| | |
|---|---|
| **CSE 522: Algorithms and Uncertainty** | Spring 2017 |

Lecture 2 — March 29, 2017

*Lecturer: Anna R. Karlin*

**These rough notes follow lectures and notes by Avrim Blum, in some parts, verbatim. However, any errors are mine.**

# 1 Bounding sample complexity using the growth function

In this lecture we develop techniques for bounding the sample complexity that will work even when the hypothesis class is large or infinite.

**Definition 1.1.** Let $S$ be a set of examples or instances, e.g., each element in $S$ is a feature vector. Define $\mathcal{H}[S]$ to be the maximum number of ways to label, that is, classify as either 0 or 1, the points in $S$ using hypotheses in $\mathcal{H}$ . We define $\mathcal{H}[m] = \max_{S||S|=m} \mathcal{H}(S)$. This is called the **growth function** of $\mathcal{H}$.

For example, when $\mathcal{H}$ is the set of initial intervals, $\mathcal{H}[m]$ is $m + 1$.

**Theorem 1.2.** *For any class $\mathcal{H}$, and distribution $D$, if we draw a sample $S$ from $D$ of size*

$$m \geq \frac{2}{\epsilon} \log_2 \left( \frac{2\mathcal{H}[2m]}{\delta} \right),$$

*(and also $m \geq \frac{8}{\epsilon}$), then with prob $1 - \delta$, all $h$ with $\mathrm{err}_D(h) > \epsilon$ have $\mathrm{err}_S(h) > 0$. In other words, if the empirical risk minimizer has $\mathrm{err}_S(h) = 0$ then $\mathrm{err}_D(h) \leq \epsilon$ with probability at least $1 - \delta$.*

*Remark* 1.3. This means that in our bounds, we can replace the number of hypotheses $|\mathcal{H}|$ with $\mathcal{H}[2m]$, i.e., the number of hypotheses "after the fact", i.e., after $S$ is drawn. This is tricky because we can't just use a union bound after we have already drawn our set $S$.

*Proof.* Given set $S$ of $m$ examples, define the following events

$$A = \{\exists h \in \mathcal{H} \text{ with } \mathrm{err}_D(h) > \epsilon \text{ but } \mathrm{err}_S(h) = 0\}.$$

We want to show $\mathbb{P}(A)$ is low. Now, consider drawing *two* sets $S$ and $S'$ of $m$ examples each. Let $A$ be defined as before. Define

$$B = \{\exists h \in \mathcal{H} \text{ with } \mathrm{err}_{S'}(h) \geq \epsilon/2 \text{ but } \mathrm{err}_S(h) = 0\}.$$

**Claim:** $\mathbb{P}(A)/2 \leq \mathbb{P}(B)$. So, if we can bound $\mathbb{P}(B)$, then we can bound $\mathbb{P}(A)$.

**Proof of claim:** $\mathbb{P}(B) = \mathbb{P}(B|A)\,\mathbb{P}(A)$. We claim that $\mathbb{P}(B|A) > 1/2$. To see this, observe that conditioned on $A$, there is an $h$ with $\mathrm{err}_D(h)$ greater than $\epsilon$, but with no error on the sample $S$. Thus, conditioned on $A$, the event $B$ certainly happens as long as when $S'$ is sampled, $\mathrm{err}_{S'}(h) \geq \epsilon/2$. For this particular $h$, $\mathbb{P}\left[\mathrm{err}_{S'}(h) < \epsilon/2\right] \leq e^{-m\epsilon/8}$ (using Chernoff). Since $m \geq 8/\epsilon$, this is less than $1/2$. This means that $\mathbb{P}(B|A) > 1/2$ and thus $\mathbb{P}(A)/2 \leq \mathbb{P}(B)$.

Next, we show that $\mathbb{P}(B)$ is low. To do this, consider related event: draw

$$S = \{x_1, x_2, ..., x_m\} \quad \text{and } S' = \{x_1', x_2', ..., x_m'\}$$

and now create sets $T, T'$ using the following procedure Swap:

For each $i$, flip a fair coin:

- If heads, put $x_i$ in $T$ and put $x_i'$ in $T'$.
- If tails, put $x_i'$ in $T$ and put $x_i$ in $T'$.

**Claim:** $(T, T')$ has the same distribution as $(S, S')$.

Thus, we will consider

$$B_{T,T'} = \{\exists h \in \mathcal{H} \text{ with } \mathrm{err}_T(h) = 0 \text{ but } \mathrm{err}_{T'}(h) \geq \epsilon/2\}.$$

What's the point of this? Instead of $\mathbb{P}_{S,S'}[B]$ we will compute $\mathbb{P}_{S,S',\mathrm{swap}}\left[B_{T,T'}\right]$ Will show this is small by showing that for *all* $S, S'$, $\mathbb{P}_{\mathrm{swap}}\left[B_{T,T'}\right]$ is small.

The key here is that even if there are infinitely many hypotheses in $\mathcal{H}$, once we have drawn $S, S'$, the number of different labelings we have to worry about is at most $\mathcal{H}[2m]$, and will argue that whp (over the randomness in "swap") none of them will hurt us.

Now, fix $S, S'$ and fix some labeling $h$.

- If, for any $i$, $h$ makes a mistake on *both* $x_i$ and $x_i'$ then
$$\mathbb{P}_{\mathrm{swap}}\left[\mathrm{err}_T(h) = 0\right] = 0.$$
- If $h$ makes a mistake on less than $\epsilon * m/2$ points total, then
$$\mathbb{P}_{\mathrm{swap}}\left[\mathrm{err}_{T'}(h) \geq \epsilon/2\right] = 0.$$
- Else,
$$\mathbb{P}_{\mathrm{swap}}\left[\mathrm{err}_T(h) = 0 \wedge \mathrm{err}_{T'}(h) \geq \epsilon/2\right] \leq 2^{-\epsilon m/2},$$
since each of the mislabeled samples has to go to $T'$; each of these events happens with probability $1/2$. Now, we apply the union bound over $h \in \mathcal{H}[2m]$ and conclude that.
$$\mathbb{P}\left[B_{T,T'}\right] \leq \mathcal{H}[2m] * 2^{-\epsilon m/2}.$$
Setting $\mathbb{P}[A] \leq \delta/2$ and solving yields the results.

$\square$

*Remark* 1.4. We could rewrite this as follows: For any realizable[1] class $\mathcal{H}$ and distribution $\mathcal{D}$, with probability at least $1-\delta$

$$\mathrm{err}_D(h) \le \frac{2}{m}\log_2\left(\frac{2\mathcal{H}[2m]}{\delta}\right).$$

In the realizable case, we say that $\mathcal{H}$ is *PAC-learnable* if the right hand side above goes to 0 as $m$ goes to infinity. Whether or not the hypothesis class is PAC-learnable depends on whether $\log_2(\mathcal{H}[2m])/m$ goes to 0 or not.

There is also a uniform convergence version:

**Theorem 1.5.** *For any class $\mathcal{H}$, distrib $D$, if:*

$$m > \frac{8}{\epsilon^2}[\ln(2\mathcal{H}[2m]) + \ln(\delta^{-1})],$$

*then with prob $1-\delta$, all $h \in \mathcal{H}$ have $|\mathrm{err}_D(h) - \mathrm{err}_S(h)| < \epsilon$.*

*Proof.* We redo the proof using Hoeffding. Given set $S$ of $m$ examples, define the following events

$$A = \{\exists h \in \mathcal{H} \text{ with } |\mathrm{err}_D(h) - \mathrm{err}_S(h)| \ge \epsilon\}.$$

We want to show $\mathbb{P}(A)$ is low. Again, we draw two sets $S, S'$ of $m$ examples each. Define

$$B = \{\exists h \in \mathcal{H} \text{ with } |\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)| \ge \epsilon/2\}.$$

As before, we have $\mathbb{P}(B|A) \ge 1/2$ so $\mathbb{P}(A) \le 2 * \mathbb{P}(B)$. To see that $\mathbb{P}(B|A) \ge 1/2$ suppose that there is an $h$ with $\mathrm{err}_D(h) - \mathrm{err}_S(h) \ge \epsilon$. Conditioned on this, the probability that there is an $h$ with $\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) \ge \epsilon/2$ is at least the probability that $\mathrm{err}_D(h) - \mathrm{err}_{S'}(h) < \epsilon/2$. This has probability at least $1 - e^{-2m\left(\frac{\epsilon}{2}\right)^2}$ by the Hoeffding bound, which is at least 0.5 since $m > 2\epsilon^{-2}$. Applying a similar argument to the case where $\mathrm{err}_S(h) - \mathrm{err}_D(h) \ge \epsilon$ yields the desired fact.

Now, show $\mathbb{P}(B)$ is low:

As before, let's pick $S, S'$ where

$$S = \{x_1, x_2, ..., x_m\} \quad \text{and } S' = \{x'_1, x'_2, ..., x'_m\}$$

and do the random procedure swap to construct $T, T'$. Let's use $y_i$ to denote the element in $\{x_i, x'_i\}$ that goes to $T$, and let $y'_i$ denote the element that goes to $T'$. We'll show that for any $S, S'$,

$$\mathbb{P}_{\text{swap}}\left[\exists h \in \mathcal{H} \text{ with } |\mathrm{err}_T(h) - \mathrm{err}_{T'}(h)| > \epsilon/2)\right]$$

Again, there are at most $\mathcal{H}[2m]$ labelings of $S \cup S'$, so fix one such $h$.

Observe that

$$|\mathrm{err}_T(h) - \mathrm{err}_{T'}(h)| = \frac{1}{m}\left|\sum_i (\mathbb{1}_{h(y_i)\ne f(y_i)} - \mathbb{1}_{h(y'_i)\ne f(y'_i)})\right|.$$

[1]Recall that this means that the true classifier is in the class $\mathcal{H}$ and therefore, there is always some $h \in \mathcal{H}$ such that $\mathrm{err}_S(h) = 0$.

For any $i$ such that both $h(y_i)$ and $h(y_i')$ are right, or both are wrong, the contribution to the sum is 0. Thus, we can think of any $i$ such that exactly one of $h(x_i)$ and $h(x_i')$ is correct as a "coin". If the correct one of $x_i$ and $x_i'$ (for which $h(x) = f(x)$) goes to $T$ the contribution to $\sum_i(\mathbb{1}_{h(y_i)\neq f(y_i)} - \mathbb{1}_{h(y_i')\neq f(y_i')})$ is $+1$, whereas if it goes to $T'$, the contribution is $-1$. In other words, we are asking: if we flip $m' \leq m$ coins, where $m'$ is the number of indices corresponding to "coins", what is $\mathbb{P}(|\text{heads-tails}| > \epsilon \cdot m/2)$. This is the same as the number of heads being off from its expectation by more than $\epsilon \cdot m/4 = (1/4)(\epsilon \cdot m/m')m'$. By Hoeffding, this probability is most $2 \cdot e^{-(\epsilon \cdot m/m')^2 m'/8}$ and this is always $\leq 2 \cdot e^{-\epsilon^2 m/8}$. Now multiply by $\mathcal{H}[2m]$ and set to $\delta$.

□

As before, we can rewrite this as follows: For any class $\mathcal{H}$, distrib $D$, with probability at least $1 - \delta$

$$\mathrm{err}_D(h) \leq \mathrm{err}_S(h) + \sqrt{\frac{8}{m}\log_2\left(\frac{2\mathcal{H}[2m]}{\delta}\right)}.$$

Thus, as long as $\log_2(\mathcal{H}[2m])/m$ goes to 0, as the number of samples grows large, the sample or training error approaches the generalization error.

## 2 Notes

For detailed expositions of this material, including references, see Kearns and Vazirani [1] (chapter 3) Shalev-Schwartz and Ben-David [3] (chapters 2-6) and Mohri, Rostamizadeh and Talwalkar [2] (chapter 3).

## References

[1] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory.* MIT press, 1994.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning.* MIT press, 2012.

[3] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[4] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.