COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire | Lecture #8
Scribe: Haochen Li | February 27, 2019

1 Review

So far, we have been assuming that the true target is from some target class. We will transition to a general case: you might not be able to find a consistent hypothesis.

Here is the new setup:

Data points and labels come in pairs. (x, y) is generated from some underlying distribution D, where $x \in X, y \in \{0, 1\}$. We could also view the data generalization process as a two-step process: it first generates x, it then assigns a label y to x according to some probability distribution conditioned on x, as reflected in the formula, $Pr(x, y) = Pr(x) \cdot Pr(y|x)$.

2 Generalizing PAC Learning

The generalization error of a hypothesis h under our new framework is

$$err_D(h) = Pr_{(x,y)\sim D}[h(x) \neq y].$$

2.1 Bayes Optimal Decision Rule

To better understand the new notion of error, consider the setting where we have the luxury of working with all possible hypotheses h. How low can our error be?

As a toy problem, consider flipping a coin that lands on heads with probability p and lands on tails with probability $1 - p$. If we want to predict the outcome correctly as often as possible, the optimal prediction rule would be:

$$\begin{cases} heads, & \text{if } p > \frac{1}{2} \\ tails, & \text{if } p < \frac{1}{2} \end{cases} \tag{1}$$

It doesn't matter whether we predict heads or tails if $p = \frac{1}{2}$.

By the same reasoning applied to each x separately, when trying to classify x, the optimal decision rule is:

$$h_{opt}(x) = \begin{cases} 1, & \text{if } Pr_D[y = 1|x] > \frac{1}{2} \\ 0, & \text{if } Pr_D[y = 1|x] < \frac{1}{2} \end{cases} \tag{2}$$

It doesn't matter what we predict if $Pr_D[y = 1|x] = \frac{1}{2}$.

This is known as the *Bayes optimal decision rule*, and its error is known as the *Bayes error*, defined as $err(h_{opt}) = \min_{\text{all h}} err_D(h)$. The Bayes error is typically strictly larger than 0, which implies that even if we know everything about the distribution that generates data and label pairs (x, y), we still cannot make perfect predictions.

2.2 PAC Learning in the New Model

The Bayes error helps facilitate our understanding of the new notion of error, but in Bayes error, we had the luxury of working with all possible hypotheses. In the real world, however, we can typically only work with a limited hypothesis space, $\mathcal{H}$, and consider $\min_{h \in \mathcal{H}} err_D(h)$.

We will consider a very natural approach to PAC learning in this new framework. Choose a sample $S = \langle (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \rangle$ where each (x_i, y_i) is generated i.i.d. from some underlying distribution D. We will define the training error or empirical error of a hypothesis h as:

$$e\hat{r}r(h) = \frac{1}{m} \sum_{i=1}^{m} \begin{cases} 1, & \text{if } h(x_i) \neq y_i \\ 0, & \text{otherwise} \end{cases} = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq y_i\} \tag{3}$$

where $1\{h(x_i) \neq y_i\}$ is an indicator variable that is 1 if $h(x_i) \neq y_i$ and 0 otherwise.

Our goal is to find a hypothesis h whose generalization error is almost as good as the generalization error of the best hypothesis in the hypothesis space $\mathcal{H}$, and a natural approach is to pick the hypothesis $\hat{h}$ with the lowest training error. That is, $\hat{h} = \text{argmin}_{h \in \mathcal{H}} e\hat{r}r(h)$.

Suppose we can show that with probablility at least $1 - \delta$, on a single random sample, for every $h \in \mathcal{H}$, $|e\hat{r}r(h) - err_D(h)| \leq \epsilon$. That is,

$$Pr[\forall h \in \mathcal{H}: \quad |e\hat{r}r(h) - err_D(h)| \leq \epsilon] \geq 1 - \delta.$$

This is known as the uniform convergence property. If we can prove such a result, then we will have that:

$$\begin{aligned} err(\hat{h}) &\leq e\hat{r}r(\hat{h}) + \epsilon \quad \text{By Assumption} \\ &\leq e\hat{r}r(h) + \epsilon \quad \hat{h} = \text{argmin}_{h \in \mathcal{H}} e\hat{r}r(h) \\ &\leq err(h) + 2\epsilon \quad \text{By Assumption} \end{aligned} \tag{4}$$

Thus, since this is true for every $h \in \mathcal{H}$, the generalization error of $\hat{h}$ is within 2ϵ of the best hypothesis in the hypothesis space if the uniform convergence property holds.

The next natural goal will be to prove the uniform convergence property. As a first step, we will prove convergence of the training error for a single hypothesis. In the process, we will prove useful Chernoff bounds including Hoeffding's Inequality, and we will also discuss relative entropy (KL-divergence).

3 Convergence for a Single Hypothesis

We will first consider the question that, given a single hypothesis h, how can we prove that with high probability the training error of h is close to the generalization error of h?

The indicator function $1\{h(x_i) \neq y_i\}$ is 1 with probability $err(h)$ and is 0 with probability $1 - err(h)$. We can thus think of $1\{h(x_i) \neq y_i\}$ as a coin flip that is heads with probability $err(h)$ and tails with probability $1 - err(h)$. In this view, the training error is the fraction of times that the coin comes up heads in m flips, and our goal is to prove that this empirical estimate of the bias converges rapidly to the actual bias of the coin.

To study this, we consider the abstract setting in which there are m i.i.d. random variables $X_1, X_2, \ldots, X_m$ where $X_i \in [0, 1]$, each with mean $p = E[X_i]$, and we want to show that their empirical average $\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i$ converges rapidly to p. Our aim will be to show that $Pr[\hat{p} \geq p + \epsilon]$ goes to zero, which implies that $\hat{p}$ converges to p. In our particular setting, we would then choose $X_i = 1\{h(x_i) \neq y_i\}$ so that $p = err(h)$ and $\hat{p} = e\hat{r}r(h)$, which would let us conclude that the training error of h converges to its generalization error.

3.1 Kullback-Leibler Divergence

We will first discuss a concept known as the Kullback-Leibler Divergence or relative entropy, which will be central to the result that follows.

The notion of Kullback-Leibler divergence comes from information theory. Consider the simple setting that Alice wants to send Bob a letter of the alphabet, but she needs to send the letter in binary. A naive way is to use 5 bits to represent each letter. However, some letters in English appear much more frequently than others, so it will be better to encode a frequently appearing letter (like "a" and "e") with fewer bits, and less frequent letters (like "q") with more bits. More generally, let x denote a message, let X denote the space of messages, and suppose $P(x)$ is the probability of a message x appearing. It turns out that the optimal number of bits to encode x is $\lg(\frac{1}{P(x)})$ (the lg here indicates logorithm base 2). So the expected length of an encoded message is:

$$\sum_{x \in X} P(x) \cdot \lg\left(\frac{1}{P(x)}\right) \tag{5}$$

This is known as the entropy of distribution P. It measures how speard out the distribution is (the more spread the distribution is, the longer the expected length of the encoded messages).

Suppose that Alice and Bob mistakenly think that the distribution of the message is Q, so that they instead use $\lg\left(\frac{1}{Q(x)}\right)$ bits to encode message x. If they make this mistake, the expected message length will be $\sum_{x \in X} P(x) \cdot \lg\left(\frac{1}{Q(x)}\right)$. We can compare this to the optimal expected message length (i.e. the entropy); the difference is:

$$\begin{aligned} &\sum_{x \in X} P(x) \cdot \lg\left(\frac{1}{Q(x)}\right) - \sum_{x \in X} P(x) \cdot \lg\left(\frac{1}{P(x)}\right) \\ =&\sum_{x \in X} P(x) \lg\left(\frac{P(x)}{Q(x)}\right) \end{aligned} \tag{6}$$

This is known as the KL-divergence or the relative entropy, written as $\mathrm{RE}(P||Q)$. This is a way to measure the difference between distribution P and Q. Note that $\mathrm{RE}(P||Q) \geq 0$ and $\mathrm{RE}(P||Q) = 0$ if and only if $P = Q$. However, the relative entropy is not a distance because it is not symmetric, that is, $\mathrm{RE}(P||Q) \neq \mathrm{RE}(Q||P)$. When we are dealing with a distribution over two outcomes(like a coin flip), we use the notation $RE(p||q)$ as shorthand for $RE((p, 1-p)||(q, 1-q))$.

Also, we used log base-2 above since we wanted to measure coding length in bits. But henceforth, we will switch to measuring entropy and relative entropy using the natural logarithm which is mathematically more convenient. The definitions are exactly the same, except lg is replaced by ln.

3.2 Hoeffding's Inequality

In order to prove the convergence for a single hypothesis, we will prove a useful Chernoff bound, called Hoeffding's inequality:

$$\begin{aligned} Pr[\hat{p} \geq p + \epsilon] &\leq e^{-2\epsilon^2 m} \\ Pr[\hat{p} \leq p - \epsilon] &\leq e^{-2\epsilon^2 m} \end{aligned} \tag{7}$$

Once we prove Hoeffding's Inequality, we can apply the union bound to the two inequalities above to get $Pr[|\hat{p} - p| \geq \epsilon] \leq 2e^{-2\epsilon^2 m}$. Setting $2e^{-2\epsilon^2 m} = \delta$, we will be able to prove the convergence result for a single hypothesis. That is, with probability at least $1 - \delta$, $|\hat{err}(h) - err(h)| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}$.

It turns out that Hoeffding's Inequality is just a special case of an even better bound. Given m i.i.d. random variables $X_1, X_2, \cdots, X_m$ with $X_i \in [0,1]$, $p = E[X_i], \hat{p} = \frac{1}{m}\sum_{i=1}^m X_i$, and $\epsilon > 0$:

$$\begin{aligned} Pr[\hat{p} \geq p + \epsilon] &\leq e^{-\mathrm{RE}(p+\epsilon||p)\cdot m} \\ Pr[\hat{p} \leq p - \epsilon] &\leq e^{-\mathrm{RE}(p-\epsilon||p)\cdot m} \end{aligned} \tag{8}$$

We will prove this bound using a very weak inequality, called Markov's inequality, which states that given a nonnegative random variable X, $Pr[X \geq t] \leq \frac{E[X]}{t}$. The proof of Markov's inequality is the following:

$$E[X] = Pr[X \geq t] \cdot E[X|X \geq t] + Pr[X < t] \cdot E[X|X < t] \tag{9}$$

Since

$$\begin{aligned} &Pr[X < t] \geq 0 \\ &E[X|X < t] \geq 0 \\ &E[X|X \geq t] \geq t \\ &E[X] \geq Pr[X \geq t] \cdot E[X|X \geq t] \geq Pr[X \geq t] \cdot t \end{aligned} \tag{10}$$

We get $Pr[X \geq t] \leq \frac{E[X]}{t}$.

We will now proceed to prove the first bound in (8). Let $q = p + \epsilon$. Our goal is to get a bound on $Pr[\hat{p} \geq q]$. If f is a strictly increasing function, then $\hat{p} \geq q \Leftrightarrow f(\hat{p}) \geq f(q)$. Let $f(x) = e^{\lambda m x}$, where $\lambda > 0$ will be chosen later. Note that $f(x)$ is a strictly increasing function.

$$\begin{aligned} Pr[\hat{p} \geq q] &= Pr[e^{\lambda m \hat{p}} \geq e^{\lambda m q}] \\ &\leq e^{-\lambda m q} E[e^{\lambda m \hat{p}}] \quad \text{By Markov's Inequality} \end{aligned} \tag{11}$$

We will now compute $E[e^{\lambda m \hat{p}}]$.

$$\begin{aligned} E[e^{\lambda m \hat{p}}] &= E[e^{\lambda \sum_{i=1}^m X_i}] \\ &= E[\prod_{i=1}^m e^{\lambda X_i}] \\ &= \prod_{i=1}^m E[e^{\lambda X_i}] \quad \text{Since } X_i\text{'s are independent} \end{aligned} \tag{12}$$

Next, we will use an inequality: if $0 \leq x \leq 1$, then $e^{\lambda x} \leq 1 - x + x \cdot e^{\lambda}$. Figure 1 is an illustration of this inequality for $\lambda = 1$.

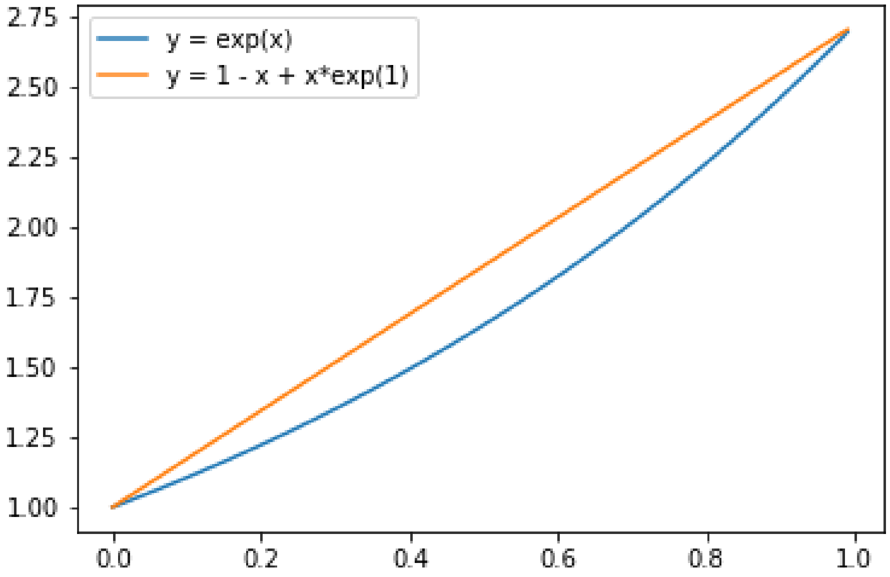

Figure 1: if $0 \le x \le 1$, then $e^{\lambda x} \le 1 - x + x \cdot e^{\lambda}$ for $\lambda = 1$

$$\begin{aligned} E[e^{\lambda m \hat{p}}] &= \prod_{i=1}^{m} E[e^{\lambda X_i}] \\ &\le \prod_{i=1}^{m} E[1 - X_i + X_i \cdot e^{\lambda}] \\ &= \prod_{i=1}^{m} [1 - p + p \cdot e^{\lambda}] \quad E[X_i] = p \\ &= [1 - p + p \cdot e^{\lambda}]^m \end{aligned} \tag{13}$$

Combining (11) and (13), which gives:

$$\begin{aligned} Pr[\hat{p} \ge q] &\le e^{-\lambda m q} E[e^{\lambda m \hat{p}}] \\ &\le e^{-\lambda m q} [1 - p + p \cdot e^{\lambda}]^m \\ &= [e^{-\lambda q}(1 - p + p \cdot e^{\lambda})]^m \end{aligned} \tag{14}$$

The above bound holds for all $\lambda > 0$. To find the value of λ that gives the best bound, we differentiate $[e^{-\lambda q}(1-p+p \cdot e^{\lambda})]^m$ and set the derivative to 0, which gives $\lambda_{min} = \ln\left(\frac{q(1-p)}{p(1-q)}\right)$. Plugging λ_{min} into $[e^{-\lambda q}(1 - p + p \cdot e^{\lambda})]^m$, we get $e^{-\mathrm{RE}(q||p) \cdot m}$. Since $q = p + \epsilon$, we get

$$Pr[\hat{p} \ge p + \epsilon] \le e^{-\mathrm{RE}(p+\epsilon||p) \cdot m} \tag{15}$$

We have thus proven the bound in (8).