CSE 522: Algorithms and Uncertainty Spring 2017

Lecture 4 — April 7, 2017

Lecturer: Anna R. Karlin

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

Our growth function/VC dimension bounds are nice but have two drawbacks that we'd like to address:

1. computability/estimability: If we don't understand $\mathcal{H}$ very well, might be hard to compute the VC-dimension exactly or otherwise get a good estimate of $\mathcal{H}[2m]$.
2. Tightness: There are two sources of loss: (a) We did a union bound over the labellings of the double-sample S, S', which is overly pessimistic if many of the splittings are very similar to each other. (b) We did worst-case over S, whereas we would rather do expected case over S, or even just have a bound that depends on our actual training set.

The next set of results will address both of these. What we will see is that we can capture how rich a family of hypotheses is by measuring the degree to which it can fit random noise.

1 Rademacher complexity

Definition 1.1. For a given set of data $S = x_1, ..., x_m$ and class $\mathcal{H}$ of functions from X to $\{-1, 1\}$, define the **empirical Rademacher complexity** of $\mathcal{H}$ as:

$$R_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\max_{h \in \mathcal{H}} \frac{1}{m} (\sum_i \sigma_i \cdot h(x_i)) \right]$$

where $\sigma = (\sigma_1, ..., \sigma_m)$ is a random {-1,1} labeling.

This measures the correlation of the best function in the class to a random labeling. The Rademacher complexity captures the complexity of the class on the sample itself and is therefore distribution dependent, in contrast to VC dimension.

- If $|\mathcal{H}(S)| = 1$, then the Rademacher complexity is 0.
- If $|\mathcal{H}(S)| = 2^{|S|}$, then the Rademacher complexity is 1.

Definition 1.2. We define the (distributional) **Rademacher complexity** of $\mathcal{H}$ as:

$$R_D(\mathcal{H}) = \mathbb{E}_S [R_S(\mathcal{H})].$$

We will prove the following theorem:

Theorem 1.3. *For any class $\mathcal{H}$ and distribution D over samples, if we see m examples then with probability at least $1-\delta$ every $h \in \mathcal{H}$ satisfies*

$$\begin{aligned} err_D(h) &\leq err_S(h) + R_D(\mathcal{H}) + \sqrt{(ln(2/\delta)/2m)} \\ &\leq err_S(h) + R_S(\mathcal{H}) + 3 \cdot \sqrt{(ln(2/\delta)/2m)}. \end{aligned}$$

Remark 1.4. We will see that this bound in the worst case is the same as the VC dimension bound, but in some cases, much much better.

Remark 1.5. Notice that the first line takes an expectation over S. The second line looks at the actual S. So this is a distribution dependent bound. However, the bound is still worst case over true hypothesis, so we don't expect overfitting less than $1/\sqrt{m}$, e.g. if the true function is the coin flip function.

For the proof, we will need the following useful tool called McDiarmid's inequality. This generalizes Hoeffding bounds to the case where, rather than considering the average of a set of independent RVs, we are considering some other function of them. Specifically,

Theorem 1.6 (McDiarmid's inequality)**.** *Say $\mathbf{X} := (X_1, ..., X_m)$ are independent RVs, and $\phi(X_1, ..., X_m)$ is some real-valued function. Assume that ϕ satisfies the Lipschitz condition. That is,*[1]

$$|\Phi(x_i', \mathbf{x}_{-i}) - \Phi(x_i, \mathbf{x}_{-i})| \leq c_i.$$

if X_i is changed, ϕ can change by at most c_i. Then:

$$Pr[\phi(\mathbf{X}) > E[\phi(\mathbf{X})] + \epsilon] \leq e^{-2\cdot\epsilon^2/\sum_i c_i^2}$$

Remark 1.7. For example, if all $c_i \leq 1/m$ (which would be the case if $\phi(\mathbf{X}) = \sum_i X_i/m$ and each $X_i \in \{0, 1\}$), we get:

$$Pr[\phi(\mathbf{X}) > E[\phi(\mathbf{X})] + \epsilon] \leq e^{-2\cdot\epsilon^2\cdot m}.$$

just like Hoeffding.

Proof. The first step of the proof is to simplify the quantity we care about. Specifically, let's define

$$\text{MaxGap}(S) = \max_{h\in\mathcal{H}}[\text{err}_D(h) - \text{err}_S(h)].$$

We want to show that with probability at least $1-\delta$, MaxGap(S) is at most some ϵ.

As a first step, we can use McDiarmid to say that with high probability, MaxGap(S) will be close to its expectation. In particular, the examples x_j are independent random variables and MaxGap(S) can change by at most $1/m$ if any individual x_j in S is replaced (because the gap for any specific

[1] We use the notation $f(x_i, \mathbf{x}_{-i})$, as a shorthand for $f(x_1, \ldots, x_i, \ldots, x_n)$.

h can change by at most $1/m$). So, using MaxGap as our "ϕ" in McDiarmid's Inequality, with probability at least $1 - \delta/2$, we get:

$$\text{MaxGap}(S) \leq \mathbb{E}_S[\text{MaxGap}(S)] + \sqrt{ln(2/\delta)/(2m)}.$$

So, to prove the first line of the theorem, we just need to show that

$$\mathbb{E}_S[\text{MaxGap}(S)] \leq R_D(\mathcal{H}). \tag{1}$$

Note that the second line of the theorem follows immediately from the first line plus an application of McDiarmid to the random variable $R_S(\mathcal{H})$ (since a single example changes $R_S(\mathcal{H})$ by at most $2/m$). So, we just need to prove the first line.

The next step is to do a double-sample argument like we did before. Specifically, let's rewrite $\text{err}_D(h)$ as $\mathbb{E}_{S'}[\text{err}_{S'}(h)]$, where S' is a new set of m points drawn from D. So, we can rewrite $\mathbb{E}_S[\text{MaxGap}(S)]$ as:

$$\mathbb{E}_S\left[\max_{h\in\mathcal{H}} \mathbb{E}_{S'}[\text{err}_{S'}(h) - \text{err}_S(h)]\right] \leq \mathbb{E}_{S,S'}\left[\max_{h\in\mathcal{H}}(\text{err}_{S'}(h) - \text{err}_S(h))\right]$$

The inequality follows from the fact that in the rightmost term we get to pick h after seeing both S and S'.

If we let $S = x_1, ..., x_m$ and let $S' = x'_1, ..., x'_m$ then we can rewrite this as:

$$\mathbb{E}_{S,S'}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_i [\text{err}_{x'_i}(h) - \text{err}_{x_i}(h)]\right]$$

where

$$\text{err}_x(h) = \mathbb{1}_{h(x)\neq f(x)}.$$

Now, as in the VC proof, let's imagine that for each index i, we flip a coin to decide whether to swap x_i and x'_i or not before taking the max. This doesn't affect the expectation since everything is i.i.d. So, letting $\sigma_i \in \{-1, 1\}$ at random, we can rewrite our quantity as:

$$\mathbb{E}_{S,S',\sigma}\left[\max_{h\in\mathcal{H}} \mathbb{E}\left[\frac{1}{m}\sum_i \sigma_i[\text{err}_{x'_i}(h) - \text{err}_{x_i}(h)]\right]\right]$$

Thus, we have

$$\mathbb{E}_S[\text{MaxGap}(S)] \leq \mathbb{E}_{S',\sigma}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_i \sigma_i \text{err}_{x'_i}(h)\right] + \mathbb{E}_{S,\sigma}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_i -\sigma_i \text{err}_{x_i}(h)\right]$$

since the gap is only larger if allow the two h's to differ

$$= 2\mathbb{E}_{S,\sigma}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_i \sigma_i \text{err}_{x_i}(h)\right]$$

by symmetry, since σ_i is random $\{-1, 1\}$.

Finally, we will show that this is equal to $R_D(\mathcal{H})$. To see this, we observe that, since x_i is random from distribution $\mathcal{D}$ and σ_i is random in $\{-1, 1\}$, the following random variables for each fixed x_i have the same distribution. (In what follows, only σ_i is random.)

$$\begin{aligned}\sigma_i h(x_i) &= \sigma_i f(x_i) h(x_i) = \sigma_i(1 - 2\text{err}_{x_i}(h)) \\ &= -\sigma_i(1 - 2\text{err}_{x_i}(h)) \\ &= 2\sigma_i \cdot \text{err}_{x_i}(h) - \sigma_i\end{aligned}$$

Therefore,

$$R_D(\mathcal{H}) = E_{S,\sigma}[\max_{h\in\mathcal{H}} \frac{1}{m}[\sum_i \sigma_i h(x_i)]] = \cdot E_{S,\sigma}[\max_{h\in\mathcal{H}} \frac{1}{m} \sum_i (2\sigma_i \cdot \text{err}_{x_i}(h) - \sigma_i)] = 2 \cdot E_{S,\sigma}[\max_{h\in\mathcal{H}} \frac{1}{m}[\sum_i \sigma_i \cdot \text{err}_{x_i}(h)]].$$

This completes the proof of (1).

□

2 Discussion

Relating Rademacher complexity and VC dimension

Relating Rademacher and VC: We first observe that the bounds based on Rademacher complexity are essentially as good as our VC bounds (and sometimes much better). In particular, let's consider how big $R_S(\mathcal{H})$ can be? Fix some h. Then since

$$\sigma_i h(x_i) = 2\mathbb{1}_{h(x_i)=\sigma_i} - 1,$$

$$\begin{aligned}\mathbb{P}\left[\frac{1}{m}\sum_i \sigma_i h(x_i) \geq 2\epsilon\right] &= \mathbb{P}\left[\frac{1}{m}\sum_i (2\mathbb{1}_{h(x_i)=\sigma_i} - 1) \geq 2\epsilon\right] \\ &= \mathbb{P}\left[\frac{1}{m}\sum_i \mathbb{1}_{h(x_i)=\sigma_i} \geq \frac{1}{2} + \epsilon\right] \\ &\leq e^{-2m\epsilon^2}\end{aligned}$$

by Hoeffding. Therefore

$$\mathbb{P}\left[\exists h \in \mathcal{H}[m] \text{ s.t. } \sum_i \sigma_i h(x_i) \geq 2\epsilon m\right] \leq \mathcal{H}[m] e^{-2m\epsilon^2}$$

setting this to be at most δ means that, with probability at least $1 - \delta$, if

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

then

$$R_S(\mathcal{H}) \leq 2\epsilon.$$

So, $R_S(\mathcal{H})$, which is the expected maximum correlation really can't be much higher than what we had before and probably is lower.

An example

[This example is from lecture notes by Clayton Scott.]

Suppose that $\{A_1, A_2, \ldots, A_k\}$ is a fixed partition over instances X. Let $\mathcal{H}$ be the set of hypotheses that are constant on each part A_i.Then $|\mathcal{H}| = 2^k$. We will abuse notation below and refer to $h(A_i)$ as the value h takes on each element in A_i.

The VC-dimension of $\mathcal{H}$ is clearly k. Let's see what we get from the Rademacher complexity.

We have

$$\begin{aligned} R_S(\mathcal{H}) &= \mathbb{E}_\sigma \left[\max_{h \in \mathcal{H}} \frac{1}{m} (\sum_i \sigma_i \cdot h(x_i)) \right] \\ &= \frac{1}{m} \mathbb{E}_\sigma \left[\sum_{j=1}^{k} \max_{h \in \mathcal{H}} \; h(A_j) \sum_{i \mid x_i \in A_j} \sigma_i \right]. \end{aligned}$$

Now observe that

$$\mathbb{E} \left[\max_{h \in \mathcal{H}} \; h(A_j) \sum_{i \mid x_i \in A_j} \sigma_i \right] = \mathbb{E} \left[\left| \sum_{i \mid x_i \in A_j} \sigma_i \right| \right],$$

since each $\sigma_i \in \{-1, +1\}$. Thus,

$$\mathbb{E}_\sigma \left[\max_{h \in \mathcal{H}} \; h(A_j) \sum_{i \mid x_i \in A_j} \sigma_i \right] = \mathbb{E} \left[\sqrt{\left(\sum_{i \mid x_i \in A_j} \sigma_i \right)^2} \right] \leq \sqrt{\mathbb{E} \left[\left(\sum_{i \mid x_i \in A_j} \sigma_i \right)^2 \right]}$$

by Jensen's Inequality applied to the square root function (which is concave). Finally, since

$$\mathbb{E} \left[\sigma_i \sigma_j \right] = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

$$\sqrt{\mathbb{E} \left[\left(\sum_{i \mid x_i \in A_j} \sigma_i \right)^2 \right]} = \sqrt{m_j}, \qquad \text{where} \qquad m_j = |i \mid x_i \in A_j| \,.$$

Therefore,

$$R_S(\mathcal{H}) = \frac{1}{m} \sum_{j=1}^{k} \sqrt{m_j}.$$

Thus, if $\mathcal{D}$ generates examples uniformly at random and, say, $m_j = m/k$, then $R_S(\mathcal{H}) = \sqrt{k/m}$. Unless $m > k$, this doesn't give any bound on the generalization error.

On the other hand, if the distribution $\mathcal{D}$ is such that almost all of the examples are from one part, say A_1, then

$$R_S(\mathcal{H}) \approx \frac{1}{\sqrt{m}}.$$

3 Notes

For detailed expositions of this material, including references, see Shalev-Schwartz and Ben-David [3] (chapter 26) 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.