Learning Bound based on VC dimension

Recall $\Pi_{\mathcal{H}}(m) \leq \frac{ed}{d}\left(\frac{em}{d}\right)^d$

Thm Let $\mathcal{H}$ be a hypothesis set with VC dimension d. Let L be a consistent learner for $\mathcal{H}$ that outputs $h \in \mathcal{H}$. Then L is a PAC-learning algorithm for $\mathcal{H}$ provided it is given as input a sample of size m drawn from X with distribution $\mathcal{D}$ where $m \geq K_0\left[\frac{1}{\varepsilon}\log\frac{1}{\delta} + \frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\right]$ for some universal constant K_0.

Proof Let h, h' be two hypotheses from $\mathcal{H}$. We define

$$(h \Delta h')(x) = \begin{cases} 1 & \text{if } h(x) \neq h'(x) \\ 0 & \text{if } h(x) = h'(x) \end{cases}$$

If h is the "real" hypothesis we use h' then $\operatorname{err}(h') = \mathbb{P}_{x \sim \mathcal{D}}\left[(h \Delta h')(x) = 1\right]$

Define $\Delta_h(\mathcal{H}) = \{h' \Delta h \mid h' \in \mathcal{H}\}$

Claim $VC(\mathcal{H}) = VC(\Delta_h(\mathcal{H}))$

Fix any finite $S \subseteq X$. Let $k \in \Pi_{\mathcal{H}}(S)$

Map k to $\hat{k} := \hat{k} \Delta h \in \Pi_{\Delta_h(\mathcal{H})}(S)$

This map is a bijection (routine case check) so $|\Pi_{\mathcal{H}}(S)| = |\Pi_{\Delta_h(\mathcal{H})}(S)|$ for any S.

Thus if $\mathcal{H}$ can shatter S so can $\Delta_h(\mathcal{H})$ & vice versa. Thus we have $VC(\mathcal{H}) = VC(\Delta_h(\mathcal{H}))$.

Now we define the bad hypotheses wrt $\epsilon > 0$

$$\Delta_{h,\epsilon}(\mathcal{H}) = \{\hat{h} \in \Delta_h(\mathcal{H}) \mid \mathbb{P}_{x \sim \mathcal{D}}[\hat{h}(x) = 1] \geq \epsilon\}$$

If we pick a hypothesis h' & $h' \Delta h \in \Delta_{h,\epsilon}(\mathcal{H})$ then h' will give an unacceptable error.

We say S is an ϵ-net for $\Delta_h(\mathcal{H})$ if $\forall\, \hat{h} \in \Delta_{h,\epsilon}(\mathcal{H})\ \exists\, x \in S$ s.t $\hat{h}(x) = 1$.

Now we are in the realizable case so it is always true that some hypothesis exists to give an empirical error (loss, risk) of 0 & we have a consistent learning algorithm so we find this hypothesis. Thus our algorithm would never choose a concept in $\Delta_{h,\epsilon}(\mathcal{H})$ if S is an ϵ-net.

→ So we need to bound the probability that S is not an ϵ-net.

Draw the sample in 2 stages of size m. We will view the first sample as training & the second as testing.

S_1 : first sample of size m.

Let A be the event : S_1 is not an ϵ-net.

If A happens then ~~$\exists x \in S_1$ s.t.~~ $\exists\, \tilde{h} \in \Delta_{h,\epsilon}(\mathcal{H})$ s.t. $\forall x \in S_1\ \ \tilde{h}(x) = 0$.

Fix $\tilde{h}$ & draw a second sample S_2 of size m.

Now we will aim to get a lower bound on the number of elements of S_2 s.t. $\tilde{h}(x) = 1$.

Let X_i be the random variable

$$X_i = \begin{cases} 1 \text{ if } i^{th} \text{ element of } S_2 \text{ satisfies } \tilde{h}(x)=1 \\ 0 \text{ otherwise} \end{cases}$$

$$X = \sum_{i=1}^{m} X_i$$

$$\mathbb{P}_{x\sim D}[X_i = 1] \geq \varepsilon \text{ since } \tilde{h} \in \Delta_{h,\varepsilon}(\mathcal{H}).$$

So $\mathbb{E}[X] \geq m\varepsilon$

$$\mathbb{P}[X < \varepsilon m/2] \leq \mathbb{P}[|X - \mathbb{E}[X]| > \frac{\mathbb{E}[X]}{2}]$$

So we can use Chernoff bounds

$$\mathbb{P}[|X-\mu| > \alpha\mu] \leq 2\exp[-\mu\alpha^2/3]$$

& obtain

$$\mathbb{P}[X < \varepsilon m/2] \leq 2e^{-\varepsilon m/12}$$

So if $\varepsilon m \geq 24$ the probability that at least $\varepsilon m/2$ points in S_2 satisfy $\tilde{h}(x)=1$ is at least $1/2$.

Now we define a new event B:

A sample $S = S_1 \cup S_2$ of size $2m$ is drawn with $|S_1| = |S_2| = m$ according to D, and there exists $\tilde{h} \in \Pi_{\Delta_{H,\varepsilon}(\mathcal{H})}(S)$ such that

$$|\{x \in S \mid \tilde{h}(x)=1\}| \geq \varepsilon m/2 \;\&\; \tilde{h}(x)=0 \;\forall x \in S_1.$$

Note B subsumes A so $\mathbb{P}[A \wedge B] = \mathbb{P}[B]$

& hence $\mathbb{P}[B] = \mathbb{P}[B|A] \cdot \mathbb{P}[A]$

Thus we have calculated $\mathbb{P}[B|A]$ already it is $\frac{1}{2} \geq \frac{1}{2}$.

Thus $\mathbb{P}[A] \leq 2\,\mathbb{P}[B]$

But we can bound $\mathbb{P}[B]$ by pure combinatorics.
We are given $2m$ balls out of which $r (\geq \varepsilon m/2)$ are red & the rest are black.
If we divide the set into two equal disjoint subsets what is the probability that the first set has no red balls and the second set has all of them?

$$\binom{m}{r} \Big/ \binom{2m}{r} \leq \left(\frac{1}{2}\right)^r$$

Thus $\mathbb{P}[A] \leq 2 \cdot \mathbb{P}[B] \leq$

$$\leq 2 \cdot \left| \Pi_{\Delta_{h_\varepsilon}(H)}(S) \right| \cdot 2^{-\varepsilon m/2}$$

UNION BOUND

$$\leq 2 \cdot \left| \Pi_{\Delta_h(H)}(S) \right| \cdot 2^{-\varepsilon m/2}$$

$$\leq 2 \cdot \left(\frac{2em}{d}\right)^d \cdot 2^{\varepsilon m/2}$$

Now collecting factors ind. of d & m into K_0 & setting this bound $= \delta$ & solving for m we get that if $m \geq K_0 \left(\frac{1}{\varepsilon}\log\frac{1}{\delta} + \frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\right)$

$$\mathbb{P}[A] < \delta$$