COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Vidhya Ramaswamy

Lecture # 9
March 05, 2018

---

1 Fixed Hypothesis Space

The following theorem presents a generalization bound for the error when $|\mathcal{H}|$ is finite.

Theorem 1. *Given m random examples from a distribution D, with probability $1-\delta$, the following holds for all $h \in \mathcal{H}$:*

$$|err_D(h) - \widehat{err}(h)| \leq \epsilon$$

if $m = O(\frac{\ln|\mathcal{H}+\ln(\frac{1}{\delta})|}{\epsilon^2})$.

Proof. To prove this theorem, we use a method similar to previous proofs - we bound the term for a fixed $h \in \mathcal{H}$, and use union bound to bound the probability that any h does not satisfy the required condition. Using Hoeffding's inequality, we know that for a fixed $h \in \mathcal{H}$,

$$\mathsf{Pr}[|err_D(h) - \widehat{err}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 m}$$

Using union bound,

$$\mathsf{Pr}[\exists h \in \mathcal{H}, |err_D(h) - \widehat{err}(h)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 m}$$

We want this to be less than δ. Solving for m, we get

$$m = O(\frac{\ln|\mathcal{H} + \ln(\frac{1}{\delta})|}{\epsilon^2})$$

□

These uniform convergence bounds show that we can expect to get good generalization error, when the training error is low. We also observe the following :

- We notice that, in comparison to the consistent case, this bound is significantly worse - we require $O(\frac{1}{\epsilon^2})$ samples to get similar generalization error, compared to $O(\frac{1}{\epsilon})$ examples in the consistent case, and the rate of convergence of the additional error is $O(\frac{1}{\sqrt{m}})$ as opposed to $O(\frac{1}{m})$.

- This is mainly due to the use of Hoeffding's inequality - the ϵ^2 term carries over to the bounds on m. Moreover, this inequality is tight if the error is close to $\frac{1}{2}$. The bound using relative entropy that was proved in the last class captures this intuition: $RE(p+\epsilon||p)$ is close to ϵ^2 if p is close to $\frac{1}{2}$, and is close to ϵ if p tends to either 0 or 1.

1.1 Dependence of error on training error, complexity and number of examples

In most cases, we only care about providing an upper bound on the generalization error. Theorem 1 gives us the following:
With high probability, $\forall h \in \mathcal{H}$,

$$err(h) \leq \widehat{err}(h) + O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln(\frac{1}{\delta})}{m}}\right)$$

If we view $\ln|\mathcal{H}|$ as a complexity measure of the hypothesis space, we see that this equation demonstrates how the error depends on

- The training error: The lower the training error, the lower the error
- The complexity of $\mathcal{H}$, as measured by $\ln|\mathcal{H}|$: The more complicated the hypothesis is, the more likely we are to overfit the data, causing the generalization error to increase.
- The number of examples: Clearly, the more training examples we have, the better our hypothesis actually fits the distribution.

This bound also demonstrates the tradeoff between how well we fit the training data, versus how complicated our hypothesis space is. Figure 1 demonstrates this tradeoff. Increasing the complexity of the hypothesis allows us to reduce the training error. Initially, this lowers the generalization error as well, but eventually, the complexity of the hypothesis becomes too large, causing the generalization error to increase. This is often called *overfitting*.

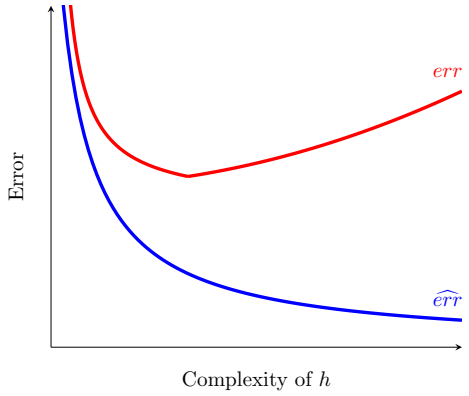

Figure 1: Tradeoff between error and complexity

2 Rademacher Complexity

In the consistent model, we have used various ways to measure the complexity of the hypothesis space, like the size of the hypothesis class, the growth function and the VC dimension. In the remaining of this lecture and the following lecture, we will discuss a new method to measure the complexity of a hypothesis space.

We first remap the labels of our samples from $\{0,1\}$ to $\{-1,1\}$. Hence, we have a set $\mathcal{S} = \langle (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \rangle$ all drawn independently from a distribution D on $(X \times \{-1,1\})$. Given a fixed hypothesis $h : X \to \{-1,1\}$, a natural question to ask is how well this hypothesis fits the data, and we have used $\widehat{err}(h)$ to measure this.

Over the new labels, we can re-write $\widehat{err}(h)$ in an easier way:

$$\begin{aligned}
\widehat{err}(h) &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(\frac{1 - y_i h(x_i)}{2}\right) \\
&= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} y_i h(x_i) \\
\Rightarrow \frac{1}{m}\sum_{i=1}^{m} y_i h(x_i) &= 2\left(\frac{1}{2} - \widehat{err}(h)\right) \\
&= 1 - 2\widehat{err}(h)
\end{aligned}$$

Hence, we can measure how well a hypothesis space $\mathcal{H}$ fits the data set by finding the minimum error possible over all $h \in \mathcal{H}$ or equivalently by finding the maximum possible value of $\frac{1}{m}\sum_{i=1}^{m} y_i h(x_i)$. Mathematically, we can measure how well $\mathcal{H}$ fits the dataset using

$$\max_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} y_i h(x_i)$$

We now consider the following experiment - rather than using the given labels, we use pure random noise. That is, we replace the y_i's with independent random variables σ_i's, called the Rachemacher random variables, where

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

We consider the following quantity:

$$R = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right]$$

The intuitive idea of this definition is that if the hypothesis class is rich, random labels can be fit reasonably well by this class. Hence, minimizing the training error by too much could lead to overfitting.

As a sanity check, we compute R for extreme values of $\mathcal{H}$.

- Suppose $|\mathcal{H}| = 1$. In this case, we know that the maximum is achieved by the only element in $\mathcal{H} = \{h\}$

$$\begin{aligned}
R &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\sigma_i] h(x_i) \\
&= 0
\end{aligned}$$

- Suppose S is shattered by $\mathcal{H}$. In this case, we know that for all values of $\boldsymbol{\sigma}$, there is an $h \in \mathcal{H}$ which correlates perfectly. Hence, $R = 1$.

We notice that R is at least 0 (since $\mathbb{E}(\max f) \geq \max(\mathbb{E}[f])$), and at most 1.

We change our definition to work with arbitrary real valued functions. Let $\mathcal{F}$ be a family of functions, where each $f \in \mathcal{F}$ is defined from some set Z to $\mathbb{R}$, that is, $f : Z \to \mathbb{R}$. Our set $S = \langle z_1, z_2, \ldots, z_m \rangle$ is a set of independent samples from S drawn from some distribution D. We define the empirical Rademacher complexity as

$$\widehat{\mathcal{R}_S}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

We define

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_S[\widehat{\mathcal{R}_S}(\mathcal{F})]$$

as the expected Rademacher complexity. Intuitively, the Rademacher complexity measures how likely a hypothesis class is to overfit a dataset.

We want to prove a uniform convergence result for functions from $\mathcal{F}$. That is, we want to show that, with high probability, $\forall f \in \mathcal{F}$, $\frac{1}{m} \sum_{i=1}^{m} f(z_i)$ converges to $\mathbb{E}_{z \sim D}[f(z)]$. Formally, we prove the following theorem. Let $\hat{\mathbb{E}}_S[f] = \frac{1}{m} \sum_{i=1}^{m} f(z_i)$ and $\mathbb{E}[f] = \mathbb{E}_{z \sim D}[f(z)]$.

Theorem 2. *Let $S = \langle z_1, z_2, \ldots, z_m \rangle$ be random variables drawn independently at random from a distribution D. Let $\mathcal{F}$ be a family of functions defined from Z to $[0, 1]$. With probability at least $1 - \delta$, for all $f \in \mathcal{F}$,*

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

In terms of the empirical Rademacher complexity, we have

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

Moreover, the term $O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$ is generally of a lower order (in both bounds) than the Rademacher complexity.

Proof. We notice that, in order to make a claim about all $f \in \mathcal{F}$, it suffices to bound

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

Step 1

$\Phi(S)$ is a random variable, which is hard to work with. Instead, we would prefer to work with $\mathbb{E}_S[\Phi(S)]$, where the expectation is taken over all possible samples $S \sim D^m$. We can do this provided that $\Phi(S)$ is not too far from its expected value. Hence, we first show that with probability at least $1 - \delta$,

$$\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

To do so, we use McDiarmid's inequality. To use the inequality, we need to ensure that $\Phi(S)$ has the following properties:

- It must be a function of independent random variables. Since

$$\Phi(S) = \Phi(z_1, z_2, \ldots, z_m)$$

and $z_1, z_2, \ldots z_m$ are all independent, this condition is satisfied.

- Perturbation of any one random variable does not change the value of the function. That is, we need to show that

$$|\Phi(z_1, z_2, \ldots, z_i, \ldots, z_m) - \Phi(z_1, z_2, \ldots, z_i', \ldots, z_m)|$$

is bounded. However, since changing z_i to z_i' changes $\widehat{\mathbb{E}}_S[f]$ by at most $\frac{1}{m}$,

$$|\Phi(z_1, z_2, \ldots, z_i, \ldots, z_m) - \Phi(z_1, z_2, \ldots, z_i', \ldots, z_m)| \leq \frac{1}{m}$$

Applying McDiarmid's inequality immediately gives us the desired result.

Step 2

We see that $\mathbb{E}[f]$ is hard to work with, whereas $\widehat{\mathbb{E}}_S[f]$ is much easier to work with. We use the double sample trick, where we choose a ghost sample $S' = \langle z_1', z_2', \ldots, z_m' \rangle$ independently from D^m. We need to show that we can use $\widehat{\mathbb{E}}_{S'}[f]$ rather than $\mathbb{E}[f]$, and hence, we want to prove that

$$\mathbb{E}_S[\Phi(S)] \leq \mathbb{E}_{S,S'}\left[\sup_{f \in \mathcal{F}}(\widehat{\mathbb{E}}_{S'}[f] - \widehat{\mathbb{E}}_S[f])\right]$$

We first make the following observations:

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

$$\mathbb{E}_{S'}[\widehat{\mathbb{E}}_S[f]] = \widehat{\mathbb{E}}_S[f]$$

The first is simply due to the definition of $E[f]$, whereas the second is because the expectation is taken over the ghost sample S', and $\widehat{\mathbb{E}}_S[f]$ does not depend on S'.

Hence, we get

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

In general, it can be shown that $\sup(\mathbb{E}[f]) \leq \mathbb{E}[\sup f]$. Hence,

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

Given the symmetric nature of the equation, for large enough m, we expect this difference to be small, for all functions $f \in \mathcal{F}$.

Step 3

Similar to proving the generalization bound in the consistent model, we randomly swap entries from S and S'. That is, for each $i \in [m]$, we swap z_i and z'_i with probability $\frac{1}{2}$, independently. The rest of the proof will be covered in Lecture 10. $\square$