Sample complexity of PAC learning

- We will deal with PAC model first.
- We deal with agnostic model later.

- Upper bound on sample complexity
- Lower bound on sample complexity

- Upper bound: If ERM is given
  $m \gtrsim \frac{VC(H) + \log(1/\delta)}{\varepsilon}$ samples
  then it (ε, δ)-learns any target
  under any distribution

- Lower bound: Any algorithm requires
  at least $m \gtrsim \frac{VC(H) + \log(1/\delta)}{\varepsilon}$ samples
  to learn some target under some
  distribution.

- lower bound $\leq m(H, \varepsilon, \delta) \leq$ upper bound

- Similar results will hold for agnostic PAC model instead of $\frac{1}{\varepsilon}$ the sample complexity will be $\frac{1}{\varepsilon^2}$.

Upper bounds

- We define notions of

ε-net and ε-approximation (ε-sample)

Definition Let S be a finite subset of a domain X. Let $H \subseteq \mathcal{P}(X)$ and D be probility distribution over X. Let $\varepsilon \in [0,1]$.

- S is called ε-net for D, H iff

$$\forall h \in H \quad D(h) > \varepsilon \Rightarrow S \cap h \neq \emptyset$$

- S is called ε-approximation (or ε-sample) for D, H iff

$$\forall h \in H \quad \left| D(h) - \frac{|S \cap h|}{|S|} \right| \le \varepsilon$$

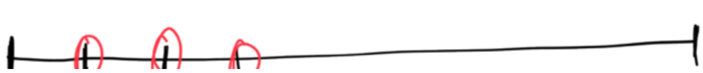
- ($S \cap h$ is interpreted as a multi-set.)

- If S is an ε-approximation then S is an ε-net.

- The reverse implication is not true.

If $X = \mathbb{R}$ and D is uniform distribution over $[0, 1)$ then

$S = \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$ is $\frac{1}{n}$-approximation with respect to the class of all intervals.



$0 \quad \frac{1}{n} \quad \frac{2}{n} \quad \cdots$

- For given $h: X \to Y$ and distribution D over X, consider the distribution D_h over $X \times Y$ defined as

$$D_h(x,y) = \begin{cases} D_h(x) & \text{if } y = h(x) \\ 0 & \text{if } y \neq h(x). \end{cases}$$

- Graph $h: X \to Y$ is

$$g(h) = \{(x,y) \in X \times Y : h(x) = y\}$$

- $g^c(h)$ is the complement of $g(h)$
- $G(H) = \{g(h) : h \in H\}$
- $G^c(H) = \{g^c(h) : h \in H\}$

<u>Lemma</u>: Let X be a domain. Let D be a distribution over X. Let $H \subset Y^X$ and let $h^* \in H$

If $S \subseteq X \times Y$ is an ε-net for $D_{h^*}, \mathcal{G}^c(H)$ and $\widehat{err}_S(h^*) = 0$ then

$$err_{D,h^*}(ERM(S)) \leq \varepsilon.$$

Proof:

$$err_{D,h^*}(h) = \Pr_{x \sim D}[h^*(x) \neq h(x)]$$

$$= \Pr_{(x,y)\sim D_{h^*}}[h(x) \neq y]$$

$$= D_{h^*}(\{(x,y) \in X \times Y : h(x) \neq y\})$$

$$= D_{h^*}(g_h^c)$$

Consider any $h \in H$.

- If $err_{D,h^*}(h) > \varepsilon$ then

$$D_{h^*}(g_h^c) > \varepsilon.$$

- If $D_{h^*}(g_h^c) > \varepsilon$ then
  $g_h^c \cap S \neq \emptyset$
- If $g_h^c \cap S \neq \emptyset$ then
  $\widehat{err}_S(h) > 0$.
- If $\widehat{err}_S(h) > 0$ then ERM will not output h.

So if $err_{D,h^*}(h) > \varepsilon$, ERM will NOT output h.

Thus ERM outputs $h \in H$ such that $err_{D,h^*}(h) \leq \varepsilon$.

$\blacksquare$

<u>Lemma</u>: If $H \subseteq Y^X$ and $F \subseteq \mathcal{P}(X)$

- $VC(H) = VC(G(H))$
- $VC(F) = VC(F^c)$

<u>Proof</u>:

- If $(x,0), (x,1) \in S$ then $G(H)$ cannot shatter S.
- $S \subseteq X \times Y$ is shattered iff $S|_X$ is shattered
- The second part is trivial

$\blacksquare$

<u>Theorem</u>: Let $H \subseteq \mathcal{P}(X)$ with

$d = VC(H)$ where $1 \le d < \infty$.

Let $\varepsilon \in (0, \frac{1}{4})$, $\delta \in (0,1)$.

Let S be an i.i.d sample from a distribution $\mathcal{D}$ over X of size

$$m \ge \max\left(\frac{4}{\varepsilon} \log_2 \frac{2}{\delta}, \frac{8d}{\varepsilon} \log_2 \frac{8d}{\varepsilon} \right).$$

Then, with probability at least $1-\delta$, S is an ε-net for $H, \mathcal{D}$.

Proof:

Let's define event

$$E = \{ S \in X^m : \exists h \in H \text{ s.t. } \mathcal{D}(h) > \varepsilon \text{ and } S \cap h = \emptyset \}$$

We need to prove $\Pr[E] \le \delta$.

- Consider an i.i.d. sample T from D of size m, independent of S.
- Sample T is called <u>ghost sample</u>

Let's define event

$$F = \{(S,T) \in X^{2m} : \exists h \in H \text{ s.t. } D(h) > \varepsilon,\ h \cap S = \emptyset,\ |h \cap T| \geq \frac{\varepsilon m}{2}\}$$

- Clearly $F \subseteq E$, thus $\Pr[F] \leq \Pr[E]$

Claim 1:

$$\Pr[F] \geq \frac{1}{2} \Pr[E]$$

<u>Proof:</u>

- It suffices to prove $\Pr[F \mid E] \geq \frac{1}{2}$.

- If E occurs, there is $h \in H$ s.t. $D(h) > \varepsilon$ and $S \cap h = \emptyset$.

$$\Pr[F \mid E] \geq \Pr\left[|h \cap T| \geq \frac{\varepsilon m}{2}\right]$$

- $|h \cap T|$ is a binomial random variable $B(p, m)$ where $p = D(h) \geq \varepsilon$.

$P[|h \cap T| < \varepsilon m$ ...

$\bullet \; \Pr\left[|h \cap I| < \frac{\cdot}{2}\right]$

$$\leq \Pr\left[|h \cap T| < \frac{pm}{2}\right]$$

$$\leq \Pr\left[\Big|\,|h \cap T| - pm\,\Big| > \frac{pm}{2}\right]$$

$$\leq \frac{\operatorname{Var}(|h \cap T|)}{\frac{1}{4}p^2m^2} \qquad \text{(Chebyshev)}$$

$$\leq \frac{p(1-p)m}{\frac{1}{4}p^2m^2}$$

$$= \frac{4(1-p)}{pm}$$

$$\leq \frac{4}{pm} \qquad \boxed{1-p \leq 1}$$

$\leq \frac{4}{\varepsilon m}$ $\boxed{\varepsilon \leq p}$

$\leq \frac{1}{2d \log_2\left(\frac{8d}{\varepsilon}\right)}$

$\leq \frac{1}{2}$ $\boxed{\boxed{d \geq 1}}$

- So $\Pr[E] \leq 2 \Pr[F]$

Claim 2:

$$\Pr[F] \leq \Pi_H(2m) \cdot 2^{-\frac{\varepsilon m}{2}}$$

Proof:

- We can think of drawing samples S, T as follows:

Draw sample R of size $2m$.

$$R = \underbrace{\boxed{\qquad\qquad\qquad\qquad\qquad\qquad}}_{2m}$$

Flip m fair coins.

- If i-th coin falls heads, put x_i into S and x_{i+m} into T.
- If i-th coin falls tails, put x_i into T and x_{i+m} into S.

- We condition on the choice of R.

- It suffices to prove

$$\Pr[F \mid R] \le \Pi_H(2m) \cdot 2^{-\frac{\varepsilon m}{2}}$$

- Note there are only <u>finitely</u> many behaviors of H on R!

- $$F = \bigcup_{\substack{h \in H \\ D(h) > \varepsilon}} \left\{ S, T \in X^m : S \cap h = \emptyset,\ |T \cap h| \ge \frac{\varepsilon m}{2} \right\}$$

- $$H' = \{ h \in \Pi_H(R) : \exists h' \in H \text{ s.t. } D(h') > \varepsilon,\ h' \cap R = h \cap R \}$$

- Note that $H' \subseteq \Pi_H(R)$

- For $h \in H'$ define

$$F_h = \{ S, T \in X^m : S \cap h = \emptyset, |T \cap h| \geq \frac{\varepsilon m}{2} \}$$

- $\Pr[F \mid R]$

$$= \Pr\left[\bigcup_{h \in H'} F_h \;\middle|\; R \right]$$

$$\leq \sum_{h \in H'} \Pr[F_h \mid R]$$

- It suffices to show

$$\Pr[F_h \mid R] \leq 2^{-\frac{\varepsilon m}{2}}$$

since $|H'| \leq |\Pi_H(R)| \leq \Pi_H(2m)$

- If $|h \cap R| < \frac{\varepsilon m}{2}$

  $\Pr[F_h \mid R] = 0$

- Suppose $|h \cap R| \geq \frac{\varepsilon m}{2}$

- Let $a = |h \cap R|$

$$\Pr[F_h \mid R] \leq \Pr[S \cap h = \emptyset \mid R]$$

$$= 2^{-a}$$

$$\leq 2^{-\frac{\varepsilon m}{2}}$$

We have

$$\Pr[E] \le 2\Pr[F]$$
$$\le 2\,\Pi_H(2m)\cdot 2^{-\frac{\varepsilon m}{2}}$$
$$\le 2\left(\frac{2em}{d}\right)^{d}\cdot 2^{-\frac{\varepsilon m}{2}}$$

We need to show

$$2\left(\frac{2em}{d}\right)^{d}\cdot 2^{-\frac{\varepsilon m}{2}} \le \delta$$

It suffices to show

$$\frac{2}{\delta}(2me)^{d} \le 2^{\frac{\varepsilon m}{2}}$$

$$\log_2\left(\frac{2}{\delta}\right) + d\log_2(2em) \le \frac{\varepsilon m}{2}$$

By assumption $\log\left(\frac{2}{\delta}\right) \le \frac{\varepsilon m}{4}$.

It suffices to prove

$$d\log_2(2em) \le \frac{\varepsilon m}{4} \qquad (*)$$

$$\Updownarrow$$

$$\frac{m}{\log_2(2em)} \ge \frac{4d}{\varepsilon}$$

Since left-hand side is increasing in m, it suffices to check $(*)$ for

$$m = \frac{8d}{\varepsilon}\log_2\frac{8d}{\varepsilon}$$

$$d \log_2(2em) \leq \frac{\varepsilon}{4} m$$

$$\Updownarrow$$

$$d \log_2\left(\frac{16ed}{\varepsilon} \log \frac{8d}{\varepsilon}\right) \leq \frac{\varepsilon}{4} \cdot \frac{8d}{\varepsilon} \cdot \log_2\left(\frac{8d}{\varepsilon}\right)$$

$$\Updownarrow$$

$$\log_2\left(\frac{16ed}{\varepsilon} \log_2 \frac{8d}{\varepsilon}\right) \leq 2 \log_2\left(\frac{8d}{\varepsilon}\right)$$

$$\Updownarrow$$

$$\frac{16ed}{\varepsilon} \log_2 \frac{8d}{\varepsilon} \leq \left(\frac{8d}{\varepsilon}\right)^2$$

$$\Updownarrow$$

$$2e \log_2 \frac{8d}{\varepsilon} \leq \frac{8d}{\varepsilon}$$

$$\Updownarrow$$

$$2e \log_2 A \leq A$$

which is true for $A \geq 32$

and clearly $A = \frac{8d}{\varepsilon} \geq \frac{8}{\frac{1}{4}} = 32$

$\blacksquare$