COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire — Lecture #9
Scribe: Utsav Popat — March 4, 2019

1 Recap

We had started looking at the case where the data points and labels came in pairs, and were drawn from some distribution $\mathcal{D}$. Given a sample of i.i.d. random variables $X_1, \ldots, X_m$, with $X_i \in [0,1]$, we defined $p = \mathbb{E}[X_i]$ and $\hat{p} = \frac{1}{m}\sum_{i=1}^m X_i$ and sought to show that $\hat{p}$ converges uniformly to p. To do this, we proved Hoeffding's Inequality:

$$\Pr\left[\hat{p} \geq p + \varepsilon\right] = \Pr\left[\hat{p} \leq p - \varepsilon\right] \leq e^{-2\varepsilon^2 m}$$

and found stricter bounds for these quantities using Relative Entropy [1] [2] :

$$\Pr\left[\hat{p} \geq p + \varepsilon\right] \leq e^{-\mathrm{RE}(p+\varepsilon \| p)m}$$
$$\Pr\left[\hat{p} \leq p - \varepsilon\right] \leq e^{-\mathrm{RE}(p-\varepsilon \| p)m}$$

2 McDiarmid's Inequality

We now look at a generalisation of Hoeffding's Inequality — McDiarmid's Inequality. While constructing Hoeffding's Inequality, we had considered $\hat{p} = \frac{1}{m}\sum_{i=1}^m X_i$, and had shown that $\frac{1}{m}\sum_{i=1}^m X_i = \hat{p} \to p = \mathbb{E}\left[X_i\right] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m X_i\right]$. Suppose we now wanted to consider a general case where we replace $\hat{p}$ by some function of the sample, $f\left(X_1, \ldots, X_m\right)$. Could we always claim that $f\left(X_1, \ldots, X_m\right) \to \mathbb{E}\left[f\left(X_1, \ldots, X_m\right)\right]$? For this to hold, we need a special property that changing one input to the function f does not change its value by much. Formally, we assume that $\forall i, \forall x_1, \ldots, x_m$ and x_i' (where $x_1, \ldots, x_m, x_i'$ are possible values for the input variables of the function f)

$$|f\left(x_1, \ldots, x_i, \ldots, x_m\right) - f\left(x_1, \ldots, x_i', \ldots, x_m\right)| \leq c_i$$

where c_i is some constant.

Theorem 1 (McDiarmid's Inequality)**.** *Assume $X_1, \ldots, X_m$ are independent (not necessarily identical) random variables, and f is some function that satisfies the property above. Then,*

$$Pr\left[f\left(X_1, \ldots, X_m\right) \geq \mathbb{E}\left[f\left(X_1, \ldots, X_m\right)\right] + \varepsilon\right] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

Hoeffding's Inequality is a special case of McDiarmid's Inequality. We require that the random variables $X_1, \ldots, X_m$ are i.i.d, and $X_i \in [0,1]$. Then, we define $f\left(X_1, \ldots, X_m\right) = \frac{1}{m}\sum_{i=1}^m X_i$. Note that because the X_i's are constrained to be either 0 or 1, changing one of these values will change the value of $f\left(X_1, \ldots, X_m\right)$ by at most $\frac{1}{m}$. So, we set $c_i = \frac{1}{m}$ in McDiarmid's Inequality to get the required result.

[1] While the second inequality was not proven in class, its proof resembles that for the first inequality, using the random variables $1 - X_1, \ldots, 1 - X_m$ instead of $X_1, \ldots, X_m$.

[2] Hoeffding's inequality is a special case of the first inequality, using the identity that $\mathrm{RE}(p+\varepsilon \parallel p) \geq 2\varepsilon^2$.

3 Learning in a Finite Hypothesis Space

Theorem 2. *Let $|\mathcal{H}| < \infty$. Given a sample of m points $\mathcal{S} = \langle x_1, \ldots, x_m \rangle$ from some distribution $\mathcal{D}$, we have that with probability $\geq 1 - \delta, \forall h \in \mathcal{H}$,*

$$|\mathit{err}\,(h) - \widehat{\mathit{err}}\,(h)\,| \leq \varepsilon$$

if $m \geq \frac{\ln 2|\mathcal{H}| + \ln \frac{1}{\delta}}{2\varepsilon^2}$.

Proof. For a fixed hypothesis $h \in \mathcal{H}$, Hoeffding's inequality gives us that $\Pr\left[|\hat{p} - p| > \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$. As we are dealing with a finite hypothesis space, we can use the Union Bound:

$$\Pr\left[\exists h \in \mathcal{H} : |\mathrm{err}\,(h) - \widehat{\mathrm{err}}\,(h)\,| > \varepsilon\right] \leq 2|\mathcal{H}|e^{-2\varepsilon^2 m}$$

Setting the RHS to δ gives us that

$$m = \mathcal{O}\left(\frac{\ln 2|\mathcal{H}| + \ln \frac{1}{\delta}}{2\varepsilon^2}\right). \tag{1}$$

Equivalently, we can say that with probability $\geq 1 - \delta, \forall h \in \mathcal{H}$,

$$\mathrm{err}\,(h) \leq \widehat{\mathrm{err}}\,(h) + \mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln \frac{1}{\delta}}{m}}\right). \tag{2}$$

□

Note that we dropped the two-sided inequality in favour of the one-sided inequality in (2) because for our purposes, it suffices to consider only the direction shown here.

We observe the following from the bounds above:

- The error reduces at a rate of $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$ in (2), compared to a rate of $\mathcal{O}\left(\frac{1}{m}\right)$ when working with consistent hypotheses.
- The amount of data needed in (1) increases from being $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ when working with consistent hypotheses, to $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$. This is reflected in the Relative Entropy version of the inequality as well — when p is close to $\frac{1}{2}$, $\mathrm{RE}(p + \varepsilon \parallel p)$ is close to $\frac{1}{\varepsilon^2}$; otherwise it is close to $\frac{1}{\varepsilon}$ when p is close to 0 or 1.

This distinction arises because of the difference in the upper bounds we are using — we used $e^{-\varepsilon m}$ in the consistency model, while we use $e^{-2\varepsilon^2 m}$ in this case.

Now suppose that we were encoding the hypothesis space $\mathcal{H}$ by bits. Then, we can replace $\ln|\mathcal{H}|$ in the error bound with $|h|$. In this scenario, the inequality (2) manages to capture the three required properties for learning:

- Simplicity versus Complexity: the lower the value of $|h|$, the lower the generalization error $\mathrm{err}\,(h)$.
- Large amount of data: the higher the value of m, the lower the error.
- Good fit to the dataset: the lower the training error $\widehat{\mathrm{err}}\,(h)$, the lower the generalization error.

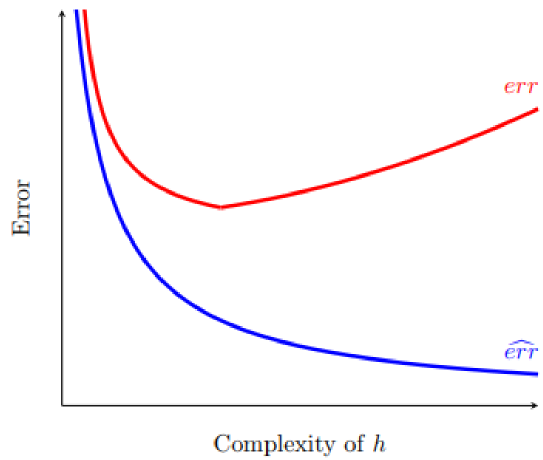

Figure 1: Trade-off between complexity and error

We can plot the errors as a function of the complexity of the hypotheses, and get the graph above[3]. We can see that increasing the complexity reduces the training error — with more complex hypotheses, we can fit the training data better. It reduces the generalization error as well initially; however, we fall prey to overfitting as the complexity increases, causing an increase in generalization error.

4 Learning in an Infinite Hypothesis Space

In previous lectures, we have used complexity measures such as the growth function and VC-dimension to help us prove learnability in an infinite hypothesis space. However, we will now look at a new measure of complexity that subsumes those that we've seen previously — namely, the Rademacher Complexity.

4.1 Rademacher Complexity

We start with a sample of m points $\mathcal{S} = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ where $x_1, \ldots, x_m \in \mathcal{X}$ and $y_1, \ldots, y_m \in \{-1, 1\}$, drawn from some distribution $\mathcal{D}$. Then we can use the training error to measure how well a fixed hypothesis h fits the training data:

$$\begin{aligned}
\widehat{\text{err}}(h) &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\} \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2} \\
&= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(x_i) \\
\implies \frac{1}{m} \sum_{i=1}^{m} y_i h(x_i) &= 1 - 2\widehat{\text{err}}(h)
\end{aligned}$$

[3]Reference: http://www.cs.princeton.edu/courses/archive/spring18/cos511/scribe_notes/0305.pdf

This shows that $\frac{1}{m}\sum_{i=1}^{m} y_i h(x_i)$ can be used as a measure of how well h fits the data set, and that this measure is equivalent to training error.

As the best hypothesis in $\mathcal{H}$ minimizes $\widehat{\text{err}}\,(h)$, we can measure how well the entire hypothesis space $\mathcal{H}$ fits the sample using

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

Now, consider the following experiment: suppose the labels y_i are given at random. We are interested in finding how well $\mathcal{H}$ will fit pure noise. Formally, we replace the labels y_i with independent random variables σ_i (also known as Rademacher random variables) such that

$$\sigma_i = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5. \end{cases}$$

Define

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

Intuitively, we can see that if a hypothesis class is rich enough, then it is more likely to fit the random labels, and hence, have a higher value of R. However, this also exposes us to the dangers of overfitting the given sample.

Let us consider some extreme cases to check for the values of R:

- Suppose $\mathcal{H} = \{h\}$. Then

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

  This is the minimum possible value of R as $\mathbb{E}\left[\max_f f\right] \geq \max_f \mathbb{E}\left[f\right]$ (by the argument given last lecture) — so R can never be negative.

- Suppose $\mathcal{S}$ is shattered by $\mathcal{H}$. Then, we know that for any labelling $\boldsymbol{\sigma}$, there exists a hypothesis $h \in \mathcal{H}$ such that $h\left(x_i\right) = \sigma_i$ for $i = 1, \ldots, m$. In this case, $R = 1$. This is the maximum value R can take.

We will study these topics in a more general and abstract setting. Assume now that we have a family $\mathcal{F}$ of real-valued functions where $f : \mathcal{Z} \to \mathbb{R}$ for some set $\mathcal{Z}$. Let $\mathcal{S} = \langle z_1, \ldots, z_m\rangle$ where $z_1, \ldots, z_m \in \mathcal{Z}$ are independently drawn from some distribution $\mathcal{D}$. We define the empirical Rademacher Complexity as

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

Note that we are now using the supremum (the least upper bound) instead of the maximum in our definition, and that the empirical Rademacher Complexity is defined with respect to a particular sample $\mathcal{S}$. Similarly, we define the expected Rademacher Complexity as

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

We want to prove that $\forall f \in \mathcal{F}, \frac{1}{m}\sum_{i=1}^{m} f(z_i) \to \mathbb{E}_{z\sim\mathcal{D}}[f(z)]$ with high probability. We will make use of the shorthand $\hat{\mathbb{E}}_{\mathcal{S}}[f] = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$ and $\mathbb{E}[f] = \mathbb{E}_{z\sim\mathcal{D}}[f(z)]$.

Theorem 3. *Let $\mathcal{F}$ be a family of functions $f : \mathcal{Z} \to [0,1]$, and suppose $\mathcal{S} = \langle z_1, \ldots, z_m\rangle$ where $z_i \sim \mathcal{D}$. Then, with probability $\geq 1-\delta$*

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

In terms of the empirical Rademacher Complexity, we have

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

Proof. We define

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

and see that it suffices to consider a bound for $\Phi(\mathcal{S})$, as that would apply for all $f \in \mathcal{F}$. The proof consists of three steps. We will show the first step and introduce the second step, but the proof will be completed in the next lecture.

Step 1

$\Phi(\mathcal{S})$ is a random variable that is cumbersome to work with. We would prefer to use the constant $\mathbb{E}_S[\Phi(S)]$. In order to do so, we need to first prove that with probability $\geq 1-\delta$,

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

This inequality can be proven using McDiarmid's Inequality. However, we need to first check whether the conditions for McDiarmid's Inequality are satisfied:

- The inputs to Φ must be independent random variables: As $\Phi(\mathcal{S}) = \Phi(z_1, \ldots, z_m)$, and $z_1, \ldots, z_m$ are independently distributed from $\mathcal{D}$, this condition is satisifed.
- A change in the input should not change the value of Φ by much: If we change z_i for some $i \in \{1, \ldots, m\}$, $\mathbb{E}[f]$ does not change. $\hat{\mathbb{E}}_{\mathcal{S}}[f] = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$, and $z_i \in [0,1]$, so the value of $\hat{\mathbb{E}}_{\mathcal{S}}[f]$ changes by at most $\frac{1}{m}$, and hence, changes the value of $\Phi(S)$ by at most $\frac{1}{m}$. Therefore, setting $c_i = \frac{1}{m}$ in McDiarmid's Inequality gives us the required bound.

Step 2

The term $\mathbb{E}[f]$ is cumbersome to work with as well; we will make use of the double-sampling trick to help us find a replacement. Suppose $\mathcal{S}' = \langle z_1', \ldots, z_m' \rangle$ where z_i' are independently chosen from $\mathcal{D}$. We want to replace $\mathbb{E}[f]$ with $\hat{\mathbb{E}}_{\mathcal{S}'}[f]$. In order to do so, we will prove that

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