COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Changyan Wang

Lecture #8
February 28, 2018

1 Review of The Learning Setting

Last class, we moved beyond the PAC model: in the PAC model we assumed that the data is labeled according to some function, the concept c from some known concept class C. In our new setting, the data and labels could certainly be correlated, but we are no longer assuming the same data is always labeled identically.

In our new setting, we assume that our data points and labels come in pairs (x, y), and the pairs (x, y) are drawn from a distribution D. As in PAC, x is from a domain $\mathcal{X}$ and $y \in \{0, 1\}$. Our "true" error measurement of a hypothesis h is $err_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y]$ because both examples and labels come from a distribution D. Our random "training" sample is a set of random samples $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ which we provide to our learning algorithm. The training error (denoted $\hat{err}(h)$) of a hypothesis h is the fraction of examples from S that it misclassifies:

$$\hat{err}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}$$

where $\mathbb{1}\{\cdot\}$ equals 1 if its argument (in this case, $h(x_i) \neq y_i$) is true, and equals 0 otherwise.

In our previous PAC setting, we were interested in consistent hypotheses, those whose training error is zero. In our current setting, such a consistent hypothesis may not exist, so we are instead interested in the hypothesis $\hat{h}$ which minimizes the training error. That is, $\hat{h} = \arg\min_{h \in H} \hat{err}(h)$. This approach is called "empirical risk minimization" (ERM).

If we can show that, with high probability, the training error of every hypothesis is within ϵ of its "true" error, this is called "uniform convergence". In other words, uniform convergence is when, with probability $\geq 1-\delta$, $\forall h \in H : |err_D(h) - \hat{err}(h)| \leq \epsilon$. Last lecture, we showed that if uniform convergence is true, then $\forall h \in H$ we know with probability $\geq 1-\delta$ that $err_D(\hat{h}) \leq \min_{h \in H} err_D(h) + 2\epsilon$. In other words, if uniform convergence is true, then we know that with high probability, the error of $\hat{h}$ (the hypothesis returned by ERM, which minimizes the training error) is at most 2ϵ more than the true error of the best hypothesis in H.

2 Overview

Thus, if we can prove uniform convergence in a given training setting, we will have a valuable bound on how well the ERM hypothesis $\hat{h}$ performs relative to the truly best hypothesis in the hypothesis class H. In proving a bound for ERM, uniform convergence is the difficult thing to prove, so this lecture focuses on proving uniform convergence. We will begin by trying to prove convergence for a single given hypothesis. In the process, we discuss helpful Chernoff bounds including Hoeffding's Inequality, and we will touch on relative entropy to help us understand Chernoff bounds. Then, we apply Hoeffding's Inequality to derive a useful bound on the ERM hypothesis, and finally we touch on a more general version of Hoeffding's Inequality, called McDiarmid's Inequality.

3 Convergence for a Single Hypothesis

To prove uniform convergence, we first consider convergence for a single hypothesis. That is: given a hypothesis h, can we prove that the training error of h is probably close to the true error of h?

We are interested in the indicator variable $\mathbb{1}\{h(x) \neq y\}$ (whether h is incorrect on an example (x, y)). For a random $(x, y) \sim D$, $\mathbb{1}\{h(x) \neq y\}$ equals 1 with probability $err_D(h)$ and equals 0 otherwise. When estimating the true error based on training error, we are only interested in how often h is incorrect, rather than the details of what the examples (x, y) look like. Consider a coin that lands heads with probability $err_D(h)$ and tails otherwise. Since the examples are randomly drawn from a distribution D, each time we draw a sample it is like flipping that coin. Thus, for a single h, trying to estimate the true error of h from the training error of h is essentially like estimating the bias of that coin based on a sample of flips! Let's drill down to this kind of problem.

Let us represent the coin flips as random variables $X_1, \ldots, X_m$ which are i.i.d., where X_i can take values 0 or 1. We're interested in estimating the probability $p = \mathrm{E}[X_i]$ of getting 1 on any given flip. (Since the X_i are i.i.d. $\mathrm{E}[X_i]$ is the same for all i.) Based on our m samples $X_1, \ldots, X_m$, our natural estimate of p is $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$, the proportion of examples that are equal to 1. How good is $\hat{p}$ compared to p?

Because X_i are random variables, $\hat{p}$ is a random variable. We can imagine the distribution of $\hat{p}$ as some distribution between 0 and 1, with the distribution of p being roughly centered at p. We're interested in how quickly $\hat{p}$ converges to p: the probability $\Pr[\hat{p} \geq p+\epsilon]$ and $\Pr[\hat{p} \leq p-\epsilon]$. Theorems concerning this rate of convergence are sometimes called "tail bounds" or "concentration inequalities".

4 Hoeffding's Inequality

Let us begin with Hoeffding's Inequality[1]. Hoeffding's Inequality states that

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2 m}$$

and similarly $\Pr[\hat{p} \leq p - \epsilon] \leq e^{-2\epsilon^2 m}$. Why are we interested in Hoeffding's Inequality? Once we prove Hoeffding, we can use union bound to combine the two bounds to show that $\Pr[|\hat{p} - p| \geq \epsilon] \leq 2e^{-2\epsilon^2 m}$. If we set $\delta = 2e^{-2\epsilon^2 m}$, then solving for ϵ informs us that with probability $1 - \delta$,

$$|\hat{p} - p| \leq \sqrt{\frac{\ln 2/\delta}{2m}}$$

This implies that $\hat{p}$ converges to p at a rate of $\sqrt{\frac{1}{m}}$!

Applying this to learning, if we let $X_i = \mathbb{1}\{h(x_i) \neq y_i\}$, $p = \mathrm{E}[X_i] = err_D(h)$, and $\hat{p} = \frac{1}{m}\sum_i X_i = \hat{err}(h)$, then, with probability $1 - \delta$:

$$|err_D(h) - \hat{err}(h)| \leq \sqrt{\frac{\ln 2/\delta}{2m}}$$

[1] Hoeffding's Inequality is a "Chernoff bound". A Chernoff bound is an exponentially decreasing tail bound for distributions of sums of independent random variables, so Hoeffding's Inequality is a special case of a Chernoff bound.

which is a bound on the difference between $err_D(h)$ and $\hat{err}(h)$ as we desired. Thus, Hoeffding's Inequality is pretty valuable.

Now we should prove Hoeffding's Inequality, but first we will prove even better bounds, from which Hoeffding's Inequality is a special case. These bounds are:

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

$$\Pr[\hat{p} \leq p - \epsilon] \leq e^{-\mathrm{RE}(p-\epsilon||p)m}$$

where

$$\mathrm{RE}(p + \epsilon||p)$$

is the "relative entropy" of $p + \epsilon$ from p. In order to prove this bound, let's discuss relative entropy, also known as Kullback-Leibler divergence.

5 Relative Entropy

Relative entropy is an idea from information theory. Information theory ideas often arise in machine learning because, broadly speaking, we're often interested in how much information is conveyed through the examples given to us.

We will illustrate relative entropy through an example. Suppose Alice wants to send Bob a message in English. This problem boils down to deciding how to send one letter at a time using bits, so we'd like to decide on an encoding from letters to bits. An obvious option would be to use a code of equal length for every letter. Since there are 26 letters, we'd need 5 bits. So we could encode A as "00000", B as "00001", and so on.

But this isn't a smart way to do it. Some letters are more common than others, so if we're trying to reduce the message length in bits, we could encode common letters with shorter codes and rarer letters with longer codes. For example, we could encode A as "00" and Q as "011001". The question is, if we're trying to minimize message length, what is the optimal way to encode letters as bits?

For a random message M, let each letter x appear with probability $P(x)$. That is, $P(\text{“}A\text{”}) > P(\text{“}Q\text{”})$ because A is more likely to appear in M than Q. The optimal encoding turns out to use $\log_2(\frac{1}{P(x)})$ bits for the letter x. (Ignore the fact that this isn't an integer for now: there are ways to get around that that we won't discuss here.) Given this optimal encoding, what is the expected encoded message length? It should be the expected length of M when encoded. This is:

$$\mathrm{E}[\text{encoded length of } M] = \sum_x P(x) \log_2(\frac{1}{P(x)})$$

This value is known as the "entropy" of P, where P is the distribution of letters. However, this description *assumes* that we know the distribution P before we start encoding. What if we encoded the message based on our estimate Q of the distribution P? Q, too, is a distribution of how often a letter appears in a message, and it may not be exactly equal to P. Suppose we encode our message using Q, when the real distribution is P. Then, the expected message length is $\sum_x P(x) \log_2(\frac{1}{Q(x)})$. We want to compare this value to the entropy $\sum_x P(x) \log_2(\frac{1}{P(x)})$, so we find the difference between these two lengths and call this the "relative entropy". So, the relative entropy is:

$$\sum_x P(x) \log_2(\frac{1}{Q(x)}) - \sum_x P(x) \log_2(\frac{1}{P(x)}) = \sum_x P(x) \log_2(\frac{P(x)}{Q(x)})$$

We will make a small substitution here, and use ln instead of $\log_2$. Natural log ln is more common in machine learning, and only differs by a constant factor from $\log_2$.

One can prove that

$$\mathrm{RE}(P||Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)} \geq 0$$

so relative entropy is always non-negative. One can also prove that relative entropy is positive unless $P = Q$:

$$\sum_x P(x) \ln \frac{P(x)}{Q(x)} = 0 \text{ iff } P = Q$$

Note that we define $0 \ln 0$ as 0, so in relative entropy we ignore letters whose probability $P(x) = 0$. Also note that relative entropy is not a metric, because it is not symmetric. That is, $\mathrm{RE}(P||Q) \neq \mathrm{RE}(Q||P)$ in general. Sometimes, we are interested in the special case where $P(x)$ is over two outcomes, so it only has two probabilities p and $1-p$ and Q only has two probabilities, q and $1-q$. Then, the relative entropy is $\mathrm{RE}((p, 1-p)||(q, 1-q))$ which is simplified as $\mathrm{RE}(p||q)$. Side note: The Alice/Bob message encoding scenario isn't particularly relevant to our purposes, but it is useful to illustrate the meaning of relative entropy.

6 Proving Our Bounds

Let's return to proving $\Pr[\hat{p} \geq p + \epsilon] \leq e^{-\mathrm{RE}(p+\epsilon||p)m}$. We will use a surprisingly weak inequality to accomplish this: Markov's Inequality. Markov's inequality states that, if some random variable $X \geq 0$, then

$$\Pr[X \geq t] \leq \frac{\mathrm{E}[X]}{t}$$

The proof uses the law of total probability:

$$\mathrm{E}[X] = \Pr[X \geq t] \, \mathrm{E}[X|X \geq t] + \Pr[X < t] \, \mathrm{E}[X|X < t]$$

Since $\mathrm{E}[X|X \geq t] \geq t$ and $\Pr[X < t] \geq 0$ and $\mathrm{E}[X|X < t] \geq 0$, we find that

$$\mathrm{E}[X] \geq t \cdot \Pr[X \geq t]$$

which yields Markov's Inequality if we divide by t on both sides.[2]

Here's a first try at a proof of $\Pr[\hat{p} \geq p + \epsilon] \leq e^{-\mathrm{RE}(p+\epsilon||p)m}$ using Markov's Inequality:

Let $q = p + \epsilon$. If we use Markov's inequality immediately, we find that $\Pr[\hat{p} \geq q] \leq \frac{\mathrm{E}[\hat{p}]}{q} = \frac{p}{q} = \frac{p}{p+\epsilon}$, which is a true bound but not very useful, and doesn't even depend on m. To prove our desired result we do something clever before using Markov: we pass both sides of the inequality through a strictly increasing fuction.

If f is strictly increasing, then $\hat{p} \geq q \iff f(\hat{p}) \geq f(q)$. The key step is to let $f(x) = e^{\lambda m x}$, where $\lambda > 0$ is a constant (we'll pick it later). Before applying Markov, we note that $\Pr[\hat{p} \geq q] = \Pr[e^{\lambda m \hat{p}} \geq e^{\lambda m q}]$. We now use Markov:

$$\Pr[e^{\lambda m \hat{p}} \geq e^{\lambda m q}] \leq e^{-\lambda m q} \cdot \mathrm{E}[e^{\lambda m \hat{p}}]$$

[2]Markov's Inequality is pretty weak, as the following example illustrates. Suppose the average US woman's height is 5 feet and 4 inches, that is 64 inches. Then what fraction of women is at least 10 feet (120 inches) tall? Let X be the height of a US woman. Markov's inequality tells us that $\Pr[X \geq 120] \leq \frac{\mathrm{E}[X]}{120} = \frac{64}{120} \approx 53\%$, which is true but quite a weak bound, of course.

Now we simplify. Recall that $\hat{p} = \sum_i X_i \cdot \frac{1}{m}$. Then,

$$\mathrm{E}[e^{\lambda m \hat{p}}] = \mathrm{E}[e^{\lambda m \sum_i X_i \cdot \frac{1}{m}}] = \mathrm{E}[e^{\lambda \sum_i X_i}] = \mathrm{E}[\prod_{i=1}^{m} e^{\lambda X_i}] = \prod_{i=1}^{m} \mathrm{E}[e^{\lambda X_i}]$$

In the last step above, we used the fact that the X_i are independent, so the $e^{\lambda X_i}$ are also independent, meaning the expected value of the product is equal to the product of the expected values. Next, let's use an inequality: if $0 \leq x \leq 1$, then $e^{\lambda x} \leq 1 - x + x \cdot e^{\lambda}$. Plugging this in, we find

$$\mathrm{E}[e^{\lambda m \hat{p}}] = \prod_{i=1}^{m} \mathrm{E}[e^{\lambda X_i}] \leq \prod_{i=1}^{m} \mathrm{E}[1 - X_i + X_i \cdot e^{\lambda}]$$

$$= \prod_{i=1}^{m} (1 - p + p \cdot e^{\lambda}) = (1 - p + p \cdot e^{\lambda})^m$$

We plug this in to find that

$$\Pr[\hat{p} \geq q] \leq e^{-\lambda m q} \cdot (1 - p + p \cdot e^{\lambda})^m = \left(e^{-\lambda q} \cdot (1 - p + p \cdot e^{\lambda})\right)^m$$

and this is true for any λ. We want to find the value of λ that minimizes this value, so if we think of $\left(e^{-\lambda q} \cdot (1 - p + p \cdot e^{\lambda})\right) = \phi(\lambda)$ as a function $\phi(\lambda)$, we want to find the minimum. Setting $\phi'(\lambda) = 0$ yields the following value of λ to minimize $\phi(\lambda)$:

$$\lambda' = \ln\left(\frac{q(1-p)}{(1-q)p}\right)$$

Plugging in and simplifying, we get: $\phi(\lambda') = e^{-\mathrm{RE}(q||p)}$ so

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

Finally, we plug in $q = p + \epsilon$ to find

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

as desired, so we have proven the bound we set out to prove.

We claimed earlier that Hoeffding's Inequality is a special case of the bound above. This is true because it is possible to prove (we do not do it here) that $\mathrm{RE}(p + \epsilon||p) \geq 2\epsilon^2$, which yields Hoeffding's Inequality. Finally, let's return to learning. Our resultant theorem is as follows:

Let H be a finite hypothesis class ($|H| < \infty$). Then, with probability $\geq 1 - \delta$:

$$\forall h \in H : |err_D(h) - \hat{err}(h)| \leq \epsilon$$

so long as we receive at least m examples where

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

We omit the proof for lack of time. The proof is similar to the homework problems and previous proofs, where we use our result for a single hypothesis h, along with union-bound, to prove a result for the hypothesis class H.

7 McDiarmid's Inequality

Having used Hoeffding's Inequality, we should mention McDiarmid's Inequality, a more general form of Hoeffding. It was not used in our proofs today, but it is very useful.

With Hoeffding, we were interested in $\hat{p}$, the average of $X_1, \ldots, X_m$. More abstractly, we can think of $\hat{p}$ as a *function* of $X_1, \ldots, X_m$ so $\hat{p} = f(X_1, \ldots, X_m)$ where f returns the average. Now, what if f were some other function? We would want to prove that $f(X_1, \ldots, X_m)$ converges to $\mathrm{E}[f(X_1, \ldots, X_m)]$ just as we proved that $\hat{p}$ converges to p when $\hat{p}$ is the average.

Proving this isn't always possible, but it is possible in many cases. The specific case that McDiarmid's Inequality deals with is the case that changing one input of f doesn't change the value very much. That is, for any values $x_1, \ldots, x_m$, if we replace x_i with x_i', the value doesn't change by more than some constant c_i:

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

McDiarmid's Inequality states that if the above condition holds for all $x_1, \ldots, x_m$ and x_i', and the random variables X_i are independent, then we can bound the difference between $f(X_1, \ldots, X_m)$ and $\mathrm{E}[f(X_1, \ldots, X_m)]$ as follows:

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

And a similar bound for $\Pr\left[f(X_1, \ldots, X_m) \leq \mathrm{E}[f(X_1, \ldots, X_m)] - \epsilon\right]$. Note that McDiarmid's Inequality assumes that the X_i are independent, but not necessarily identically distributed.

Hoeffding's Inequality is a special case of McDiarmid's Inequality, where $c_i = \frac{1}{m}$ for all i in the case of Hoeffding (because changing one of the X_i changes the average by at most $\frac{1}{m}$ since $X_i \in \{0, 1\}$). Unlike McDiarmid's Inequality, Hoeffding's Inequality assumes that the X_i are identically distributed.