COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire Lecture #9
Scribe: Josh Chen March 5, 2013

We've spent the past few classes proving bounds on the generalization error of PAC-learning algorithms for the cases of consistent and inconsistent hypotheses selected from finite and infinite hypothesis spaces. In particular, last time, we proved bounds for the case of inconsistent hypotheses selected from infinite hypothesis spaces. However, recall that each time we encountered the problem of an infinite hypothesis space, we had to resort to techniques like using ghost samples or the VC-dimension of a concept class. In this lecture, we introduce a more modern and elegant approach, using a concept called *Rademacher complexity.* This approach turns out to include each of the bounds we've proved in the past few lectures as special cases.

1 Definition of Rademacher Complexity

1.1 Some usual definitions

Before getting into the definition of Rademacher complexity, we remind ourselves of the usual setup:

- Let the sample $S = ((x_1, y_1), ..., (x_m, y_m))$ where, unlike before, $y_i = \{-1, +1\}$
- Let the hypothesis $h : X \to \{-1, +1\}$
- To measure how well h fits S, let the training error $\hat{err}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(x_i)\neq y_i}$

Note that, since we are using $y_i = \{-1, +1\}$ instead of $y_i = \{0, 1\}$ as in previous lectures (for simplicity), we can provide an alternative definition of training error:

$$\begin{aligned}
\hat{err}(h) &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\} && (1)\\
&= \frac{1}{m}\sum_{i=1}^{m} \begin{cases} 1 & \text{if } (h(x_i), y_i) = (1, -1) \text{ or } (-1, 1) \\ 0 & \text{if } (h(x_i), y_i) = (1, 1) \text{ or } (-1, -1) \end{cases} && (2)\\
&= \frac{1}{m}\sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2} && (3)\\
&= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} y_i h(x_i) && (4)
\end{aligned}$$

The term $\frac{1}{m}\sum_{i=1}^{m} y_i h(x_i)$ can be interpreted as the *correlation* of the predictions $h(x_i)$ with the labels y_i. We see that correlation is related to training error as *correlation* $= 1 - 2\hat{err}(h)$. To find a hypothesis h that minimizes training error, we can thus equivalently seek to find the h satisfying:

$$\arg\max_{h\in\mathcal{H}} \frac{1}{m}\sum_i y_i h(x_i) \qquad (5)$$

1.2 Playing with correlation

Imagine, now, an experiment where we replace a sample's true labels y_i with the *Rademacher random variables* σ_i:

$$\sigma_i = \begin{cases} +1 & \text{with prob. } 1/2 \\ -1 & \text{with prob. } 1/2 \end{cases} \tag{6}$$

This gives a modified expression for correlation:

$$\arg\max_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i h(x_i) \tag{7}$$

Instead of selecting the hypothesis in $\mathcal{H}$ that correlates best with the labels, this now selects the hypothesis h in $\mathcal{H}$ that correlates best with the random noise variables σ_i. Since h is dependent on the random variables σ_i, however, to measure how well $\mathcal{H}$ can correlate with random noise, we take the expectation of this correlation over the random variables σ_i and find:

$$E_\sigma[\max_{h \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i h(x_i)] \tag{8}$$

This intuitively measures the expressiveness of $\mathcal{H}$. We can bound this expression using two extreme cases: $|\mathcal{H}| = 1$ where we only have one choice for a hypothesis, and $|\mathcal{H}| = 2^m$ where $\mathcal{H}$ shatters S. In the first case, our expectation equals 0 since the max term disappears; in the second case our expectation equals 1 since there always exists a hypothesis matching any set of σ_i's. Thus our measure, as defined above, must fall between 0 and 1.

1.3 Generalizing correlation

Instead of working with hypotheses $h : X \to \{-1, +1\}$, let's generalize our class of functions to the set of all real-valued functions. Replace $\mathcal{H}$ with $\mathcal{F}$, which we define to be any family of functions $f : Z \to \mathbb{R}$. Now, given sample $S = (z_1, ..., z_m)$ with $z_i \in Z$, if we apply our expression from above to $\mathcal{F}$, we arrive at the *empirical Rademacher complexity* of a family of functions $\mathcal{F}$ with respect to a sample S:[1]

$$\hat{\mathcal{R}}_S(\mathcal{F}) := E_\sigma[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_i \sigma_i f(z_i)] \tag{9}$$

Again, this expression measures how well, on average, the function class $\mathcal{F}$ correlates with random noise over the sample S. However, often we want to measure the correlation of $\mathcal{F}$ with respect to a distribution D over X, rather than with respect to a sample S over X. To find this, we take the expectation of $\hat{\mathcal{R}}_S(\mathcal{F})$ over all samples of size m drawn according to D:

$$\mathcal{R}_m(\mathcal{F}) := E[\hat{\mathcal{R}}_S(\mathcal{F})] \tag{10}$$

This is the *Rademacher complexity*, or for clarity, the *expected Rademacher complexity*, of $\mathcal{F}$.

We now have the definitions we need, and are finally ready to present our first generalization bounds based on Rademacher complexity.

[1] Note: Since $\mathcal{F}$ can be the family of all real-valued functions, max may not exist. Thus we use sup instead, which is defined as the least upper bound on the elements in a set. For example, the sup of the set $\{.9, .99, .999, ...\}$ is 1.

2 Generalization bounds based on Rademacher complexity

2.1 Bounds for general function classes $\mathcal{F}$

The following theorem will serve as a very general tool for proving uniform convergence bounds via the concept of Rademacher complexity:

Theorem 1. *Let $\mathcal{F}$ be a family of functions mapping from Z to $[0,1]$, and let sample $S = (z_1, ..., z_m)$ where $z_i \sim D$ for some distribution D over Z. Define $E[f] := E_{Z \sim D}[f(z)]$, and define $\hat{E}_S[f] := \frac{1}{m}\sum_{i=1}^{m} f(z_i)$. With probability $\geq 1-\delta$, for all $f \in \mathcal{F}$:*[2]

$$E[f] \leq \hat{E}_S[f] + 2\mathcal{R}_m(\mathcal{F}) + \mathcal{O}\left(\sqrt{\frac{\ln 1/\delta}{m}}\right) \tag{11}$$

$$E[f] \leq \hat{E}_S[f] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + \mathcal{O}\left(\sqrt{\frac{\ln 1/\delta}{m}}\right) \tag{12}$$

Proof. We derive a bound for $E[f] - \hat{E}_S[f]$ for all $f \in \mathcal{F}$, or equivalently, bound $\sup_{f \in \mathcal{F}}(E[f] - \hat{E}_S[f])$. Note that this expression is a random variable that depends on S. So we want to bound the following random variable:

$$\Phi(S) = \sup_{f \in \mathcal{F}}(E[f] - \hat{E}_S[f]) \tag{13}$$

Step 1: We show, with probability $\geq 1-\delta$, $\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{\ln 1/\delta}{2m}}$. This step allows us to go from working with $\Phi(S)$ to working with $E_S[\Phi(S)]$.

Recall that McDiarmid's inequality states that, if:

$$|f(x_1, ..., x_i, ..., x_m) - f(x_1, ..., x_i', ..., x_m)| \leq c_i \tag{14}$$

then:

$$Pr[f(x_1, ..., x_m) \geq E[f(X_1, ..., X_m)] + \epsilon] \leq \exp(-2\epsilon^2 / \sum_{i=1}^{m} c_i^2) \tag{15}$$

From the definition of $\Phi(S)$, we have:

$$\begin{aligned}\Phi(S) &= \sup_{f \in \mathcal{F}}(E[f] - \hat{E}_S[f]) &(16)\\ &= \sup_{f \in \mathcal{F}}(E[f] - \frac{1}{m}\sum_i f(z_i)) &(17)\end{aligned}$$

Since $f(z_i) \in [0,1]$ for all z_i, changing any one example z_i to z_i' in the training set S will change $\frac{1}{m}\sum_i f(z_i)$ by at most $\frac{1}{m}$. Thus this changing of any one example affects $\Phi(S)$ by at most this amount, implying that $|\Phi((z_1, ..., z_i, ..., z_m)) - \Phi((z_1, ..., z_i', ..., z_m))| \leq \frac{1}{m}$. This fits the condition of McDiarmid's inequality (see (14)) with $c_i = \frac{1}{m}$, so we can apply McDiarmid's inequality and arrive at the bound shown.

[2] Note that the Big-Oh terms in the two expressions have different constants.

Step 2: Define a ghost sample $S' = (z'_1, ..., z'_m), z'_i \sim D$. We show that $E_S[\Phi(S)] \leq E_{S,S'}[\sup_{f\in\mathcal{F}}(\hat{E}_{S'}[f] - \hat{E}_S[f])]$:

$$\begin{aligned} E_S[\Phi(S)] &= E_S[\sup_{f\in\mathcal{F}}(E[f] - \hat{E}_S[f])] && (18)\\ &= E_S[\sup_{f\in\mathcal{F}}(E_{S'}[\hat{E}_{S'}[f]] - \hat{E}_S[f])] && (19)\\ &= E_S[\sup_{f\in\mathcal{F}}(E_{S'}[\hat{E}_{S'}[f] - \hat{E}_S[f]])] && (20)\\ &\leq E_{S,S'}[\sup_{f\in\mathcal{F}}(\hat{E}_{S'}[f] - \hat{E}_S[f])] && (21) \end{aligned}$$

Note that we arrive at (19) since the expected Rademacher complexity $E[f]$ is equal to the expectation over all samples S' of the empirical Rademacher complexity over those S', or $E_{S'}[\hat{E}_{S'}[f]]$. We also arrive at (21) by moving the expectation over S' in (20) outside of the sup; this can be done since the expectation of a max over some function is at least the max of that expectation over that function.

Step 3: We show $E_{S,S'}[\sup_{f\in\mathcal{F}}(\hat{E}_{S'}[f] - \hat{E}_S[f])] = E_{S,S',\sigma}[\sup_{f\in\mathcal{F}}\sum_i \sigma_i(f(z'_i) - f(z_i))]$

We use the ghost sampling technique for this step. In particular, for each pair of elements z_i, z'_i in S, S' respectively, swap the two with probability 1/2. Let the resulting two sets of examples be T, T'. Since S, S' each initially represented iid samples from D, we have that $T, T' \sim S, S'$. This implies:

$$\begin{aligned} \hat{E}_{S'}[f] - \hat{E}_S[f] &\sim \hat{E}_{T'}[f] - \hat{E}_T[f] && (22)\\ &= \frac{1}{m}\sum_i \begin{cases} f(z_i) - f(z'_i) & \text{with prob. } 1/2 \\ f(z'_i) - f(z_i) & \text{with prob. } 1/2 \end{cases} && (23)\\ &= \frac{1}{m}\sum_i \sigma_i(f(z'_i) - f(z_i)) && (24) \end{aligned}$$

Thus the expressions $\sup_{f\in\mathcal{F}}(\hat{E}_{S'}[f] - \hat{E}_S[f])$ and $\sup_{f\in\mathcal{F}}\sum_i \sigma_i(f(z'_i) - f(z_i))$ are equally distributed. The latter depends on an additional set of random variables σ_i, however, so we must take the expectation of the latter over σ as well as S, S'. Taking the expectation of the former over S, S', as well, we arrive at the expression shown.

Step 4: We show $E_{S,S',\sigma}[\sup_{f\in\mathcal{F}}\sum_i \sigma_i(f(z'_i) - f(z_i))] \leq 2\mathcal{R}_m(\mathcal{F})$

$$\begin{aligned} E_{S,S',\sigma}[\sup_{f\in\mathcal{F}}\sum_i \sigma_i(f(z'_i) - f(z_i))] &\leq E_{S,S',\sigma}[\sup_{f\in\mathcal{F}}\sum_i \sigma_i f(z'_i) + \sup_{f\in\mathcal{F}}\sum_i(-\sigma_i)f(z_i))] && (25)\\ &\leq E_{S',\sigma}[\sup_{f\in\mathcal{F}}\sum_i \sigma_i f(z'_i)] + E_{S,\sigma}[\sup_{f\in\mathcal{F}}\sum_i(-\sigma_i)f(z_i))] && (26)\\ &= \mathcal{R}_m(\mathcal{F}) + \mathcal{R}_m(\mathcal{F}) && (27) \end{aligned}$$

where we arrive at (27) because $-\sigma_i$ has the same distribution as σ_i.

Conclusion: Combining all the pieces together, we finally have that, with probability $\geq 1-\delta$, for all $f \in \mathcal{F}$:

$$E[f] - \hat{E}_S[f] \leq 2\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{2m}} \qquad (28)$$

To derive the bound involving $\hat{\mathcal{R}}_S(\mathcal{F})$, we use McDiarmid's inequality again. Recall the definition of $\hat{\mathcal{R}}_S(\mathcal{F}) := E_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_i \sigma_i f(z_i)]$. Since $f \in [0,1]$, changing one element in S changes $\hat{\mathcal{R}}_S(\mathcal{F})$ by at most $\frac{1}{m}$. We can apply McDiarmid's inequality again, finding, with probability $\geq 1-\delta$:

$$\hat{\mathcal{R}}_S(\mathcal{F}) \leq \mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{2m}} \tag{29}$$

Using a $\delta' = \delta/2$ and applying the union bound to (28) and (29), we have our result. With probability $\geq 1-\delta$, for all $f \in \mathcal{F}$:

$$E[f] \leq \hat{E}_S[f] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + \mathcal{O}(\sqrt{\frac{\ln 1/\delta}{m}}) \tag{30}$$

□

2.2 Bounds for hypothesis spaces $\mathcal{H}$

To get from this generalization bound on classes of all real-valued functions to classes of hypotheses, define the following:

$$Z = X \times \{-1,+1\} \tag{31}$$

$$f_h(x,y) = \mathbf{1}\{h(x) \neq y\} \tag{32}$$

$$\mathcal{F}_\mathcal{H} = \{f_h : h \in \mathcal{H}\} \tag{33}$$

Note that, due to (33), each $f_h \in \mathcal{F}_\mathcal{H}$ corresponds to some $h \in \mathcal{H}$. Also note that, by these definitions, we have:

$$err(h) = E_{(x,y)\sim D}[\mathbf{1}\{h(x) \neq y\}] = E[f_h] \tag{34}$$

$$\hat{err}(h) = \frac{1}{m}\sum_i \mathbf{1}\{h(x_i) \neq y_i\} = \hat{E}_S[f_h] \tag{35}$$

Evidently we can use our bound from Theorem 1 to bound $err(h) - \hat{err}(h)$:

$$\hat{\mathcal{R}}_S(\mathcal{F}_\mathcal{H}) = E_\sigma[\sup_{f_h\in\mathcal{F}_\mathcal{H}} \frac{1}{m}\sum_i \sigma_i f_h(x_i,y_i)] \tag{36}$$

$$= E_\sigma[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_i \sigma_i(\frac{1-y_i h(x_i)}{2})] \tag{37}$$

$$= E_\sigma[\frac{1}{2m}\sum_i \sigma_i + \sup_{h\in\mathcal{H}} \frac{1}{2m}\sum_i(-y_i\sigma_i)h(x_i)] \tag{38}$$

$$= \frac{1}{2}E_\sigma[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_i(-y_i\sigma_i)h(x_i)] \tag{39}$$

$$= \frac{1}{2}E_\sigma[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_i \sigma_i h(x_i)] \tag{40}$$

$$= \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{H}) \tag{41}$$

Note that we arrive at (40) since $(-y_i\sigma_i)$ has the same distribution as σ_i. Now, combining (30), (34), (35), and (41), we have:

$$err(h) \leq \hat{err}(h) + \hat{\mathcal{R}}_S(\mathcal{H}) + \mathcal{O}(\sqrt{\frac{\ln 1/\delta}{m}}) \tag{42}$$