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

Lecture 8: PAC Guarantee for Infinite Sets and Growth Function

Lecturer: Jacob Abernethy *Scribes: Yike Liu*
Editors: Luke Brandl, Nghia Nguyen, and Shuang Qiu

8.1 Review: Coin Toss

Recall the coin toss experiment, we have Bernoulli random variables $X_1, \ldots, X_n$, where:

$$X_i = \begin{cases} 1 & \text{with probability } \epsilon \\ 0 & \text{with probability } 1-\epsilon \end{cases}$$

It's obvious that:

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i = 0\right) = (1-\epsilon)^n \leq e^{-n\epsilon}$$

where the inequality is given by $\log(1-\epsilon) \leq -\epsilon$.

Fact 8.1 *Also we have:*

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i < \frac{\epsilon}{2} n\right) \leq e^{-\epsilon n/8}$$

You will show this in homework. We can see these two upper bounds are of the same order.

8.2 Review: General PAC Guarantee

Remember from last lecture, we talked about simplest general PAC guarantee. For finite set of concepts or hypotheses $\mathcal{C}$, you have an algorithm $\mathcal{A}$ that selects $h_S \in \mathcal{C}$. Giving target concept c and sample $S \sim D^m$, denote $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, $\forall h$, define:

$$R(h) = \mathbb{E}_{\mathbf{x}\sim D}[\mathbb{1}[h(\mathbf{x}) \neq c(\mathbf{x})]]$$

$$\hat{R}_S(h) = \frac{1}{m} \sum_{\mathbf{x}\in S} \mathbb{1}[h(\mathbf{x}) \neq c(\mathbf{x})]$$

Note that h_S was chosen because $\hat{R}_S(h_S) = 0$, we can bound:

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{S\sim D^m}(R(h_S) > \epsilon) &\leq \mathbf{Pr}(\exists h \in \mathcal{C} : \hat{R}_S(h) = 0 \text{ and } R(h) > \epsilon) && (8.1)\\
&\leq \sum_{h\in\mathcal{C}} \mathbf{Pr}(\hat{R}_S(h) = 0 \text{ and } R(h) > \epsilon) && (8.2)\\
&\leq \sum_{h\in\mathcal{C}} (1-\epsilon)^m && (8.3)\\
&\leq \sum_{h\in\mathcal{C}} e^{-m\epsilon} && (8.4)\\
&= |\mathcal{C}| e^{-m\epsilon} \qquad \text{(and we want this } < \delta) && (8.5)
\end{aligned}$$

This implies $\mathop{\mathbf{Pr}}_{S\sim D^m}(R(h_S) > \epsilon) < \delta$ when $m > \frac{1}{\epsilon}(\log|\mathcal{C}| + \log\frac{1}{\delta})$.

8.3 Hard Case of General PAC Guarantee

Now we consider $|\mathcal{C}| = \infty$. This comes in many examples such as the learning rectangles. The first step is that we only need to know how big $\mathcal{C}$ is when **restricted** to subsets. Denote $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, $\mathcal{C}|_S = \{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{C}\}$.

Definition 8.2 (growth function) *The **growth function** for $\mathcal{C}$ is:*

$$\Pi_{\mathcal{C}}(m) = \max_{S \subseteq \mathbb{X}, |S|=m} |\mathcal{C}|_S|$$

We want $\Pi_{\mathcal{C}}(m) = 2^{o(m)}$, but how should we handle $|\mathcal{C}| = \infty$?

Trick 8.3 *Take two samples S, S'.*

- *Step 1: Sample $S \sim D^m$*
- *Step 2: Find h_S such that $\hat{R}_S(h_S) = 0$*
- *Step 3: For analysis, sample $S' \sim D^m$*

Consider $\mathop{\mathbf{Pr}}\limits_{S \sim D^m, S' \sim D^m}(R(h_S) > \epsilon)$. We will look at the performance of h_S on the independent sample S', and consider 2 cases:

- (A) $\hat{R}_{S'}(h_S) > \frac{\epsilon}{2}$
- (B) $\hat{R}_{S'}(h_S) \leq \frac{\epsilon}{2}$

so we have:

$$\mathop{\mathbf{Pr}}_{S \sim D^m, S' \sim D^m}(R(h_S) > \epsilon) \tag{8.6}$$

$$\leq \underbrace{\mathop{\mathbf{Pr}}_{S \sim D^m, S' \sim D^m}\left(R(h_S) > \epsilon \wedge \hat{R}_{S'}(h_S) \leq \frac{\epsilon}{2}\right)}_{P_1} + \underbrace{\mathop{\mathbf{Pr}}_{S \sim D^m, S' \sim D^m}\left(R(h_S) > \epsilon \wedge \hat{R}_{S'}(h_S) > \frac{\epsilon}{2}\right)}_{P_2} \tag{8.7}$$

It may seem surprising that we separate the probability calculation in this way, but this was done for a very specific reason: we can apply different tricks to get bounds on P_1 and P_2. To start, note that to bound P_1 we have the useful fact that S and S' are uncorrelated. This means that we may as well assume S is fixed when we calculate the probability that $R(h_S) > \epsilon \wedge \hat{R}_{S'}(h_S) \leq \frac{\epsilon}{2}$. Then we can use the trick in Section 8.1. Mathematically, this means:

$$P_1 = \mathbf{Pr}(R(h_S) > \epsilon \wedge \hat{R}_{S'}(h_S) \leq \mathop{\mathbb{E}}_S[e^{-m\epsilon/8}] = e^{-m\epsilon/8}$$

where in the last inequality we used Fact 8.1.

How to bound P_2? The key is noting that I can sample S and S' in the following way:

1. First sample a set $U \sim D^{2m}$;
2. Then randomly partition U into two disjoint sets S, S' of size m, i.e. $U = S \cup S'$ (notationally, let's write this as $S \sqcup S' \sim U$).

Why did we do this? First we show that all that matters to bound this probability is to consider the hypothesis set $\mathcal{C}$ when restricted to U. But this is a finite set, so now we can use the union bound! Precisely:

$$\begin{aligned}
P_2 = \mathbf{Pr}\left(\hat{R}_{S'}(h_S) > \frac{\epsilon}{2} \wedge \hat{R}_S(h_S) = 0\right) &= \mathop{\mathbb{E}}_{U\sim D^{2m}}\left[\mathop{\mathbf{Pr}}_{S\sqcup S'\sim U}\left(\hat{R}_{S'}(h_S) > \frac{\epsilon}{2} \wedge \hat{R}_S(h_S) = 0\right)\right] \\
&\le \mathop{\mathbb{E}}_{U}\left[\mathop{\mathbf{Pr}}_{S\sqcup S'\sim U}\left(\exists h \in \mathcal{C} : \hat{R}_{S'}(h) > \frac{\epsilon}{2} \wedge \hat{R}_S(h) = 0\right)\right] \\
\text{(need only consider } \mathcal{C}|_U \text{ not all } \mathcal{C}) \qquad &= \mathop{\mathbb{E}}_{U}\left[\mathop{\mathbf{Pr}}_{S\sqcup S'\sim U}\left(\exists h \in \mathcal{C}|_U : \hat{R}_{S'}(h) > \frac{\epsilon}{2} \wedge \hat{R}_S(h) = 0\right)\right] \\
&\le \mathop{\mathbb{E}}_{U}\left[\sum_{h\in\mathcal{C}|_U}\mathop{\mathbf{Pr}}_{S\sqcup S'\sim U}\left(\hat{R}_{S'}(h) > \frac{\epsilon}{2} \wedge \hat{R}_S(h) = 0\right)\right]
\end{aligned}$$

We're almost there. Now we can use a simple balls-and-bins type analysis to get a bound on this probability value. Let us imagine we are dividing a larget set of blue balls and a small set of red balls into two equally-sized categories. Assume that there are at least $\frac{\epsilon m}{2}$ red balls and no more than $2m - \frac{\epsilon m}{2}$ blue balls, and these two sets are randomly parititioned into two bins of size m. What is the probability that the first bin got NO red balls? Each time you took a ball to place in the first bin, you had *at least* $\epsilon/4$ chance of getting a red ball. So after m rounds, you had *no more than* a chance of $(1 - \frac{\epsilon}{4})^m \le e^{-m\epsilon/4}$ of never seeing a red ball placed in the first bin. This calculation gives us:

$$P_2 \le \mathop{\mathbb{E}}_{U}\left[\sum_{h\in\mathcal{C}|_U}\mathop{\mathbf{Pr}}_{S\sqcup S'\sim U}\left(\hat{R}_{S'}(h) > \frac{\epsilon}{2} \wedge \hat{R}_S(h) = 0\right)\right] \le \mathop{\mathbb{E}}_{U}[|\mathcal{C}|_U|\, e^{-m\epsilon/4}] \le \Pi_{\mathcal{C}}(2m)e^{-m\epsilon/4}$$

Putting it all together, and assuming that m is large enough, it's easy to see that:

$$\mathop{\mathbf{Pr}}_{S\sim D^m, S'\sim D^m}(R(h_S) > \epsilon) \le \Pi_{\mathcal{C}}(2m)e^{-m\epsilon/8},$$

and we note that the constants can be significantly improved with more care.

8.4 Controlling $\Pi_{\mathcal{C}}(2m)$

To control the growth of $\Pi_{\mathcal{C}}(2m)$, we use the following trick.

Trick 8.4 (Vapnik-Chervonenkis dimension)

Definition 8.5 (shatter) $\mathcal{C}$ ***shatters*** $S \subseteq \mathbb{X}$ *if* $|\mathcal{C}|_S| = 2^{|S|}$

Some examples of shattering and impossible to be shattered are in Figure 8.1.

Definition 8.6 (VC-dimension) *The* ***VC-dimension*** *of* $\mathcal{C}$ *is* $\max\{d : \exists S \subseteq \mathbb{X}, |S| = d \text{ and } \mathcal{C} \text{ shatters } S\}$

Lemma 8.7 (Sauer-Shelah lemma) *If* $\mathcal{C}$ *has VC-dimension* d, $m > d$, *then:*

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

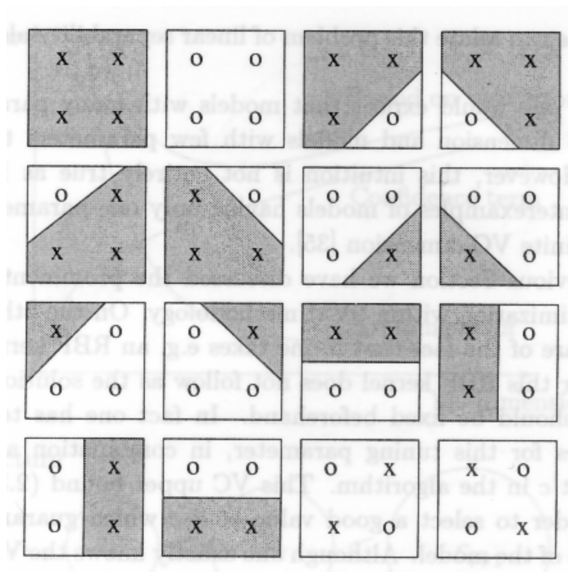

Figure 8.1: 4 points shattered and not shattered