CSE 522: Algorithms and Uncertainty Spring 2017

Lecture 3 — April 3, 2017

Lecturer: Anna R. Karlin

1 Recap

Last time, we proved the following theorem.

Theorem 1.1. *For any class $\mathcal{H}$, and distribution D, if we draw a sample S from D of size*

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

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 $1-\delta$.

Idea of the proof

We wanted to bound the probability that there is a hypothesis with no training error but large generalization error.

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

Instead, we considered a double sample S, S' and asked for the probability of event B:

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

namely that there is a hypothesis with large training error on S', but no training error on S. We argued that the probability of B is close to the probability of A because S' is a new random sample.

Then to bound the probability of B we fixed the two samples and a particular labelling of the elements of S and S'. We then considered a random swapping process, observed that the resulting pair of sets T, T' you get after swapping have same distribution as S, S', and argued that for any fixed S and S', once we did random swapping, event

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

was very unlikely to occur. Finally, we did a union bound over the labelings.

Where did we lose in this argument? First, whenever you do a union bound, you're overestimating the probability of the bad event. In particular, this can be a gross overestimate if many labelings are similar to each other, because then if the random swapping of one pair is unlikely to cause harm, then swapping another close one is unlikely to cause harm. Next lecture, we will find a way to get better bounds in some cases.

Remark 1.2. One of the corollaries ofwhat we did last time is the following: for any class $\mathcal{H}$, distrib D, with probability at least $1 - \delta$

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

Observe that even if $\mathcal{H}[2m]$ is constant, we don't expect the overfitting, i.e., $(\mathrm{err}_D(h) - \mathrm{err}_S(h))$, to be less than $c/\sqrt{m}$. To see this, suppose that there is one hypothesis in the class, and that the true function is a coin flip function, i.e. maps each example to a uniformly random value. Then $\mathrm{err}_D(h) = 1/2$, whereas $\mathrm{err}_S(h)$ is the fraction of errors on the sample of size m. When you toss m unbiased coins, the fraction of tails will be about $1/\sqrt{m}$ off from $1/2$, since X/m (with $X \sim \mathrm{Bin}(m, p)$) has mean $1/2$ and variance $1/4m$ and is well approximated by a normal distribution. So with constant probability, $\mathrm{err}_S(h)$ will be off from its mean by at least one standard deviation, which is $\Theta(1/\sqrt{m})$.

2 VC dimension

$\mathcal{H}[m]$ is sometimes hard to calculate exactly, but we'll see next that we can get a good bound using "VC-dimension". (VC-dimension is roughly the point at which $\mathcal{H}$ stops looking like it contains all labelings.)

Definition 2.1. A set of points S is **shattered** by $\mathcal{H}$ if there are hypotheses in $\mathcal{H}$ that label S in all of the $2^{|S|}$ possible ways.

In other words, all possible ways of classifying points in S are achievable using hypotheses in $\mathcal{H}$.

Definition 2.2. The **VC-dimension** of a hypothesis class $\mathcal{H}$ is the size of the largest set of points that can be shattered by $\mathcal{H}$.

So if the VC-dimension of a hypothesis class is d, that means there exists a set of d points that can be shattered, but there is no set of $d + 1$ points that can be shattered.

2.1 VC dimension (and $\mathcal{H}[m]$) examples

- finite class of functions $\mathcal{H}$. Then a set of points S can't be shattered if $|\mathcal{H}| \leq 2^{|S|}$. Therefore

$$\mathrm{VCdim} \leq \log_2(|\mathcal{H}|).$$

- Threshold functions over $\mathbb{R}^+$: that is each h is of the form $h(x) = \mathbb{1}_{0<x\leq a}$. Then VCdim is 1, since with 2 points, the labeling 0 followed by 1 is not possible. Also, $\mathcal{H}[m] = m + 1$.
- Intervals over $\mathbb{R}$: that is, each h of the form $h(x) = \mathbb{1}_{a\leq x\leq b}$ Then the VCdim is 2, since for any 3 points, the labeling 1, 0, 1 is not possible and $\mathcal{H}[m] = O(m^2)$.
- Axis aligned rectangles: VCdim $= 4$. If you take 5 points, and consider leftmost, rightmost, lowest and highest. Then can't make all of those $= 1$ and the remaining point, which is in the rectangle defined by the previous four $= 0$. Here $\mathcal{H}[m] = O(m^4)$.

- Convex polygons in the plane. If you place n points on the unit circle, then any subset of the points are vertices of a convex polygon that doesn't contain any points not in the subset. Therefore arbitrarily large sets can be shattered and the VC-dimension is infinite. This also implies that $\mathcal{H}[m] = 2^m$.
- half spaces in d dimensions. There is a set of $d+1$ points that can be shattered. E.g., take the vertices of a simplex in $\mathbb{R}^d$, say $\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_d$. Suppose that you want to label positively a subset A that includes the origin. Then letting $\mathbf{w}$ be the vector that has 1's corresponding to coordinates that are not in A and taking $\mathbf{w}^T\mathbf{x} \leq 0.5$ works labels A positively and the rest negatively. On the other hand, if A is a subset that excludes the origin, let $\mathbf{w}$ be the vector with 1's corresponding to coordinates in A. Then $\mathbf{w}^T\mathbf{x} \geq 0.5$ works.

  Claim 2.3. *No set of $d+2$ points in d dimensions can be shattered.*

  This can be proved using **Radon's Theorem** which says that any set of $d+2$ points can be partitioned into 2 distinct subsets A and B of points whose convex hulls intersect. This will prove the claim since any linear separator with A on one side must have its entire convex hull on that side, but then some point in B is also on that side.

  Proof of Radon's Thm: Let $\mathbf{x}_1, \ldots, \mathbf{x}_{d+2}$ be points in $\mathbb{R}^d$, and define $\mathbf{y}_i := (\mathbf{x}_i, 1) \in \mathbb{R}^{d+1}$. Since the $d+2$ vectors $\mathbf{y}_i$ are in $\mathbb{R}^{d+1}$, they are linearly dependent, so there are numbers $\alpha_1, \ldots, \alpha_{d+2}$ not all 0 such that $\sum_i \alpha_i \mathbf{y}_i = 0$. Let A be the set of points $\mathbf{y}_i$ such that $\alpha_i > 0$, and B be the rest. Then considering the final coordinate, we have

$$\sum_{i \in A} \alpha_i = \sum_{i \in B} |\alpha_i| \triangleq C,$$

  so

$$\sum_{i \in A} \frac{\alpha_i}{C} \mathbf{x}_i = \sum_{i \in B} \frac{|\alpha_i|}{C} \mathbf{x}_i.$$

3 Sample complexity bounds in terms of growth function and VC dimension

Next we will prove a bound on the growth function in terms of VC dimension that will allow us to show that the sample complexity essentially grows linearly in the VC dimension.

Theorem 3.1 (Sauer's Lemma). *Let $\mathcal{H}$ be a hypothesis class where* $\mathrm{VCdim}(\mathcal{H}) = \mathrm{d}$. *Then*

$$\mathcal{H}[m] \leq \sum_{i=0}^{d} \binom{m}{i} \triangleq \binom{m}{\leq d}.$$

To prove Sauer's lemma, we prove the following:

Lemma 3.2. *Fix any set S, and consider a set H of 0/1 labelings of S. Then H shatters at least $|H|$ subsets of S.*

Remark 3.3. For example, the labellings of x_1, x_2, x_3 consisting of 011, 110, 101, 100, and 111, shatters the empty set, $\{x_1\}, \{x_2\}\{x_3\}\{x_2, x_3\}$.

Proof. By induction on $|H|$. Base case: If $|H| = 1$, then H shatters the empty set.

Induction step: Fix an element $x \in S$, and let $H = H_0 \cup H_1$, where H_i is the subset of labelings in H that label element x with i (in $\{0, 1\}$). By hypothesis, H_i shatters at least $|H_i|$ sets, where i is either 0 or 1. Let T_i be the subsets shattered by H_i. Clearly no subset in T_i contains x. If a subset S' is shattered by both H_0 and H_1, then $S' \cup x$ is also shattered by H. Therefore the number of sets shattered by H is $|H_0| + |H_1|$. □

Proof of Sauer's Lemma: The contrapositive to the lemma is that if a set of labellings H shatters fewer than k subsets of a set S, then $|H| < k$. Now suppose that the largest subset of a set of size m that is shattered has size d. Then the number of sets shattered is at most

$$\binom{m}{\leq d}.$$

This means that that $\mathcal{H}[m]$ has cardinality at most

$$\binom{m}{\leq d}.$$

Corollary 3.4. *For any class $\mathcal{H}$, distrib D, if:*

$$m > \Omega(\frac{1}{\epsilon^2}[d \ln\left(\frac{2em}{d}\right) + \ln(1/\delta)])$$

then with prob $1 - \delta$, all $h \in \mathcal{H}$ have $|\mathrm{err}_D(h) - \mathrm{err}_S(h)| < \epsilon$. Since $m \geq \epsilon^{-1}$, we could also say that as long as

$$m = \Omega\left(\frac{d}{\epsilon^2} \ln\left(\frac{1}{\epsilon\delta}\right)\right)$$

the same conclusion holds.

Proof. Use Sauer's Lemma, so

$$\mathcal{H}[2m] \leq \binom{2m}{\leq d} \leq (2em/d)^d.$$

The rightmost inequality follows from simple algebra. □

Remark 3.5. Rewriting the corollary, we get that for any class $\mathcal{H}$, distrib D, with probability at least $1 - \delta$

$$\mathrm{err}_D(h) \leq \mathrm{err}_S(h) + O\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right).$$

Remark 3.6. These results are essentially tight in the sense that for any $\mathcal{H}$ and any learning algorithm, there is a distribution over instances and a target hypothesis in $\mathcal{H}$ for which

$$\mathbb{E}\left[\mathrm{err}_D(h_S)\right] \geq c\frac{\mathrm{VCdim}}{m}.$$

See homework.

4 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. 5

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning.* MIT press, 2012. 5

[3] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014. 5

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