10-806 Foundations of Machine Learning and Data Science

Maria-Florina Balcan Lecture 6: September 28th, 2015

---

1 Concentration Inequalities (Tail Inequalities)

Consider a coin of bias p flipped m times. Let S be the number of observed number of heads. So $\mathbf{E}[S/m] = p$.

Hoeffding bounds state that for any $\epsilon \in [0, 1]$,

1. $\Pr[\frac{S}{m} > p + \epsilon] \leq e^{-2m\epsilon^2}$, and
2. $\Pr[\frac{S}{m} < p - \epsilon] \leq e^{-2m\epsilon^2}$.

Chernoff bounds state that under the same conditions,

1. $\Pr[\frac{S}{m} > p(1 + \epsilon)] \leq e^{-mp\epsilon^2/3}$, and
2. $\Pr[\frac{S}{m} < p(1 - \epsilon)] \leq e^{-mp\epsilon^2/2}$.

Hoeffding bounds and Chernoff bounds are great tools that we will often use in our analyses.

2 Sample Complexity Lower Bounds

Recall that we earlier proved the following theorem:

Theorem 1 *Let C be an arbitrary hypothesis space of VC-dimension d. Let D be an arbitrary unknown probability distribution over the instance space and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size m satisfying*

$$m \geq \frac{8}{\epsilon}\left[d \ln\left(\frac{16}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right].$$

then with probability at least $1 - \delta$, all the hypotheses in C with $err_D(h) > \epsilon$ are inconsistent with the data, i.e., $err_S(h) \neq 0$.

So it is possible to PAC-learn a class C of VC-dimension d with parameters δ and ϵ given that the number of samples m is at least $m \geq c\left(\frac{d}{\epsilon}\log\frac{1}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ where c is a fixed constant. So, as long as $VCdim(C)$ is finite, it is possible to PAC-learn concepts from C even though $|C|$ might be infinite. We now show that this sample complexity result is tight within a factor of $O(\log(1/\epsilon))$.

Theorem 2 *Any algorithm for PAC-learning a concept class of VC dimension d with parameters ϵ and δ must use $\Omega(\frac{1}{\epsilon}[d + \log(1/\delta)])$ examples in the worst case.*

We will prove here the $\Omega(\frac{d}{\epsilon})$ part of the lower bound. The $\Omega(\frac{\log 1/\delta}{\epsilon})$ part will be in your homework.

Theorem 3 *Any algorithm for PAC-learning a concept class of VC dimension d with parameters ϵ and $\delta \leq 1/15$ must use more than $(d-1)/(64\epsilon)$ examples in the worst case.*

Proof: Consider a concept class C with VC dimension d. Let $X = \{x_1, \ldots, x_d\}$ be shattered by C. To show a lower bound we construct a particular distribution that forces any PAC algorithm to take that many examples. The support of this probability distribution is X, so we can assume WLOG that $C = C(X)$, so C is a finite class, $|C| = 2^d$. Note that we have arranged things such that for all possible labelings of the points in X, there is exactly one concept in C that induces that labeling. Thus, choosing the target concept uniformly at random from C is equivalent to flipping a fair coin d times to determine the labeling induced by c on X.

Let $m = (d-1)/(64\epsilon)$, and A be an algorithm that uses at most m i.i.d. examples and then produces a hypothesis h. We need to show that there exist a distribution D on X and a concept $c \in C$ such that the $err(h) > \epsilon$ with probability at least 1/15.

We first define D independently of A:

$$p(x_1) = 1 - 16\epsilon$$
$$p(x_2) = p(x_3) = \cdots = p(x_d) = \frac{16\epsilon}{d-1}$$

In the following we assume that S is a random i.i.d sample from D of size m. We want to establish that there is a c so that $\Pr_S[err(h) > \epsilon] > \frac{1}{15}$.

Let $X' = \{x_2, \ldots, x_d\}$. For any fixed $c \in C$ and hypothesis h, let

$$err'(h) = \Pr[c(x) \neq h(x) \wedge x \in X'].$$

For technical reasons, it is easier to prove that $\Pr_S[err'(h) > \epsilon] > 1/15$, which is enough since $err'(h) \leq err(h)$.

We pick a random $c \in C$ and show that with positive probability c is hard to learn for A, thereby showing that there must be some fixed c that is hard to learn for A.

Let us now define the event:

$B:$ S contains less than $(d-1)/2$ points in X'.

We have:

$$\Pr_S[B] \quad \geq \quad 1/2 \tag{1}$$

To see this, let Z be the number of points in S that are from X'. Clearly, $\mathrm{E}[Z] = 16\epsilon m = (d-1)/4$. We have $\Pr_S[B] \geq 1 - \Pr[Z \geq (d-1)/2] \geq 1/2$, since by Markov's inequality we have $\Pr[Z \geq (d-1)/2] \leq 1/2$.

We can also show:

$$\mathrm{E}_{c,S}[\, err'(h) \mid B\,] \quad > \quad 4\epsilon \tag{2}$$

Let S be the set of points that A gets. Choosing a random c is equivalent to flipping a fair coin for each point in X to determine its label. Since h is independent of the labeling of $X' - S$, the contribution to $err'(h)$ is expected to be $16\epsilon/(2(d-1))$ for each point in $X' - S$. When B occurs, we have $|X' - S| > (d-1)/2$; thus the expected value of $err'(h)$ given B is strictly greater than 4ϵ.

Using (1) and (2) we get a lower bound on $\mathrm{E}_{c,S}[err'(h)]$.

$$\mathrm{E}_{c,S}[err'(h)] \geq \Pr_S[B] \cdot \mathrm{E}_{c,S}[\, err'(h) \mid B\,] > \frac{1}{2} \cdot 4\epsilon = 2\epsilon.$$

So there must exist some $c^* \in C$ such that $\mathrm{E}_S[err'(h)] > 2\epsilon$. We take c^* as the target concept and show that A is likely to produce a hypothesis with high error rate.

Using the fact that for any h we have $err'(h) \leq \Pr[x \in X'] = 16\epsilon$ we note that

$$\mathrm{E}_S[\, err'(h) \mid err'(h) > \epsilon\,] \quad \leq \quad 16\epsilon \ \text{ for any fixed } c. \tag{3}$$

We have:

$$\begin{aligned} 2\epsilon &< \mathrm{E}_S[err'(h)] \\ &= \Pr_S[err'(h) > \epsilon] \cdot \mathrm{E}_S[\, err'(h) \mid err'(h) > \epsilon\,] \\ &\quad + (1 - \Pr_S[err'(h) > \epsilon]) \cdot \mathrm{E}_S[\, err'(h) \mid err'(h) \leq \epsilon\,]. \end{aligned}$$

Next we apply (3) to get

$$\begin{aligned} 2\epsilon < \mathrm{E}_S[err'(h)] &\leq \Pr_S[err'(h) > \epsilon] \cdot 16\epsilon + (1 - \Pr_S[err'(h) > \epsilon]) \cdot \epsilon \\ &= 15\epsilon \Pr_S[err'(h) > \epsilon] + \epsilon, \end{aligned}$$

which implies $\Pr_S[err'(h) > \epsilon] > 1/15$, as desired. ■

3 Recent results

As mentioned in class, there have been several fairly recent results on the general sample complexity of learning. First, Auer and Ortner [1] show that Theorem 1 is tight for arbitrary consistent learners. That is, there exist classes C and distributions D such that $\Omega(\frac{1}{\epsilon}[d \ln(1/\epsilon) + \ln(1/\delta)])$ examples are needed to ensure that every hypothesis $h \in C$ with $err_S(h) = 0$ has $err_D(h) \leq \epsilon$, where $d = VCdim(C)$.

However, Simon [2] shows that for any integer $k \geq 1$ there exist algorithms that require only $O(\frac{1}{\epsilon}[d \log^{(k)}(1/\epsilon) + \ln(1/\delta)])$ examples to learn to error ϵ with probability $1 - \delta$. Here, we define $\log^{(k)}(x) = \log(\log(\ldots \log(x)))$ where the log is iterated k times. The constant hidden by the "O" depends on k however.

References

[1] Peter Auer and Ronald Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2-3):151–163, 2007.

[2] Hans Ulrich Simon. An almost optimal PAC algorithm. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 1552–1563, 2015.