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 \( H[S] \) to be the maximum number of ways to label, that is, classify as either 0 or 1, the points in \( S \) using hypotheses in \( H \). We define \( H[m] = \max_{S| |S|=m} H(S) \). This is called the growth function of \( H \).

For example, when \( H \) is the set of initial intervals, \( H[m] \) is \( m + 1 \).

Theorem 1.2. For any class \( H \), and distribution \( D \), if we draw a sample \( S \) from \( D \) of size

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

(and also \( m \geq \frac{8}{\epsilon^2} \)), then with prob \( 1 - \delta \), all \( h \) with \( \text{err}_D(h) > \epsilon \) have \( \text{err}_S(h) > 0 \). In other words, if the empirical risk minimizer has \( \text{err}_S(h) = 0 \) then \( \text{err}_D(h) \leq \epsilon \) with probability at least \( 1 - \delta \).

Remark 1.3. This means that in our bounds, we can replace the number of hypotheses \( |H| \) with \( H[2m] \), 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

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

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

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

Claim: \( P(A) / 2 \leq P(B) \). So, if we can bound \( P(B) \), then we can bound \( P(A) \).

Proof of claim: \( P(B) = P(B|A) P(A) \). We claim that \( P(B|A) > 1/2 \). To see this, observe that conditioned on \( A \), there is an \( h \) with \( \text{err}_D(h) \) 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, $\text{err}_S'(h) \geq \epsilon/2$. For this particular $h$, $\Pr[\text{err}_S'(h) < \epsilon/2] \leq e^{-me/8}$ (using Chernoff). Since $m \geq 8/\epsilon$, this is less than 1/2. This means that $\Pr(B|A) > 1/2$ and thus $\Pr(A)/2 \leq \Pr(B)$.

Next, we show that $\Pr(B)$ is low. To do this, consider related event: draw $S = \{x_1, x_2, ..., x_m\}$ 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 $B_{T,T'} = \{\exists h \in \mathcal{H} \text{ with } \text{err}_T(h) = 0 \text{ but } \text{err}_{T'}(h) \geq \epsilon/2\}$.

What’s the point of this? Instead of $\Pr_{S,S'}[B]$ we will compute $\Pr_{S,S',\text{swap}}[B_{T,T'}]$ Will show this is small by showing that for *all* $S, S'$, $\Pr_{\text{swap}}[B_{T,T'}]$ 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 $\Pr_{\text{swap}}[\text{err}_T(h) = 0] = 0$.

- If $h$ makes a mistake on less than $\epsilon \ast m/2$ points total, then $\Pr_{\text{swap}}[\text{err}_{T'}(h) \geq \epsilon/2] = 0$.

- Else, $\Pr_{\text{swap}}[\text{err}_T(h) = 0 \land \text{err}_{T'}(h) \geq \epsilon/2] \leq 2^{-cm/2}$, 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 $\Pr[B_{T,T'}] \leq \mathcal{H}[2m] \ast 2^{-cm/2}$.

Setting $\Pr[A] \leq \delta/2$ and solving yields the results. 

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

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

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:

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

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

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

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

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

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

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

Now, show $P(B)$ is low:

As before, let’s pick $S, S'$ where

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

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 $S, S'$,

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

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

Observe that

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

1Recall 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 $\text{err}_S(h) = 0.$
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 $\sum_i(1_{h(y_i)\neq f(y_i)} - 1_{h(y'_i)\neq f(y'_i)})$ 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 $2 \cdot e^{-\epsilon^2 m'/8}$ 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 $1 - \delta$

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

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