COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Zoe Ashwood

Lecture #10
March 6, 2019

1 Summary and Goal for Class

Previously we provided bounds on generalization error in terms of three different complexity measures – the cardinality of the hypothesis space, $|\mathcal{H}|$, the growth function $\Pi_{\mathcal{H}}(m)$ and the VC-dimension, VC-dim($\mathcal{H}$). In this lecture, we will introduce a fourth measure of complexity, the Rademacher complexity. We would like to provide a uniform convergence theorem of the form: with probability $1-\delta$, $\forall h \in \mathcal{H}$, $|\text{err}(h) - \hat{\text{err}}(h)| \leq \epsilon$, but we will first prove a more general result. After obtaining a bound on the generalization error in terms of the Rademacher complexity, we will show that the Rademacher complexity measure subsumes all three of our previous complexity measures and that we can get back many of the bounds that we saw earlier in class by evaluating Rademacher complexity for different hypothesis spaces. We will conclude part 1 of the course (writing bounds for the generalization error in terms of training error and complexity measures) and we will start part 2, which introduces learning algorithms such as boosting.

2 Generalization Bounds with Rademacher Complexity

In the last lecture, we were in the middle of proving the following theorem:

2.1 Theorem

Let $\mathcal{F}$ be a family of functions $f : \mathcal{Z} \to [0,1]$, $\mathcal{S} = \langle z_1, ..., z_m \rangle$ where $z_i \sim \mathcal{D}$. Then with probability $\geq 1-\delta$, $\forall f \in \mathcal{F}$:

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_{\mathcal{S}}[f] + 2\mathbb{R}_m(\mathcal{F}) + O\left(\sqrt{\frac{ln(\frac{1}{\delta})}{m}}\right) \tag{1}$$

and

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_{\mathcal{S}}[f] + 2\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F}) + O\left(\sqrt{\frac{ln(\frac{1}{\delta})}{m}}\right) \tag{2}$$

where

$$\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\sigma}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m}\sum_i \sigma_i f(z_i)\right)\right]$$

is the empirical value of the Rademacher complexity for Rademacher Random Variables

$$\sigma_i = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

and

$$\mathbb{R}_m(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}\left[\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F})\right]$$

is the expected value of the Rademacher complexity. Furthermore,

$$\mathbb{E}[f] = \mathbb{E}_{z\sim\mathcal{D}}\big[f(z)\big]$$

and

$$\hat{\mathbb{E}}_{\mathcal{S}}[f] = \frac{1}{m}\sum_i f(z_i)$$

2.2 Proof

2.2.1 Step 1

In the previous lecture, we proved that, with probability $\geq 1-\delta$,

$$\Phi(\mathcal{S}) \leq \mathbb{E}_{\mathcal{S}}\big[\Phi(\mathcal{S})\big] + \sqrt{\frac{ln(\frac{1}{\delta})}{m}}$$

where

$$\Phi(\mathcal{S}) = \sup_{f\in\mathcal{F}}\left(\mathbb{E}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]\right)$$

2.2.2 Step 2

We will now show that

$$\mathbb{E}_{\mathcal{S}}\big[\Phi(\mathcal{S})\big] \leq \mathbb{E}_{\mathcal{S},\mathcal{S}'}\big[\sup_{f\in\mathcal{F}}\big(\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]\big)\big] \tag{3}$$

for ghost sample $\mathcal{S}' = \langle z'_1, ..., z'_m\rangle$.

Proof: Observe that:

$$\mathbb{E}_{\mathcal{S}'}[\hat{\mathbb{E}}_{\mathcal{S}'}[f]] = \mathbb{E}[f]$$

and

$$\mathbb{E}_{\mathcal{S}'}[\hat{\mathbb{E}}_{\mathcal{S}}[f]] = \hat{\mathbb{E}}_{\mathcal{S}}[f]$$

where the last identity follows because $\hat{\mathbb{E}}_{\mathcal{S}}[f]$ does not depend on $\mathcal{S}'$, so this is equivalent to taking the expectation of a constant.

Then we can write:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\big[\Phi(\mathcal{S})\big] &= \mathbb{E}_{\mathcal{S}}\big[\sup_{f\in\mathcal{F}}\big(\mathbb{E}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]\big)\big] \\
&= \mathbb{E}_{\mathcal{S}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{S}'}\left[\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]\right]\right)\right] \\
&\leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]\right)\right]
\end{aligned}$$

as required. Here we used our observations to obtain the second equality, and we used the result that, for random variables, $X_1, \ldots, X_n$, $\mathbb{E}[\max_i\{X_1, ..., X_n\}] \geq \max\{\mathbb{E}[X_1], ..., \mathbb{E}[X_n]\}$ to obtain the last inequality. (The proof of this result: $\max\{X_1, ..., X_n\} \geq X_i\ \forall i \implies \mathbb{E}[\max\{X_1, \ldots, X_n\}] \geq \mathbb{E}[X_i]\ \forall i$. The argument generalizes even when working with infinitely many random variables).

2.2.3 Step 3

We would now like to examine the difference of empirical averages, $\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]$, more closely.

Let's create two new sets, T and T', as follows:

for $i = 1, \ldots, m$:
swap z_i and z_i' with probability $\frac{1}{2}$
do nothing else

Let's now examine $\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f]$ (call this (1)) and we note that $\hat{\mathbb{E}}_{\mathcal{T}'}[f] - \hat{\mathbb{E}}_{\mathcal{T}}[f]$ (call this (2)) has the same distribution as (1). This is because all of the samples are i.i.d., so permuting the samples does not change their distribution.

We can rewrite (1) as:

$$\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f] = \frac{1}{m} \sum_i \Big(f(z_i') - f(z_i) \Big)$$

and we can rewrite (2) as:

$$\begin{aligned} \hat{\mathbb{E}}_{\mathcal{T}'}[f] - \hat{\mathbb{E}}_{\mathcal{T}}[f] &= \frac{1}{m} \sum_i \begin{cases} \big(f(z_i) - f(z_i')\big) & \text{if } z_i \text{ and } z_i' \text{ were swapped} \\ \big(f(z_i') - f(z_i)\big) & \text{no swap} \end{cases} \\ &= \frac{1}{m} \sum_i \sigma_i \big(f(z_i') - f(z_i)\big) \end{aligned}$$

where $\sigma_i = \begin{cases} 1 & \text{no swap} \\ -1 & \text{swap} \end{cases}$ is a Rademacher Random Variable. Substituting our new expression for the difference of empirical averages into the right hand side of Equation 3, we obtain that

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'} \big[\sup_{f \in \mathcal{F}} \big(\hat{\mathbb{E}}_{\mathcal{S}'}[f] - \hat{\mathbb{E}}_{\mathcal{S}}[f] \big) \big] = \mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma} \Big[\sup_{f \in \mathcal{F}} \Big(\frac{1}{m} \sum_i \sigma_i \big(f(z_i') - f(z_i) \big) \Big) \Big] \tag{4}$$

2.2.4 Step 4

We would now like to show that

$$\mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma} \Big[\sup_{f \in \mathcal{F}} \Big(\frac{1}{m} \sum_i \sigma_i \big(f(z_i') - f(z_i) \big) \Big) \Big] \leq 2\mathbb{R}_m(\mathcal{F}) \tag{5}$$

We see this by rewriting

$$\begin{aligned} \mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma} \Big[\sup_{f \in \mathcal{F}} \Big(\frac{1}{m} \sum_i \sigma_i \big(f(z_i') - f(z_i) \big) \Big) \Big] &\leq \mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma} \Big[\sup_{f \in \mathcal{F}} \Big(\frac{1}{m} \sum_i \sigma_i f(z_i') \Big) \Big] \\ &\quad + \mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma} \Big[\sup_{f \in \mathcal{F}} \Big(\frac{1}{m} \sum_i (-\sigma_i) f(z_i) \Big) \Big] \qquad (6) \\ &= \mathbb{R}_m(\mathcal{F}) + \mathbb{R}_m(\mathcal{F}) \\ &= 2\mathbb{R}_m(\mathcal{F}) \end{aligned}$$

where, to derive the inequality, we used the fact that $\max(A+B) \leq \max(A) + \max(B)$ and for the first equality, we used the fact that $-\sigma_i$ has the same distribution as σ_i. We also used that the first expectation after the '$\leq$' sign does not depend on sample $\mathcal{S}$ and the second does not depend on sample $\mathcal{S}'$. We have now obtained the first bound of the Theorem.

2.2.5 Step 5

To obtain the second bound in terms of $\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F})$, it's enough to use McDiarmid's inequality to show that $\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F})$ and $\mathbb{R}_m(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}[\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F})]$ are "close" (i.e. their difference is of the same order as $(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}})$ term).

3 Generalization Bounds with Rademacher Complexity

Let's now use the Theorem that we just proved to derive some generalization bounds. Reminder:

$$\begin{aligned}\hat{\text{err}}(h) =& \frac{1}{m}\sum_i \mathbb{1}\{h(x_i) \neq y_i\} \\ =& \hat{\mathbb{E}}_{\mathcal{S}}[\mathbb{1}\{h(x_i) \neq y_i\}]\end{aligned}$$

is the training error and

$$\begin{aligned}\text{err}(h) =& \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \\ =& \mathbb{E}[\mathbb{1}\{h(x_i) \neq y_i\}]\end{aligned}$$

is the generalization error.

Let $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ and $\mathcal{F}_{\mathcal{H}} = \{f_h : h \in \mathcal{H}\}$, where $f_h(x, y) = \mathbb{1}\{h(x) \neq y\}$, for hypothesis space $\mathcal{H}$ and for instance space $\mathcal{X}$. If we plug in the function space $\mathcal{F}_{\mathcal{H}}$ and the space $\mathcal{Z}$ into the theorem we just proved then, with probability $\geq 1-\delta$, $\forall f_h \in \mathcal{F}_{\mathcal{H}}$ (equivalently $\forall h \in \mathcal{H}$, since there is a one-to-one mapping between $\mathcal{F}_{\mathcal{H}}$ and $\mathcal{H}$):

$$\text{err}(h) = \mathbb{E}[f_h] \leq \hat{\mathbb{E}}_{\mathcal{S}}[f_h] + 2\mathbb{R}_m(\mathcal{F}_{\mathcal{H}}) + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right) = \hat{\text{err}}(h) + 2\mathbb{R}_m(\mathcal{F}_{\mathcal{H}}) + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right)$$

or

$$\text{err}(h) \leq \hat{\text{err}}(h) + 2\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F}_{\mathcal{H}}) + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right)$$

if we rewrite in terms of the empirical Rademacher complexity.

But what is the Rademacher complexity, $\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F}_{\mathcal{H}})$?

$$\begin{aligned}\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F}_{\mathcal{H}}) &= \mathbb{E}_\sigma\left[\sup_{f_h \in \mathcal{F}_{\mathcal{H}}}\left(\frac{1}{m}\sum_i \sigma_i f_h(x_i, y_i)\right)\right] \\ &= \mathbb{E}_\sigma\left[\sup_{h \in \mathcal{H}}\left(\frac{1}{m}\sum_i \sigma_i\left(\frac{1 - y_i h(x_i)}{2}\right)\right)\right] \\ &= \frac{1}{2}\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i \sigma_i + \sup_{h \in \mathcal{H}}\left(\frac{1}{m}\sum_i (-\sigma_i y_i) h(x_i)\right)\right] \\ &= \frac{1}{2}\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{H})\end{aligned}$$

In the last step, we used that $\forall i\ \mathbb{E}_\sigma[\sigma_i] = 0$, and that $-\sigma_i y_i$ is distributed like a standard Rademacher variable (-1 or +1 with equal probability).

Pulling everything together, we can finally obtain generalization bounds in terms of Rademacher complexities of the hypothesis space, $\mathcal{H}$:

$$\text{err}(h) \leq \hat{\text{err}}(h) + \mathbb{R}_m(\mathcal{H}) + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right) \tag{7}$$

and

$$\text{err}(h) \leq \hat{\text{err}}(h) + \hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{H}) + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right) \tag{8}$$

4 Evaluating Rademacher Complexities

We see now that we have proved uniform convergence of the training error to the generalization error for every $h \in \mathcal{H}$ in terms of the Rademacher complexity of $\mathcal{H}$. Hence, we will now focus on how to compute $\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{H})$.

4.1 Finite hypothesis spaces

We begin by considering finite hypothesis spaces.

Theorem: (Massart's Lemma) For hypothesis space $\mathcal{H}$ with $|\mathcal{H}| < \infty$ and for binary functions with outputs in $\{-1, 1\}$, and for any $\mathcal{S}$:

$$\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{H}) \leq \sqrt{\frac{2\ln(|\mathcal{H}|)}{m}} \tag{9}$$

We will not prove this result here (we will do so later in the course; a proof is also given in the textbook).

If we now plug this result into the generalization bound we just obtained, we get back:

$$\text{err}(h) \leq \hat{\text{err}}(h) + \sqrt{\frac{2\ln(|\mathcal{H}|)}{m}} + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right) \tag{10}$$

which is a result we showed earlier in class but which we had to obtain by using a customized argument (using Union bound and Chernoff). Here, the result is a consequence of the bound and Rademacher. The $\ln(|\mathcal{H}|)$ complexity measure is subsumed by Rademacher complexity.

4.2 Infinite Hypothesis Spaces

Turning next to infinite hypothesis spaces, we next derive a bound on Rademacher complexity in terms of the growth function.

4.2.1 Bound in terms of Growth Function

$$\hat{\mathbb{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i}\sigma_i h(x_i)\right)\right] \tag{11}$$

Observe: all that matters is what is happening on training samples, $\mathcal{S}$, and while we are still working with an infinite hypothesis space, we can construct

$$\mathcal{H}' = \{\text{one representative from } \mathcal{H} \text{ for each labeling in } \mathcal{S}\}$$

Then note that:

$$|\mathcal{H}'| = |\Pi_{\mathcal{H}}(\mathcal{S})| \leq \Pi_{\mathcal{H}}(m)$$

and that replacing $\mathcal{H}$ with $\mathcal{H}'$ does not change the Rademacher complexity since $\mathcal{H}'$ includes exactly the same behaviors on $\mathcal{S}$ as $\mathcal{H}$. Hence, if we replace $\mathcal{H}$ with $\mathcal{H}'$ in Equation 11:

$$\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}) = \hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}') \leq \sqrt{\frac{2\ln(|\mathcal{H}'|)}{m}} = \sqrt{\frac{2\ln(|\Pi_{\mathcal{H}}(\mathcal{S})|)}{m}} \leq \sqrt{\frac{2\ln(|\Pi_{\mathcal{H}}(m)|)}{m}} \tag{12}$$

and we can write a bound for the generalization error in terms of the growth function.

4.2.2 Bound in terms of VC-dimension

Let $d = \text{VC-dim}(\mathcal{H})$. From Sauer's Lemma, we have:

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$$

for $m \geq d \geq 1$. Plugging this into Equation 12, we get a generalization error bound in terms of the VC-dimension:

$$\text{err}(h) \leq \hat{\text{err}}(h) + \sqrt{\frac{2d\,\ln(\frac{em}{d})}{m}} + O\left(\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}\right) \tag{13}$$

5 Learning Algorithms

We have now finished Part 1 of the course (writing bounds for the generalization error in terms of the training error and complexity measures), and we will now begin Part 2, which introduces learning algorithms. Let's now modify PAC-learning (where the color blue is used to indicate a modification):

Let $\mathcal{C}$ be a concept class. We say $\mathcal{C}$ is *weakly* learnable if $\exists\gamma > 0$, $\exists$ algorithm $\mathcal{A}$, $\forall c \in \mathcal{C}$, $\forall$ distributions $\mathcal{D}$, $\forall\delta > 0$ such that, when $\mathcal{A}$ is given poly$(\frac{1}{\delta}, ...)$ examples, it outputs h such that

$$Pr(err(h) \leq \frac{1}{2} - \gamma) \geq 1 - \delta$$

Whereas in strong learning (the version of PAC learning we have been considering until now), there exists an algorithm that can achieve arbitrarily low error on every concept in the class; in weak learning, the learning algorithm only needs to achieve an error that is slightly better than $\frac{1}{2}$, which is what we would get by guessing randomly.

The question is now whether weak and strong learning are equivalent, that is, whether there are some concept classes that can be weakly learned but not strongly learned, or if it is the case that every class that can be weakly learned can also be strongly learned.