COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Sina Jafarpour

Lecture #8
February 27, 2008

1 Outline

In the previous lecture, we saw that PAC learning is not enough to model real learning problems. We may not desire or even may not be able to find a hypothesis that is consistent with the training data. So we introduced a more general model in which the data is generated as a pair (x, y) from an unknown distribution D. And we defined the *generalization error* to be $err_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y]$ and also the *empirical error* of the training set $x_1, ..., x_m$ to be the fraction of mistakes on the training set . $e\hat{r}r(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i|$. We also showed that in order to get an appropriate bound for the generalization error, it is enough to show that $|err(h) - e\hat{r}r(h)| \leq \epsilon$. And today, we will introduce some powerful tools to find the desired bounds.

2 Relative Entropy and Chernoff Bounds

In order to find bounds for the error in this new model, first, we prove a set of more general bounds called **Chernoff Bounds**:

Suppose $X_1, ..., X_m$ are m *i.i.d* random variables such that $\forall i : X_i \in [0, 1]$. Let p be the common[1] expected value of X_i. Also define a new variable $\hat{p}$ to be the average of the value of these random variables:

$$\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i. \tag{1}$$

$\hat{p}$ is called the *empirical average* of the random variables $X_1, ..., X_m$. In our learning model p will be the generalization error and $\hat{p}$ the training error.

So, Chernoff bounds are general bounds on how fast $\hat{p}$ approaches to p. One example of Chernoff bounds is Hoeffding's inequality that we saw in the last lecture. Today, we are going to prove some much stronger results. However, before that, we are going to define one important way to measure the distance between two probability distributions called *Relative Entropy.*

2.1 Relative Entropy

Entropy is one of the most fundamental concepts in Information Theory. Suppose Alice wants to send a letter of the alphabet to Bob over the internet, so she has to encode the letter in bits. One obvious way to do this is to encode each letter of the alphabet with 5 bits, and since the number of letters of the alphabet is 26, 5 bits is enough. But this method is wasteful. In English the letter **e** is more frequnent than the letter **q**. So one smarter way she can use, is for example to encode the letter **e** with two bits, and the letter **q** with 7 bits. Then on average, she uses less bits to send a message to Bob.

[1] Since these random variables are i.i.d they have the same excpected value.

So by using less bits for more common letters and more bits for less common letters, the expected number of bits to send decreases. Now suppose the probability of sending each message x is $P(x)$. In information theory, it can be proved that the optimal way of coding a message with probability $P(x)$ is to use $\lg \frac{1}{P(x)}$ code length for x.
So the expected code length will be:

$$E[codelength] = \sum_x P(x) \lg \frac{1}{P(X)}. \tag{2}$$

This quantity is called *entropy.* It is obvious that the entropy is always non-negative, and also since entropy is the optimal way of sending a message, it cannot exceed $\lg(\#messages)$ (which is the naive way of sending messages). Entropy is a way to measure how spread out our distribution is. The larger the entropy, the closer the distribution to a uniform random distribution.

Now, suppose Alice and Bob think that their letter is from the French alphabet. So they think the message x has some probability *Q(x).* But in fact, the message is from the English alphabet and hence, it has distribution *P(x).* Alice thinks the distribution is $Q(x)$, so she uses $\lg \frac{1}{Q(x)}$ bits to send each message. Hence the expected codelength will be :

$$E[codelength] = \sum_x P(x) \lg \frac{1}{Q(x)}. \tag{3}$$

So Alice is using a sub-optimal way to sending messages.

The *Relative Entropy* of two probability distributions P and Q, also called *Kullback-Libler divergence*, which has the notations $RE(P\|Q)$, $D(P\|Q)$, and $KL(P\|Q)$ is a way to measure the distance between the two probability distributions P, Q. It says that on average how much worse is the code length if we use the distribution Q, than the optimal code length:

$$RE(P\|Q) = \sum_x P(x) \lg \frac{1}{Q(x)} - \sum_x P(x) \lg \frac{1}{P(x)} = \sum_x P(x) \lg \frac{P(x)}{Q(x)}. \tag{4}$$

And since we cannot do better than the optimal case always

$$RE(P\|Q) \geq 0. \tag{5}$$

Also, in order to have a well defined definition, we define $0 \lg 0 = 0$ and $0 \lg \frac{0}{0} = 0$.

However, there are two draw-backs with the relative entropy. Realtive entropy is not symmetric, $RE(P\|Q) \neq RE(Q\|P)$. So relative entropy is not a metric, it is just a way to measure the distance between two probability distributions. Also, relative entropy can be unbounded, for example, if there exists one x such that $P(x) > 0$ and $Q(x) = 0$. Also since lg and ln are equal up to a constant we will use ln instead of lg to simplify our calculations.

Moreover, to keep the notation nicer, we can write the entropy of two Bernoulli distribution $(p, 1-p)$ and $(q, 1-q)$ as follows:

$$RE(p\|q) = RE((p, 1-p)\|(q, 1-q)) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}. \tag{6}$$

2.2 Chernoff Bounds

The most general form of Chernoff bounds is:

Theorem 1. ***(Chernoff Bounds)*** *Assume random variables $X_1, ..., X_m$ are i.i.d, and $\forall i : X_i \in [0, 1]$. Let $p = E[X_i]$ and $\hat{p} = \frac{1}{m}\sum_i X_i$, then $\forall \epsilon > 0$:*

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-RE(p+\epsilon\|p)m}. \tag{7}$$

$$\Pr[\hat{p} \leq p - \epsilon] \leq e^{-RE(p-\epsilon\|p)m}. \tag{8}$$

Before proving the strong Chernoff Bounds, we are going to prove a weak result called Markov's inequality.

Theorem 2. ***(Markov's inequality)*** *Suppose $X \geq 0$ is a random variable and $t > 0$ is a real number, then*

$$\Pr[X \geq t] \leq \frac{E[X]}{t}. \tag{9}$$

Proof:

$$\begin{aligned} E[x] &= \Pr[X \geq t] \cdot E[X|X \geq t] + \Pr[X < t] \cdot E[X|X < t] \\ &\geq t \cdot \Pr[X \geq t] + 0 \end{aligned}$$

which immediately implies the inequality. However this inequality is very weak. If we try to find a bound for $\Pr[\hat{p} \geq p + \epsilon]$ we get:

$$\Pr[\hat{p} \geq p + \epsilon] \leq \frac{E[\hat{p}]}{p+\epsilon} = \frac{\frac{\sum_i E[X_i]}{m}}{p+\epsilon} = \frac{p}{p+\epsilon}. \tag{10}$$

This bound is near one, and also independent of m, so this bound is absolutely useless. A great observation makes it possible for us to use the Markov's inequality. The function $f(x) = e^{ax}$ is one-to-one and strictly increasing. So for any $a > 0$, $e^{ax} \geq e^{ay}$ iff $x \geq y$.
Let $q = p + \epsilon$ and $\lambda > 0$ then $\hat{p} \geq q$ iff $e^{\lambda m \hat{p}} \geq e^{\lambda m q}$. We have:

$$\Pr[\hat{p} \geq q] = \Pr[e^{\lambda m \hat{p}} \geq e^{\lambda m q}]. \tag{11}$$

By Markov's inequality

$$\begin{aligned} \Pr[e^{\lambda m \hat{p}} \geq e^{\lambda m q}] &\leq e^{-\lambda m q} E[e^{\lambda m \hat{p}}] \\ &= e^{-\lambda m q} E[e^{\lambda \sum X_i}] \\ &= e^{-\lambda m q} E\left[\Pi e^{\lambda X_i}\right]. \end{aligned}$$

However the X_is are independent, so:

$$e^{-\lambda m q} E\left[\prod e^{\lambda X_i}\right] = e^{-\lambda m q} \prod E[e^{\lambda X_i}]. \tag{12}$$

Now since $X_i \in [0,1]$ always, $e^{\lambda x} \leq 1 - x + xe^{\lambda}$ and so:

$$\begin{aligned} e^{-\lambda mq} \prod E[e^{\lambda X_i}] &\leq e^{-\lambda mq} \prod E[1 - X_i + X_i e^{\lambda}] \\ &= e^{-\lambda qm}(1 + p + pe^{\lambda})^m \\ &= e^{\phi(\lambda)m} \end{aligned}$$

where we define $\phi(\lambda) = \ln[e^{-\lambda q}(1 - p + pe^{\lambda})]$.
So we should find a λ to minimize $\phi(\lambda)$. By solving $\frac{d\phi}{d\lambda} = 0$ we get that $\phi(\lambda)$ minimizes when $\lambda_{min} = \ln\left(\frac{q(1-p)}{(1-q)p}\right)$ and by pluging in λ_{min} to $\phi()$ we get $\phi(\lambda_{min}) = -RE(q\|p)$ and so:

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-RE(p+\epsilon\|p)m}. \tag{13}$$

This completes the proof of the theorem.

By proving this upper bound, we can simply prove the corresponding lower bound. Let $X_i \leftarrow 1 - X_i$ so that $p \leftarrow 1 - p$ and $\hat{p} \leftarrow 1 - \hat{p}$. Then:

$$\Pr[\hat{p} \leq p - \epsilon] = \Pr[1 - \hat{p} \geq 1 - p + \epsilon] \leq e^{-RE(1-p+\epsilon\|1-p)m} = e^{-RE(p-\epsilon\|p)m}. \tag{14}$$

Other bounds come from bounding the relative entropy. For example Hoeffding's inequality comes directly from the fact that $RE(p + \epsilon\|p) \geq 2\epsilon^2$. Or we can obtain the following multiplicative inequalites:

$$\forall\gamma \in [0,1] : \Pr[\hat{p} \geq p + \gamma p] \leq e^{-mp\gamma^2/3} \tag{15}$$

$$\forall\gamma \in [0,1] : \Pr[\hat{p} \leq p - \gamma p] \leq e^{-mp\gamma^2/2} \tag{16}$$

2.3 McDiarmid's inequality

Finally, we are going to state McDiarmid's inequality, which is a generalization of Hoeffding's inequality.

Theorem 3. *Given a function f for which*

$$\forall x_1, ..., x_m, x_i' : |f(x_1, ..., x_i, ..., x_m) - f(x_1, ..., x_i', ..., x_m)| \leq c_i \tag{17}$$

and given $X_1, ..., X_m$ independent but not necessarily identically distributed random variable. Then:

$$\Pr[f(x_1, ..., x_m) \geq E[f(x_1, ..., x_n)] + \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum {c_i}^2}\right). \tag{18}$$

To show that Hoeffding's inequality is a special case of this inequality, let $f(x_1, ..., x_m) = \frac{1}{m}\sum x_i = \hat{p}$ and $E[f(x_1, ..., x_m)] = p$. Then clearly

$$\begin{aligned} \forall x_1, ...x_m, x_i' : |f(x_1, ...x_i, ...x_m) - f(x_1, ..., x_i', ..., x_m)| &\leq \frac{1}{m}|x_i - x_i'| \\ &\leq \frac{1}{m} \end{aligned}$$

so $c_i = \frac{1}{m}$ for all i. Applying McDiarmid's inequality we get:

$$\Pr[\hat{p} \geq p + \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_i \frac{1}{m^2}}\right) = \exp\left(-2\epsilon^2 m\right). \tag{19}$$

3 Bounding the generalization error

Now we are going to use Hoeffding's inequality to find a bound for the generalization error of the hypothesis. All of the results that we will find can be generalized to infinite $\mathcal{H}$, by replacing $\ln|\mathcal{H}|$ with the VC-dimention of $\mathcal{H}$, but to make it simpler, we are not going to deal with that.

Theorem 4. *Given examples $x_1, ..., x_m$, with probability at least $1-\delta$ if $m = O\left(\frac{\ln|\mathcal{H}|+\ln\frac{1}{\delta}}{\epsilon^2}\right)$ then:*

$$\forall h \in \mathcal{H} : |err(h) - \hat{err}(h)| \leq \epsilon. \tag{20}$$

Proof: First we fix h, and define the following indicator variables: X_i to be 1 if $h(x_i) \neq y_i$ and 0, otherwise. It is obvious that $err(h) = E[X_i]$ and $\hat{err}(h) = \frac{1}{m}\sum X_i$. Then by applying Hoeffding's inequality we get:

$$\Pr[|err(h) - \hat{err}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 m} \tag{21}$$

and so by the union bounds (as $\exists$ is a big or)

$$\Pr[\exists h \in \mathcal{H} : |err(h) - \hat{err}(h)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 m}. \tag{22}$$

We want this bound to be less than δ so:

$$2|\mathcal{H}|e^{-2\epsilon^2 m} \leq \delta \tag{23}$$

which hold if

$$m \geq \frac{\ln(2|\mathcal{H}|) + \ln(\frac{1}{\delta})}{2\epsilon^2}. \tag{24}$$

So given some number of examples, the error we expect is, with probability at least $1-\delta$,

$$err(h) \leq \hat{err}(h) + \sqrt{\frac{\ln(2|\mathcal{H}|) + \ln(\frac{1}{\delta})}{2m}}. \tag{25}$$

This means:

$$err(h) \leq \hat{err}(h) + O\left(\sqrt{\frac{\ln(2|\mathcal{H}|) + ln(\frac{1}{\delta})}{m}}\right). \tag{26}$$

This formula, explains the essence of the three conditions for learning:

- Large amount of training data: by increasing m, the second term becomes smaller and hence the total error decreases.
- Low training error: This is the first term of the summation. By decreasing the training error, the generalization error also decreases.
- Simple hypothesis space. The measure for simplicity is the size of the hypothesis, as measured by $\ln(|\mathcal{H}|)$. If we make this smaller, the generalization error decreases.

Finally, there are two comments on the formula that we obtained for the generalization error:

First, we saw that in the PAC model, sample size m depends on ϵ but here the sample size depends on ϵ^2, and in practice we see that when we cannot find a consistent hypothesis,

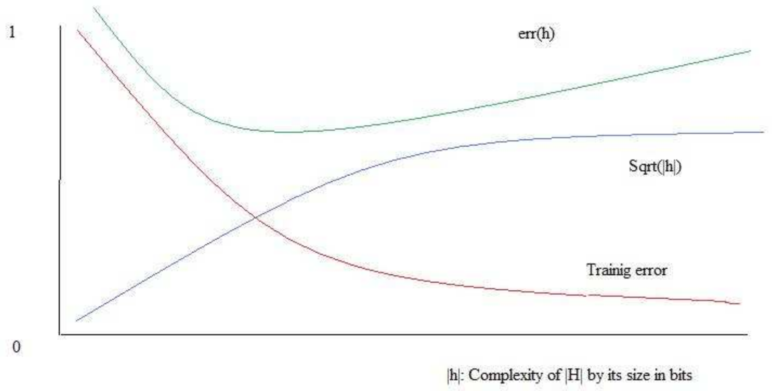

Figure 1: empirical error vs. generalization error

we need more data to obtain a hypothesis with less generalization error, and the fact that sample size depends on ϵ^2 is something real.

Second, as we can see, there exists a trade-off between decreasing the traing error and keeping the hypothesis simple. As the complexity of the hypotheses increases, the probability of finding a consistent hypothesis increases and so $e\hat{r}r(h)$ approaches 0. However, at some point, the O term begins to dominate and $err(h)$ reaches a minimum after which it begins to rise again. This is called *overfitting.*

Overfitting is one of the hardest and most important issues in practical machine learning to deal with. The major difficulty with overfitting is because in many cases only the training error $e\hat{r}r(h)$ can be observed directly. There are at least three main approaches to solving this problem that are common in machine learning:

- *Structural Risk Minimization*: The main idea in this approach is to try to find the exact value of the theoretical bounds to minimize the bound directly.

- *Cross-Validation*: This approach separates the training data to two segmants, one for training the hypothesis, and the other for testing the obtained hypothesis. As a result, we can estimate the generalization error, and have an idea on when to stop the algorithm.

- *New Algorithms*: This approach tries to find a hypothesis, that resists overfitting. It is not clear whether such an algorithm can exist at all! However, we will next study two practical algorithms that seem to have this property.