta / |\mathcal{H}|)$

so ...

Corollary 2.3 (binary class., 0-1 loss, realizable, $|\mathcal{H}| < \infty$)

$\forall \delta \in (0,1)$ $\forall \varepsilon > 0$, if $m \in \mathbb{Z}$ satisfies

$m \geq 1/\varepsilon \log(|\mathcal{H}|/\delta)$ then $\forall f, \forall D$

(as long as realizable), if S has m iid samples then $\forall h_S \in ERM_{\mathcal{H}}(S)$, $\underbrace{\text{w. prob} \geq 1-\delta}_{\text{"probably"}}$,

$$L_{D,f}(\underbrace{h_S) \leq \varepsilon}_{\text{"approximat"}}$$

m is our "sample complexity":

smaller δ and/or smaller ε $\Rightarrow$ need more samples

Later we'll drop realizability assumption and allow for "agnostic" case, and also more general loss functions: Corollary 4.6

Then $m \geq 1/\varepsilon^2 \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$

yuck!

So, from training examples, we can learn arbitrarily well (given enough data)

... if our choice of $\mathcal{H}$ (inductive bias) was good enough to include a good classifier.

more generally, (§3), define

Def A hypothesis class $\mathcal{H}$ is "PAC learnable" if

$\exists$ function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ "sample complexity"

and some learning algo "A" (eg. $ERM_{\mathcal{H}}$)

such that

$\forall D$ over X, $\forall f : X \to \{0,1\}$ → still doing binary classif.

if $(\mathcal{H}, D, f)$ is realizable, then

if $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid samples are used in algo A,

A returns a classifier/hypothesis h s.t. w.p. $\geq 1-\delta$, $L_{D,f}(h) \leq \varepsilon$.

Aside Since S is a random variable (r.v.),

so is $X_m = L_{D,f}(A(S))$.

We're saying $0 \leq X_m \leq \varepsilon$ w.p. $\geq 1-\delta$. if $m \geq m(\varepsilon,\delta)$

How does this relate to usual notions of convergence of r.v.?

You can show that existence of a finite $m(\varepsilon,\delta)$ is

iff $\lim_{m\to\infty} \mathbb{E}_{S \sim \mathcal{D}^m} X = 0$, i.e., "$L^1$ convergence"

(see exer. 4.1)

$X_m \xrightarrow{L^p} X$ means

$\lim_{m\to\infty} \mathbb{E}|X_m - X|^p = 0$

(aside still)

and conv. in measure / probability:

$X_m \xrightarrow{P} X$ means $\forall \varepsilon > 0$

$$\lim_{m \to \infty} \mathbb{P}(\{\omega : |X_m(\omega) - X(\omega)| > \varepsilon\}) = 0$$

and almost sure (a.s.) convergence:

$\mathbb{P}(\{\omega : X_m(\omega) \to X(\omega)\}) = 1$ i.e. ptwise a.e.

Agnostic PAC learning means relaxing assumption of an oracle $f : X \to Y$

Now, let D be a joint distribution over $Z := X \times Y$
and D_X its marginal wrt X,
$D(z/x)$ its conditional.

Before, in (pure) PAC learning, we're assuming
$D(z|x)$ collapses to a deterministic function $f(x)$

$\hat{L}_S$ is unchanged but (for 0-1 loss)

turn $L_{D,f}(h) = \mathbb{P}_{x \sim D}(h(x) \neq f(x))$

into $L_D(h) = \mathbb{P}_{(x,y) \sim D}(h(x) \neq y)$ } AGNOSTIC CASE

Now, it's typically too much to ask for $L_D(h) \leq \varepsilon$

...

⚠ One algo for all distributions

Def $\mathcal{H}$ is agnostic PAC learnable if $\exists\, m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$,
$\exists$ algo A, s.t. $\forall\, \varepsilon, \delta \in (0,1)$, $\forall$ distr. D over $\mathcal{X} \times \mathcal{Y}$,
A s.t. if S has $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid samples, then
w.p. $\geq 1-\delta$

$$L_D(h) \leq \varepsilon + \underbrace{\min_{h' \in \mathcal{H}} L_D(h')}_{\text{new}}$$

where $h = A(S)$

Note: Mohri puts it this way: agnostic = inconsistent + stochastic
consistent: means $\min_{h' \in \mathcal{H}} L(h') = 0$
deterministic: means $D(y|x) = f(x)$

Remark An alternative comparison would be the Bayes Optimal Predictor, $f_D(x) = \begin{cases} 1 & \text{if } P(y=1|x) \geq 1/2 \\ 0 & \text{else} \end{cases}$
uses knowledge of D!

Optimal, but
① you have to know D
①' it changes when D changes
② it's asking too much

Loss function $\hat{L}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$
Allow general risk/loss $L_D(h) = \mathbb{E}_{z \sim D}\, \ell(h, z)$

Ex: 0-1 loss for binary classif,
$$\ell_{0-1}(h, (x,y)) = \begin{cases} 0 & h(x) = y \\ 1 & \text{else} \end{cases} = \mathbb{1}_{h(x) \neq y}(h, z)$$

Ex regression, ℓ_p losses
$$\ell_p(h, (x,y)) = (h(x) - y)^p,$$
$p \geq 1$ for convex
$p = 2$ most common

Brief History:

a lot of the framework due to
Vapnik + Chervonenkis '71

PAC due to Valiant '84

Vladimir Vapnik (1936–) from Samarkand (Uzb.) PhD '64

- '71 VC theory
- '90 move to US, AT&T, develop SVMs
  - '98 "Stat. Learning Theory" book 60,000+ citations
- '02 NEC
- '03 Columbia
- '14 Facebook AI Research, '16 Vencore Labs

Alexey Chervonenkis (1938–2014)

- '71 VC theory
- Kept researching, stayed in USSR/Russia

Leslie Valiant (1949–) from Budapest

- Prof. at Harvard since '82
- Many works in Theoretical Computer Science (TCS)
  - two sons on faculty at Stanford + Brown