CS6781 - Theoretical Foundations of Machine Learning

Lecture 5: PAC Sample Complexity Proof

February 4, 2020

Lecturer: Nika Haghtalab *Readings: Chapter 6.4-6.5, UML*

??

1 Overview

We spent the last two lecture learning about the growth function, VC dimension, the relationship between them, and the following theorem. In this lecture, we formally prove these results.[1]

Theorem 1.1 (PAC Learnability of Infinite Concept Classes)**.** *Let $\mathcal{A}$ be an algorithm that learns a concept class $\mathcal{C}$ in the consistency model. Then, $\mathcal{A}$ learns the concept class $\mathcal{C}$ in the PAC learning model using a number of samples that satisfies*

$$m \geq \frac{2}{\epsilon}\left(\log_2(\Pi_{\mathcal{C}}(2m)) + \log_2(\frac{2}{\delta})\right).$$

2 Proof of Theorem 1.1

In this lecture, we define and work with three "bad" events. First, is the actual failure event, as a function of the training set $S \sim \mathcal{D}^m$, we would like to bound:

$$B(S) : \exists h \in \mathcal{C} \text{ such that } \mathrm{err}_S(h) = 0 \text{ and } \mathrm{err}_{\mathcal{D}}(h) > \epsilon.$$

Second, for the sake of analysis we also consider an independently drawn sample set $S' \sim \mathcal{D}^m$. We define the following event that is a function of S and S'.

$$B'(S, S') : \exists h \in \mathcal{C} \text{ such that } \mathrm{err}_S(h) = 0 \text{ and } \mathrm{err}_{S'}(h) > \frac{\epsilon}{2}.$$

Lastly, given two sample sets $S = \{x_1, \dots, x_m\}$ and $S' = \{x'_1, \dots, x'_m\}$, and a vector $\vec{\sigma} \in \{-1, +1\}^m$, we swap the members of S and S' as follows: For each $i \in [m]$, if $\sigma_i = +1$, we let $z_i = x_i$ and $z'_i = x'_i$, otherwise, we let $z_i = x'_i$ and $z'_i = x_i$. Then, let $T = \{z_1, \dots, z_m\}$ and $T' = \{z'_1, \dots, z'_m\}$. Given S, S', and $\vec{\sigma}$, we define the following bad event:

$$B''(S, S', \vec{\sigma}) : \exists h \in \mathcal{C} \text{ s.t., } \mathrm{err}_T(h) = 0 \text{ and } \mathrm{err}_{T'}(h) > \frac{\epsilon}{2}, \text{ where } T \text{ and } T' \text{ correspond to } S, S', \vec{\sigma}.$$

[1] Several proofs of Theorem 1.1 are known in the literature. The proof approach we cover is similar to that outlined by Robert Schapire's lecture notes.

When representing the probability of these events, we typically take $S \sim \mathcal{D}^m$, $S' \sim \mathcal{D}^m$, and $\sigma_i = +1$ or -1 with probability $1/2$ for all $i \in [m]$, all independently. When it is clear from the context, we suppress S, S', and $\vec{\sigma}$ in the statement of the probabilities.

To prove Theorem 1.1, it suffices to show that $\Pr_{S\sim\mathcal{D}}[B(S)] \leq \delta$. We do this by first bounding the probability of event B in terms of B' and then in terms of B''. We then argue that because B'' only depends on the empirical error on T and T' and not the true error, we can union bound only on the number of unique labelings produced on T and T', which is bounded by the growth function.

Claim 2.1. *If $m > \frac{8}{\epsilon}$, then*

$$\Pr_{S,S'\sim\mathcal{D}^m}[B'(S,S') \mid B(S)] \geq \frac{1}{2}.$$

Proof. Suppose $B(S)$ holds. Then take an h that is consistent with S, i.e., $\mathrm{err}_S(h) = 0$, and $\mathrm{err}_\mathcal{D}(h) > \epsilon$. Since S' is drawn i.i.d. from $\mathcal{D}$,

$$\mathbb{E}_{S'\sim\mathcal{D}^m}[\mathrm{err}_{S'}(h)] = \mathrm{err}_\mathcal{D}(h) > \epsilon.$$

Furthermore, $\mathrm{err}_{S'}(h)$ is the sample average of m i.i.d. bernoulli variables. Recall that Chernoff bound states that for $X_1, \ldots, X_m$ bernoulli random variables with expectation μ,

$$\Pr\left[\frac{1}{m}\sum_{i\in[m]} X_i \leq \frac{\mu}{2}\right] \leq \exp\left(-m\mu/8\right)$$

Replacing $\mu > \epsilon$, we have that $\Pr[\mathrm{err}_{S'}(h) \leq \epsilon/2] \leq \frac{1}{2}$. This proves the claim. $\square$

Note that Claim 2.1 immediately implies that $\Pr_{S\sim\mathcal{D}^m}[B(S)] \leq 2\Pr_{S,S'\sim\mathcal{D}^m}[B'(S,S')]$, because

$$\frac{\Pr[B'(S,S')]}{\Pr[B(S)]} \geq \frac{\Pr[B'(S,S') \cap B(S)]}{\Pr[B(S)]} = \Pr[B'(S,S') \mid B(S)].$$

Therefore, it suffices to bound $\Pr_{S,S'\sim\mathcal{D}^m}[B'(S,S')]$.

Claim 2.2. *For i.i.d. sample sets $S \sim \mathcal{D}^m$ and $S' \sim \mathcal{D}^m$, and a vector $\vec{\sigma}$, where $\sigma_i = +1$ or -1 with probability $1/2$ for all $i \in [m]$ independently, we have*

$$\Pr_{S,S'}\left[B'(S,S')\right] = \Pr_{S,S',\vec{\sigma}}\left[B''(S,S',\vec{\sigma})\right].$$

Proof. This is true because (T, T') and (S, S') are identically distributed. $\square$

Claim 2.3. *For any $S, S' \in \mathcal{X}^m$ and any h that is fixed (independently of $\vec{\sigma}$), we have*

$$\Pr_{\vec{\sigma}}\left[\mathrm{err}_T(h) = 0 \text{ and } \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} \mid S, S'\right] \leq 2^{-m\epsilon/2}$$

Proof. Consider the predictions of h on S and S' as follows.

$$h(x_1), h(x_2), \ldots, h(x_m)$$
$$h(x'_1), h(x'_2), \ldots, h(x'_m)$$

First, note that if there is a column with both predictions wrong then $\mathrm{err}_T(h) = 0$ can never happen, and the desired probability would be 0. Similarly, if more than $(1 - \frac{\epsilon}{2})m$ of the columns have both predictions right, $\mathrm{err}_{T'}(h) \leq \epsilon/2$, so again the desired probability would be 0. Thus, at least $r \geq m\epsilon/2$ columns have one correct and one incorrect prediction. If $\mathrm{err}_T(h) = 0$, it must happen that in all such columns, σ_i must ensure that the right prediction goes to T and the wrong one goes to T'. This happens with probability at most $2^{-r} \leq 2^{-m\epsilon/2}$. □

Claim 2.4. *For any $S, S' \in \mathcal{X}^m$,*

$$\Pr_{\vec{\sigma}}\left[\exists h \in \mathcal{C}, \mathrm{err}_T(h) = 0 \textit{ and } \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} \mid S, S'\right] \leq \Pi_{\mathcal{C}}(2m)2^{-m\epsilon/2}$$

Proof. Given a set S, define $\mathcal{C}'(S) \subseteq \mathcal{C}$ to be a set of size $|\mathcal{C}[S]|$ where we choose one (representative) hypothesis for each different labelings of $\mathcal{C}$ on S.

$$\begin{aligned}
L.H.S &= \Pr_{\vec{\sigma}}\left[\exists h \in \mathcal{C}, \mathrm{err}_T(h) = 0 \text{ and } \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} \mid S, S'\right] \\
&= \Pr_{\vec{\sigma}}\left[\exists h \in \mathcal{C}'(S \cup S'), \mathrm{err}_T(h) = 0 \text{ and } \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} \mid S, S'\right] \\
&\leq \sum_{h \in \mathcal{C}'(S \cup S')} \Pr_{\vec{\sigma}}\left[\mathrm{err}_T(h) = 0 \text{ and } \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} \mid S, S'\right] \\
&\leq \Pi_{\mathcal{C}}(2m)2^{-m\epsilon/2} \qquad \text{(Claim 2.3)}
\end{aligned}$$

□

Putting together Claims 2.1, 2.2, 2.3, and 2.4, it suffices to find m such that

$$2\Pi_{\mathcal{C}}(2m)2^{-m\epsilon/2} \leq \delta,$$

this gives us $m \geq \frac{2}{\epsilon}\left(\log_2(\Pi_{\mathcal{C}}(2m)) + \log_2(\frac{2}{\delta})\right)$.

3 Sauer's Lemma

In the last lecture, we demonstrated the importance of the following lemma.

Lemma 3.1 (Sauer's Lemma)**.** *Consider any hypothesis class $\mathcal{C}$ and let $d = \mathrm{VCDim}(\mathcal{C})$. For all m,*

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

In this lecture, we derive the proof of this lemma.

Proof of Sauer's Lemma. The following facts will be used in this proof:

Fact 3.2. $\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$

Fact 3.3. $\binom{m}{k} = 0$, *if* $k < 0$ *or* $k > m$.

We will prove Sauer's Lemma by induction on $m + d$. Let $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.

Base Cases

- For $m = 0$ and all d. $\Pi_{\mathcal{C}}(m) = 1 = \sum_{i=0}^{d} \binom{0}{i} = \Phi_d(m)$. This is a degenerate case, where we label the empty set.
- For $d = 0$ and all m. $\Pi_{\mathcal{C}}(m) = 1 = \binom{m}{0} = \Phi_d(m)$. Not even shattering a point, so only one labeling is possible.

Inductive steps We assume that the lemma holds for any $m' + d' < m + d$. We need to show that for any S, $|\mathcal{C}[S]| \leq \Phi_d(m)$. To prove this, we construct two new hypothesis classes that are defined on one fewer instance and apply our induction hypothesis. Take any $S = \{x_1, \ldots, x_m\}$ and let $\mathcal{X}' := S' = \{x_1, \ldots, x_{m-1}\}$ be the domain of two new hypothesis classes $\mathcal{C}_1$ and $\mathcal{C}_2$.

Consider the predictions of $h \in \mathcal{C}$ on S, by consider $\mathcal{C}[S]$. The labeling in $\mathcal{C}[S]$ are all unique and come in one of the following forms:

- Pairs: where there are h and h' such that, for all $i \in [m-1]$, $h(x_i) = h'(x_i)$ and $h(x_m) \neq h'(x_m)$. For these pairs, we construct a function $g : \mathcal{X}' \to \mathcal{Y}$, that is defined similarly as h and h', except that it is not defined on x_m. We add g to both $\mathcal{C}_1$ and $\mathcal{C}_2$.
- Singleton: For h where there is no h' that satisfies the pair condition. For these we construct a function $g : \mathcal{X}' \to \mathcal{Y}$, that is the same as h except not defined on x_m. We add g only to $\mathcal{C}_1$.

Note that, the number of unique labelings in $\mathcal{C}$ is preserved, so $|\mathcal{C}[S]| = |\mathcal{C}_1| + |\mathcal{C}_2|$. See the following figure for an example of this construction.

	$\mathcal{C}[S]$ x_1, x_2, x_3, x_4, x_5	$\mathcal{C}_1$ x_1, x_2, x_3, x_4	$\mathcal{C}_2$ x_1, x_2, x_3, x_4
h_1	0 1 1 1 0	0 1 1 1	
h_2	0 1 1 1 1		0 1 1 1
h_3	1 0 0 1 1	1 0 0 1	
h_4	1 0 0 1 0		1 0 0 1
h_5	1 1 1 0 0	1 1 1 0	

Moreover, notice that if a set is shattered by $\mathcal{C}_1$ then it is also shattered by $\mathcal{C}$ because each labeling in $\mathcal{C}[S]$ can be generated using the same labeling (while ignoring x_m) in $\mathcal{C}_1$. So,

$$\mathrm{VCDim}(\mathcal{C}_1) \leq \mathrm{VCDim}(\mathcal{C}) = d.$$

Furthermore, if some set T is shattered by $\mathcal{C}_2$, then $T \cup \{x_m\}$ is shattered by $\mathcal{C}$. This is because every labeling in $\mathcal{C}_2$ refers to two labelings in $\mathcal{C}$, where the labels on $x_1, \ldots, x_{m-1}$ are the same and x_m is labeled in two different ways. Hence, $\mathrm{VCDim}(\mathcal{C}) \geq \mathrm{VCDim}(\mathcal{C}_2) + 1$, which implies

$$\mathrm{VCDim}(\mathcal{C}_2) \leq d - 1.$$

Now, by induction we have that $|\mathcal{C}_1| = |\mathcal{C}_1[S']| \leq \Phi_d(m-1)$ and $|\mathcal{C}_2| = |\Pi_{\mathcal{C}_2}(m-1) \leq \Phi_{d-1}(m-1)$. We have

$$\begin{aligned}
|\mathcal{C}[S]| &= |\mathcal{C}_1| + |\mathcal{C}_2| \\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{m}{i} \\
&= \Phi_d(m).
\end{aligned}$$

$\square$