Lecture 6: Rademacher Complexity

Instructor: Aditya Bhaskara *Scribe: Vinu Joseph*

CS 5966/6966: Theory of Machine Learning

January 30th, 2017

Abstract

In this lecture, we discuss Rademacher complexity, which is a different (and often better) way to obtain generalization bounds for learning hypothesis classes.

1 Introduction

Recall the notions introduced in the last couple of lectures – the growth function $\tau_{\mathcal{H}}(m)$ and the Vapnik-Chervonenkis (VC) dimension.

We saw that a class is learnable iff $\frac{\log \tau_{\mathcal{H}}(m)}{m}$ tends to 0 as $m \to \infty$. Further, we saw that this condition is equivalent to saying that $\mathcal{H}$ has a finite VC dimension (via Sauer's Lemma). If $\mathcal{H}$ is learnable, it means that a hypothesis that minimizes the empirical risk also generalizes well, w.h.p. In this lecture, we will see an alternative way to obtain generalization bounds.

2 Motivation – Limitations of PAC

While the PAC model provides an elegant characterization of learnability, it is somewhat restrictive. Recall the definition of PAC learnability: there exists an algorithm, that for *every* distribution $\mathcal{D}$ and $\epsilon, \delta > 0$, finds a hypothesis that is "ϵ-optimal" with probability $1 - \delta$.

By ϵ-optimality, we simply mean that the loss is within ϵ of the best hypothesis in $\mathcal{H}$.

There are two strong requirements here. First, we need the learning algorithm to succeed for *every* distribution $\mathcal{D}$ over the inputs. Second, for *every* accuracy parameter ϵ, we need the algorithm to output an ϵ-optimal hypothesis, using $m(\epsilon, \delta)$ samples.

We can imagine relaxing both of these conditions. For instance, in practice we might not need the algorithm to work *for all* $\mathcal{D}$. Alternately, we might require the algorithm to learn a hypothesis that is "somewhat" accurate (say $\epsilon = 0.1$). The latter question will be addressed in the course when we discuss *boosting*. Today, we will focus on the first question. Specifically, can we obtain generalization bounds that *depend* on $\mathcal{D}$ (and are thus less "worst case")?

This is the setting for the so called Rademacher complexity, which is the subject of today's lecture.

Rademacher complexity is named after Hans Rademacher, a german-born mathematician known for work in mathematical analysis and number theory, this is a measure of richness of a class of real-valued functions with respect to a probability distribution.

3 Distribution dependent generalization

Suppose we have some distribution $\mathcal{D}$ (unknown) from which we obtain m i.i.d. samples S. Let $S = x_1, x_2, \cdots, x_m$, for convenience. As usual, we denote the empirical loss of a hypothesis $h \in \mathcal{H}$ by $L_S(h)$, and its *true* loss by$L_{\mathcal{D}}(h)$.

To prove a generalization bound, we require the difference between the two quantities to be small, for every $h \in \mathcal{H}$. (In this case, the ERM procedure will yield a hypothesis with near optimal loss).

Let us start with a thought experiment: suppose we are given a hypothesis h, the sample S, and we are asked to estimate $L_S(h) - L_{\mathcal{D}}(h)$. One natural idea is to try "cross-validation" – i.e., we can split the sample S into two, S_1 (training) and S_2 (test), and use the estimate

$$L_S(h) - L_{\mathcal{D}}(h) \approx L_{S_1}(h) - L_{S_2}(h).$$

Cross-validation and Rademacher averages

Let us now expand out the expression introduced above. Let $\ell(h, x_i)$ be an indicator which is 1 if hypothesis h errs on the example x_i and is 0 otherwise. The equation above can now be re-written as:

$$L_S(h) - L_{\mathcal{D}}(h) \approx \frac{1}{|S_1|} \sum_{x_i \in S_1} \ell(h, x_i) - \frac{1}{|S_2|} \sum_{x_i \in S_2} \ell(h, x_i).$$

Suppose we had $|S_1| = |S_2| = |S|/2$. The RHS above is now quite simple (recall $|S| = m$):

$$\frac{2}{m} \sum_{x_i \in S_1} \ell(h, x_i) - \sum_{x_i \in S_2} \ell(h, x_i).$$

Thus if S is given to us, then the loss $\ell(h, x_i)$ appears with a $+1$ coefficient if $x_i \in S_1$ and a -1 coefficient if $x_i \in S_2$. Suppose we denote this sign by σ_i, we have

$$L_{S_1}(h) - L_{S_2}(h) = \frac{2}{m} \left[\sum_i \sigma_i \ell(h, x_i) \right].$$

Suppose we have a fixed S, and we partition it into two halves S_1, S_2 at random, then this is equivalent to picking signs σ_i at random (independently and uniformly), and the sum above "in principle" provides an estimate of the generalization error. (There is a mild technicality – if we choose the signs independently, then we may not exactly have $|S_1| = |S_2|$, but we ignore this.)

The sum above is typically referred to as a *Rademacher average*. We will now define the notion of Rademacher complexity, which intuitively is the supremum of the quantity above, over $h \in \mathcal{H}$.

4 Rademacher Complexity and Generalization

4.1 definition. Let $\mathcal{H}$ be a hypothesis class over a domain $\mathcal{X}$, and let S be a sample. The Rademacher Complexity of $\mathcal{H}$ with respect to the sample S is defined to be

$$(1) \qquad R_S(\mathcal{H}) := \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i \ell(h, x_i) \right|$$

In the above defintion, the expectation is over the signs σ, which are chosen to be ± 1 independently. We take supremum over h, because we wish to have

a generalization guarantee ($L_S(h) \approx L_{\mathcal{D}}(h)$) for all $h \in \mathcal{H}$.

Indeed, the notion of Rademacher averages and complexity can be defined not just for loss functions of hypotheses, as above, but for arbitrary classes of functions. There are many settings in machine learning and theoretical computer science, in which we have a class of functions $\mathcal{F}$ over a domain, and our goal is to approximate the expected value of $f \in \mathcal{F}$ using samples from the domain. The goal broadly is to obtain a good estimate for *every* $f \in \mathcal{F}$, using a small number of samples.

The notion of Rademacher complexity allows us to analyze this general setting, for bounded functions.

4.2 DEFINITION. Let $\mathcal{F}$ be a class of real valued functions over a domain $\mathcal{X}$, that are take values in $[-1, 1]$. The Rademacher complexity of $\mathcal{F}$ wrt a set $S \subseteq \mathcal{X}$ is defined to be:

The bound of $[-1,1]$ is arbitrary – any bounded interval will work as well.

$$R_S(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{|S|} \left| \sum_{i=1}^{m} \sigma_i f(x_i) \right| \tag{2}$$

Next, we introduce a slight variant, where we do not have S, but simply the sample size m.

$$R_m(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{D}^m} R_S(\mathcal{F}) \tag{3}$$

This is a complexity measure for $\mathcal{F}$ that depends on $\mathcal{D}$ (unlike the notions of growth function and VC dimension).

A New Generalization Bound

As discussed earlier, we now present a bound that applies broadly to a setting in which we have a function class $\mathcal{F}$, we have points coming form distribution $\mathcal{D}$, and we want to estimate the expected value $\mathbb{E}_{x \sim \mathcal{D}} f(x)$ for *all* $f \in \mathcal{F}$.

4.3 THEOREM. *Let $S = \{x_1, x_2, \cdots, x_m\}$ be a sample from an unknown distribution $\mathcal{D}$. With probability at least $1 - \delta$, we have:*

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathcal{D}}(f) - \frac{1}{m} \sum_i f(x_i) \right| \leq 2R_m(\mathcal{F}) + O\left(\sqrt{\frac{log(1/\delta)}{m}} \right). \tag{4}$$

Further, the inequality also holds w.h.p., when $R_m(\mathcal{F})$ is replaced with $R_S(\mathcal{F})$. I.e.,

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathcal{D}}(f) - \frac{1}{m} \sum_i f(x_i) \right| \leq 2R_S(\mathcal{F}) + O\left(\sqrt{\frac{log(1/\delta)}{m}} \right). \tag{5}$$

The first term is Radamacher complexity, and the second term is the tail that one expects in a standard Chernoff bound for a single f. Thus having a small Rademacher complexity $R_m(\mathcal{F})$ implies a good generalization bound.

The second inequality above, i.e. (5) is interesting because it gives a bound that can, in principle, be evaluated *given the sample S* (with a knowledge of $\mathcal{F}$). Thus, given a sample, we can, with high probability, **estimate** the generalization error! Indeed, this easily implies the generalization bound we saw earlier in terms of the growth function. We will outline the proof in Section 5

The proof of the theorem is via a clever *symmetrization* argument. We will skip

it here, and point to Rob Schapire's notes:
`https://www.cs.princeton.edu/courses/archive/spring13/cos511/scribe_notes/0305.pdf`

Now we discuss one key ingredient in the proof, which is a concentration inequality that is widely useful.

McDiarmid Inequality

Let us start with a word on concentration inequalities. Given a random variable X that takes real values, concentration inequalities deal with the question: *what is the probability that X deviates a* lot *from its expectation?*

There has been a long line of work in probability and statistics that essentially says that if X is an *aggregate* of a large number of independent random variables without depending *too strongly* on each of the variables, then the so-called "law of large numbers" kicks in, and X is tightly concentrated around the mean. The McDiarmid inequality is a concrete way of formalizing this intuitive statement.

Suppose we have a random variable X which is a function g of a collection of independent (not necessarily identically distributed) variables $X_1, \ldots, X_m$. I.e.,

$$X = g(X_1, X_2, \ldots, X_m).$$

Now, suppose the *sensitivity* of g to the variable X_i is bounded by c_i, i.e.,

$$\forall x_1, x_2, \ldots, x_m,\ \forall x_i', \\ \left|g(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m) - g(x_1, x_2, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)\right| \leq c_i.$$

Then, the McDiarmid inequality says:

(6)

$$\Pr_{X_1, X_2, \ldots, X_m}\left[|g(X_1, X_2, \cdots, X_m) - \mathbb{E}[g(X_1, X_2, \cdots, X_m)]| > t\right] \leq \exp\left(\frac{-t^2}{c_1^2 + c_2^2 + c_3^2 + \cdots + c_m^2}\right).$$

The expectation is (once again) over the choice of $X_1, \ldots, X_m$. Note that we do not care how complicated g is – all we need for (6) to hold is that g does not depend "too much" on its individual arguments. In the HWs, we will see some interesting examples.

For now, let us see how McDiarmid's inequality is used in our setting. Suppose we define

$$g(X_1, X_2, \ldots, X_m) = \sup_{f \in \mathcal{F}}\left[\frac{1}{m}\sum_i f(X_i) - \mathop{\mathbb{E}}_{X \sim \mathcal{D}}[f(X)]\right].$$

I.e., g is the supremum over $f \in \mathcal{F}$ of the generalization error. This is well defined even when $\mathcal{F}$ is infinite. Let us now see how sensitive $g()$ is to a change in x_i. Suppose we consider some values for $x_1, \ldots, x_m$, and replace x_i with x_i'.

Now, for every f, we have $\frac{f(x_i) - f(x_i')}{m} \leq \frac{2}{m}$, and thus each term in the supremum changes by at most $2/m$. This means that the supremum itself changes by at most $2/m$. (You may want to show this formally if it is not clear.) Thus, we can set $c_i = 2/m$ for all i, and use McDiarmid's inequality. This shows

that with high probability, $g()$ will be close to *its* expectation. Thus the second (and the tricky) of the proof is to bound

$$\mathbb{E}_{X_1,\ldots,X_m}\left[g(X_1,\ldots,X_m)\right].$$

5 Comparison to Earlier Bounds

As we mentioned earlier, the Rademacher bound (Theorem 4.3) is *stronger* than the other ways of obtaining generalization bounds (including the one we saw involving the growth function, and the VC dimension).

Indeed, using Eq. (5), it follows that we only need to bound $R_S(\mathcal{F})$, in order to bound the generalization. The following lemma then implies that we can recover the theorem we saw a couple of lectures ago (bound in terms of the growth function).

5.1 lemma. *Let $\mathcal{F}$ be the class containing loss functions of a hypothesis class $\mathcal{H}$. Then*

$$R_S(\mathcal{F}) \leq \sqrt{\frac{log(\tau_{\mathcal{H}}(m))}{m}},$$

where $\tau_{\mathcal{H}}$, as before, refers to the growth function.

To prove the lemma, start by recalling the definition of $R_S(\mathcal{F})$,

$$R_S(\mathcal{F}) := \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i f(x_i)\right|.$$

In our case, $\mathcal{F}$ is the class of loss functions for $\mathcal{H}$. I.e., each f corresponds to one $h \in \mathcal{H}$, and $f(x_i)$ is 1 if h classifies x_i correctly, and 0 otherwise. Thus, each f can be viewed as a $0/1$ vector of length m. The number of distinct strings is at most the *growth function* $\tau_{\mathcal{H}}(m)$, because the growth function captures all the ways in which $h \in \mathcal{H}$ can classify the points S. Thus, the supremum over $f \in \mathcal{F}$ is essentially the supremum over $|\tau_{\mathcal{H}}(m)|$ terms!

Once we have the supremum over a finite number of f's we can simply use a combination of Chernoff and union bounds, as we have seen earlier. This is captured in the so-called Massart lemma. We will skip the full proof of the Lemma – it can be found in Chapter 26 of the textbook.

6 Conclusion

This concludes the portion of the course on statistical learning theory. In general, to prove learnability for a class $\mathcal{H}$, we prove a generalization bound (which implies that ERM is a *good* hypothesis), and then we need to *compute* the ERM. We have seen different techniques (of which the Rademacher complexity is the most powerful), to do the first of these steps. The second step is an optimization problem. This will be focus of the second part of the course. We then study the online learning framework, which is closely related to optimization.