COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Ugne Klibaite

Lecture #8
February 27, 2013

1 Review From Last Time:

Last class, we discussed the relationship between training and generalization error, and how we could relate the two.

$$
\begin{aligned}
(x, y) &\sim D \\
err_D(h) &= Pr_{(x,y)\sim D}[h(x) \neq y] \\
e\hat{r}r(h) &= \frac{1}{m}\sum_{i=1}^{m} 1\{h(x_i) \neq y_i\}
\end{aligned}
$$

Here, err_D is our true or generalization error, which is the error over all samples in the distribution D. $e\hat{r}r(h)$, on the other hand, is the training error and is calculated by finding the number of incorrectly labeled examples in the training data. We are interested in relating the two in order to see how a given hypotheses will perform on unseen data after traning on a sample set.

If we specify that the training set is $(x_1, y_1), ...(x_m, y_m)$ and **if** we can show that with probability $\geq 1 - \delta$:

$$
\begin{gathered}
\forall h \in (H) : |err(h) - e\hat{r}r(h)| \leq \epsilon \\
\text{and } \textbf{if} \text{ we can find } \hat{h} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ e\hat{r}r(h) \\
\textbf{then}\text{: } err(\hat{h}) \leq \min_{h\in\mathcal{H}} err(h) + 2\epsilon
\end{gathered}
$$

If our assumptions hold then this means that if we can find the hypothesis in $\mathcal{H}$ that minimizes the training error, we can bound the true or generalization error of this hypothesis to the minimum true error over all hypotheses in $\mathcal{H}$ plus 2ϵ. In order to prove this we will use a special case of Chernoff bound, Hoeffding's inequality.

For now we are focusing on the statistical problem of showing that the training error is close to the generalization error without worrying about computation.

As mentioned at the end of last class, let's define the random variables $X_1, ..., X_m$, i.i.d. $X_i \in [0, 1]$ for all $i = 1, ..., m$:

$$
\begin{aligned}
p &= E[X_i] \\
\hat{p} &= \frac{1}{m}\sum_{i=1}^{m} X_i
\end{aligned}
$$

2 Hoeffding's Inequality

Hoeffding's Inequality states that:

$$Pr[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2 m} \tag{1}$$

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

Combining these two inequalities using the union bound:

$$Pr[|\hat{p} - p| > \epsilon] \leq 2e^{-2\epsilon^2 m} = \delta \tag{3}$$

Another way to say this is:
with probability $\geq 1 - \delta$,

$$|\hat{p} - p| = |e\hat{r}r(h) - err(h)| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}$$

which means that the difference between training and generalization error decays by some constant over $\sqrt{m}$ as a function of the number of samples.

3 Chernoff Bounds

We will now prove a stronger form of Chernoff bounds, and Hoeffding's Inequality will follow as a corollary of this proof. This better bound says:

$$Pr[\hat{p} \geq p + \epsilon] \leq e^{-RE(p+\epsilon||p)m} \leq e^{-2\epsilon^2 m} \tag{4}$$

$$Pr[\hat{p} \leq p - \epsilon] \leq e^{-RE(p-\epsilon||p)m} \leq e^{-2\epsilon^2 m} \tag{5}$$

3.1 Relative Entropy

In order to understand Chernoff bounds we have to take a detour and first define the concept of relative entropy, also known as the Kullback-Leibler divergence.

The notion of KL divergence comes from information theory and can be illustrated using a simple example from class. Imagine that person A is trying to send a message to person B. Can we calculate the number of bits necessary to send this message? The obvious initial approach is to say that there are 26 letters in the alphabet and since k bits can describe 2^k different messages, we simply use $\log_2(26)$ bits. Rounding this up to 5, we can encode the alphabet with 5 bits where each letter takes exactly 5 bits (A = 0 0 0 0 0, B = 0 0 0 0 1, C = 0 0 0 1 0, ...).

However, not all letters are equally likely and we can come up with a more efficient encoding system where more likely letters take fewer bits and unlikely letters take more (A = 0 0, Q = 0 1 1 0 1 0 1, ...). This system would use less bits on average to send a message.

If letters are drawn from a probability distribution P, and $P(x)$ is the probability of sending a letter x, the optimal way of coding messages from P is to use $\log_2(\frac{1}{P(x)})$ bits for x. Then:

$$E[message\ length] = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

This value is called the entropy of the distribution P and it is possible to show that this is maximized when P is a uniform distribution.

Now suppose that person A makes a mistake and uses the distribution Q to send the message where $\log_2(\frac{1}{Q(x)})$ bits are used. Now

$$\begin{aligned} E[message\ length] &= \sum_x P(x) \log_2 \frac{1}{Q(x)} \\ expected - optimal &= \sum_x P(x) \log_2 \frac{1}{Q(x)} - \sum_x P(x) \log_2 \frac{1}{P(x)} \\ &= \sum_x P(x) \log_2 (\frac{P(x)}{Q(x)}) \\ &= RE(P||Q) \end{aligned}$$

Note: we will be using natural log base instead of base 2.
RE between two numbers will be written with the shorthand:

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

3.2 Markov's Inequality

Markov's inequality is a simple inequality used to compare the expected value of a random variable to the probability of that variable being very large in relation to the expected value.

Say $X \geq 0$, Markov's Inequality states:

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

Proof:

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

3.3 Proof

We are now ready to prove equation 4. As a first attempt, we let $q = p + \epsilon$, where p and ϵ, and therefore q are fixed. We are trying to upper bound $Pr[\hat{p} \geq q]$. Using Markov's inequality we show:

$$Pr[\hat{p} \geq q] \leq \frac{E[\hat{p}]}{q} = \frac{p}{q} = \frac{p}{p+\epsilon} \tag{6}$$

This is equal to less than one but is very weak and doesn't have the dependence on m we are looking for. Next, we try a clever trick to make this happen and pass both sides of our original inequality through a monotonically increasing function to get an equivalent inequality to work with. This is possible because if f is strictly increasing then $\hat{p} \geq q \Leftrightarrow f(\hat{p}) \geq f(q)$.

We will use the form $f(x) = e^{\lambda m x}$ where $\lambda > 0$. Our new inequality becomes:

$$\begin{aligned}
Pr[\hat{p} \geq q] &= Pr[e^{\lambda m \hat{p}} \geq e^{\lambda m q}] \\
&\leq \frac{E[e^{\lambda m \hat{p}}]}{e^{\lambda m q}} && \text{Markov's Ineq.} \\
&= e^{-\lambda m q} \cdot E[exp(\lambda \sum_{i=1}^{m} X_i)] \\
&= e^{-\lambda m q} \cdot \prod_{i=1}^{m} E[e^{\lambda X_i}] \\
&\leq e^{-\lambda m q} \cdot \prod_{i=1}^{m} E[(1 - X_i) + e^{\lambda} X_i] && \text{bound } e^{\lambda x} \text{ by a line} \\
&= e^{-\lambda m q} \cdot \prod_{i=1}^{m} [1 - p + e^{\lambda} p] \\
&= e^{-\lambda m q} \cdot [1 - p + e^{\lambda} p]^m \\
&= (e^{-\lambda q}[1 - p + e^{\lambda} p])^m \\
&= e^{\phi(\lambda) m}
\end{aligned}$$

We set $\phi(\lambda) = \ln[e^{-\lambda q} \cdot (1 - p + e^{\lambda} p)]$ to get our final result.

Minimizing over λ we find:

$$\lambda_{min} = \ln\left(\frac{q(1-p)}{(1-q)p}\right)$$
$$\phi(\lambda_{min}) = -RE(q||p)$$

This is true for all λ, true for λ_{min}, and gives the earlier bound.

4 McDiarmid's Inequality

Hoeffding's inequality shows that the average of a set of random variables is connected to the expected value of individual random variables. Hoeffding's Inequality as a corollary of the bound involving relative entropy can be proven using Taylor's theorem.

$$\frac{1}{m}\sum_{i=1}^{m} x_i \rightarrow E[\frac{1}{m}\sum_{i=1}^{m} x_i] AVG(x_1, ..., x_m) \rightarrow E[AVG(x_1, ..., x_m)] \tag{7}$$

This tells us the rate at which the average converges to the expected value. What other functions could the average be replaced with and still give these results? Whenever f has the property that changing one argument does not change f by a lot, that is:

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

Let $X_1, ..., X_m$ be independent, but not necessarily identical. Then:

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

$c_i = \frac{1}{m}$ for AVG, and we get Hoeffding's inequality for this special case.

4.1 Putting it Together

Theorem: Given $m = O(\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{\epsilon^2})$ examples, then with probability $\geq 1 - \delta$,

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

Proof: for any particular $h \in \mathcal{H}$, we have already argued that

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

$$Pr[\exists h \in \mathcal{H} : |err(h) - e\hat{r}r(h)| > \epsilon] \leq 2|\mathcal{H}|e^{-2m\epsilon^2} \leq \delta \qquad \text{for given } m$$

We can take these bounds and rewrite them by solving for ϵ, where:

$$|err(h) - e\hat{r}r(h)| \leq O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m}}\right) \tag{8}$$

This combines the number of training examples, complexity of hypotheses, and how well the hypothesis fits the training samples.