EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 10: PAC Learning Lower Bounds

Lecturer: Jacob Abernethy *Scribes: Ryen Krusinga; Editor: Peng Liao*

10.1 Review of Sauer's Lemma

Sauer's Lemma says that given a concept class $\mathcal{C}$ with VC-dimension d, we have

$$\Pi_{\mathcal{C}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = O(m^d)$$

Proof Sketch: Take $\mathcal{C}|_S$ with $S \subseteq X$, $|S| = m$.

	x_1	x_2	...	x_m
n_1	0	1	1	0
n_2		0	1	
$\vdots$		1	0	
		0	0	

Table 1: Shifting Table for Sauer's Lemma

Steps:

1. Modify the table using "shifting" until no more shifting possible.
2. Show three claims
   (a) The number of unique rows is the same after shifting
   (b) The shifting operation did not increase the VC-dim of the table
   (c) If a row contains columns $i_1, ..., i_k$ with 1s, then those columns are shattered in the table.

Conclusion: there are no more than d 1s in any row, hence the number of rows is at least the number of subsets of $[m]$ of size at most d. ■

10.2 Big Theorem

Theorem 10.1. *Let $\mathcal{C}$ be a class with VC-dim d. Given any consistent learning algorithm $\mathcal{A}$ that returns $h_S \in \mathcal{C}$ on sample $S \sim D^m$, there is a constant c_0, such that $R(h_s) \leq \epsilon$ with probability at least $1 - \delta$ as long as*

$$m \geq c_0 \left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon} \right)$$

Proof: Use the two-sample trick. For any class $\mathcal{C}$ we have

$$\Pr[R(h_S) > \epsilon] \leq \Pi_{\mathcal{C}}(2m) e^{-m\epsilon/4} + e^{-m\epsilon/8}$$

It suffices to bound each term on the right hand side by $\delta/2$. This is achieved when

$$m \geq \frac{8\log(2/\delta)}{\epsilon} \tag{10.1}$$

for the second term, $e^{-m\epsilon/8}$. Now, Sauer's lemma says that

$$\Pi_{\mathcal{C}}(m) \leq m^d$$

so the first sum term is upper bounded by

$$(2m)^d e^{-m\epsilon/4} \leq \delta/2$$

Taking log on both sides and solving for m here, we get

$$m \geq 4\left(\frac{d\log 2m + \log(2/\delta)}{\epsilon}\right) \tag{10.2}$$

It is easy to check that the inequalities 10.1 and 10.2 are satisfied when

$$m = c_0 \frac{d\log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

for some constant c_0, as desired. ∎

By solving for ϵ, we establish the following corollary:

Corollary 10.2. *When m is fixed, we can guarantee an error rate*

$$\epsilon = 8\left(\frac{d\log m + \log(1/\delta)}{m}\right)$$

10.3 Lower Bounds

One might ask whether we can get error rate ϵ with $m = \sqrt{d/\epsilon}$, or possibly $\log d/\epsilon$. The answer is no, and we can construct "hard" (counter) examples of this. That is, there exist $\mathcal{X}, \mathcal{C}, \mathcal{D}$, where one needs at least $O(d/\epsilon)$ sample to even have 50% chance of error ϵ.

Two tricks:

1. Clearly, if VC-dim$(\mathcal{C}) = d$, you need d samples to learn the target. Let U be the set shattered by $\mathcal{C}$. Let $\mathcal{D}$ be a uniform distribution on U. Let $\epsilon = \frac{1}{2d}$ (the 2 is somewhat arbitrary). Then to get ϵ error we need a sample S to contain *all* U before $R(h_s) \leq \epsilon$, $|S| \geq d$.

2. Also, to achieve error ϵ, we also need $m \geq O(1/\epsilon)$. Let $\mathcal{X} = \{x_0, x_1\}$. Let $\mathcal{C}$ be all functions $\mathcal{X} \to \{0, 1\}$. Let $\mathcal{D}$ be the distribution on $\{x_0, x_1\}$ where $Pr[\{x_0\}] = 1 - 2\epsilon$ and $Pr[\{x_1\}] = 2\epsilon$.

$$\begin{aligned}\Pr\left(R(h_s) \geq \epsilon\right) &\geq \Pr\left(x_1 \text{ not observed after } m \text{ samples}\right)\\ &= (1-2\epsilon)^m\end{aligned}$$

   Set $m = 1/(2\epsilon)$ and we get $(1-2\epsilon)^{1/(2\epsilon)} \approx 1/e$.

Putting it all together Let $\mathcal{X} = \{x_0, x_1, ..., x_{d-1}\}$, where the VC-dim of $\mathcal{C}$ is d, so that $\mathcal{C}$ shatters $\mathcal{X}$. Construct a distribution $\mathcal{D}$ over $\mathcal{X}$ such that $Pr[\{x_0\}] = 1 - 4\epsilon$, and for all $i \geq 1$, $Pr[\{x_i\}] = 4\epsilon/(d-1)$. Fact: to ensure error rate $\leq \epsilon$, need to see half of the $d-1$ rare points. So

$$\Pr\left(R(h_s) \leq \epsilon\right) \leq \Pr\left(|S - \{x_0\}| \geq \frac{d-1}{2}\right) \tag{*}$$

Aside Let

$$Z_i = \begin{cases} 1 & \text{w.p. } \epsilon \\ 0 & \text{w.p. } 1-\epsilon \end{cases}$$

We know that

$$\Pr\left(\sum_{i=1}^{m} Z_i = 0\right) \le e^{-m\epsilon}$$

and we saw

$$\Pr\left(\sum_{i=1}^{m} Z_i \le \frac{\epsilon m}{2}\right) \le e^{-m\epsilon/4}$$

I need

$$\Pr\left(\sum_{i=1}^{m} Z_i \ge 2\epsilon m\right) \le e^{-m\epsilon/3}$$

(challenge: prove this).

Now, let $Z_i = 1$ if sample i was rare, 0 if not. (So $Pr[Z_i = 1] = 4\epsilon$). Let $m = \frac{d-1}{16\epsilon}$. So

$$\begin{aligned}(*) &\le \Pr\left(\sum Z_i \ge \frac{d-1}{2}\right) = \Pr\left(\sum_{i=1}^{m} Z_i \ge 2\mathbb{E}\sum Z_i\right) \\ &\le e^{-m(4\epsilon)/3} = e^{-(d-1)\epsilon/(16\epsilon)} = e^{-(d-1)/16}\end{aligned}$$

(since $\mathbb{E}\sum Z_i = 4\epsilon m = \frac{4\epsilon(d-1)}{16\epsilon} = \frac{d-1}{4}$).

Given d, this is some constant less that $1 - \frac{1}{K}$ where you can have $K = 100, 1000$, etc. This shows that if you choose m samples for m as above, then you have have a reasonable probability of having a high error rate.