COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: E. Glen Weyl

Lecture #3
February 14, 2006

1 Probably Approximately Correct Learning

One of the most important models of learning in this course is the PAC model. This model seeks to find algorithms which can learn concepts, given a set of labeled examples, with a hypothesis that is likely to be about right. This notion of "likely to be about right" is usually called Probably Approximately Correct (PAC). We can define the concept of PAC learning formally, as we did in the last lecture. Let us repeat that definition here, for memory's sake. We will say that a concept class $\mathcal{C}$ is "PAC learnable" by a hypothesis class $\mathcal{H}$ if and only if there exists an algorithm A which can preform the following task: given any target concept $c \in \mathcal{C}$, target distribution D over the set of possible examples X, and strictly positive pairs of real numbers (δ, ϵ) (note: any (δ, ϵ) pair we will actually consider will have both components bounded above by 1), A takes as input a set (called the training set) m of independently drawn random labeled examples $(x_i, c(x_i))$, where $x_i \in X$ is drawn according to D and m is bounded above by a polynomial in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$, and outputs an hypothesis $h \in \mathcal{H}$ about which we can say, with confidence (probability over all possible choices of the training set) greater than $1 - \delta$ that the error of the hypothesis (that is $Pr_D[h(x) \neq c(x)]$ where $x \in X$ is drawn randomly according to D) is $\leq \epsilon$.

2 A Very Simple Example of PAC Learning

The first question that comes to mind in the context of PAC learning is whether such a thing is even ever possible. That we might be able to say that for any concept and any distribution that we can always find a hypothesis that is likely to be about right if we have enough evidence seems a quite ambitious claim; in a sense, this is a claim to which much of thousands of years of epistemological thought has been devoted to adressing. Despite this apparent difficulty, we will see that under the assumptions we have made (an unchanging distribution from which all examples are drawn and an unchanging concept that labels them) a number of intersting concept classes are provably PAC learnable. The best way to start building an intuition for PAC learning, however, is to consider a somewhat less interesting, but simple and instructive example. Let our sample space X be the real line ($\mathcal{R}$) and let our concept space be the set of all positive "half-lines"; that is, every concept or hypothesis is a real number paired with the indicator function of the relationship $\geq$. For example, the concept might be that all real numbers $\geq \pi$ are labeled 1. Let us consider an algorithm for learning this concept class (which we call, as usual, $\mathcal{C}$) and try to prove that it satisfies the requirements of PAC learning and therefore proves that $\mathcal{C}$ is learnable by $\mathcal{H} = \mathcal{C}$.

Theorem 1 *$\mathcal{C}$ is PAC learnable using $\mathcal{C}$.*

Consider the algorithm that first, after seeing a training set S which contains m labeled examples of the form $(x_i, c(x_i))$ where $x_i \in \mathcal{R}$, selects the greatest example labeled 0, which we call $\underline{x}$ ($\underline{x} \equiv max_{i:c(x_i)=0} x_i$), and the smallest example labeled 1, which we call

$\overline{x}$ ($\overline{x} \equiv min_{i:c(x_i)=1} x_i$). We know that $\underline{x} < \overline{x}$ because, otherwise, there is an example labeled 1 that is smaller than an example labeled 0, contradicting that c is a positive ray. The algorithm outputs an hypothesis h that is the positive ray corresponding to a point arbitarily selected (the analysis below does not depend on how it is selected) from the open interval $(\underline{x}, \overline{x})$.

Now suppose we are given any $\epsilon > 0$. Let us define k_c as the real number marking the lower boundary of the positive ray c. Also, let us define $\overline{k}_c \equiv max\{k : D([k_c, k)) \leq \epsilon\}$. That is, $\overline{k}_c$ is the greatest value of k for which the upper half-open interval $[k_c, k)$ has no more than ϵ probability weight under the sampling distribution function D. If D is smooth (has no discrete "lumps" anywhere), this reduces to choosing $\overline{k}_c$ so that $[k_c, \overline{k}_c]$ has exactly ϵ probability weight. Intutively, $\overline{k}_c$ is exactly (or for discrete distributions, as close as we can get to exactly) ϵ above k_c in "probability distance"; that is to say, under the metric where the distance between two points is measured by the proability under D of an example lying in that interval. Let us define $R_+ \equiv [k_c, \overline{k}_c]$. We also can define R_- and $\underline{k}_c$ in a symmetric fashion.

Let us now define k_h as the real number marking the lower boundary of the positive ray h. If $k_h \leq \overline{k}_c$, then there will be a probability, under D, of no more than ϵ that h misclassifies a *positive* example. This is clear because h misclassifies those, and only those, examples on which it disagrees with c and positive examples must have probability weight of no more than ϵ if $k_h \leq \overline{k}_c$ by construction. We want to show this, that the *error* of h is less than ϵ, with some confidence (probability). We can do this as follows.

Define the event that $k_h > \overline{k}_c$ as b_+. Also define symmetrically (I omit details here) the event b_- for $\underline{k}_c$. b_+ will only occur if there is no training example $x_i \in R_+$. Otherwise, A would have chosen x_i as the least example labeled 1 and thus we would have had $k_h \leq x_i \leq \overline{k}_c$. That is, b_+ will only occur if none of the m independent training examples lie inside R_+. Because the probability of an example lying in R_+ is $\geq \epsilon$ by construction, the probability of a training example not lying in R_+ is $\leq 1 - \epsilon$. Because all of the training examples are independent, the probability that *none* of the m training examples are in R_+ is $\leq (1-\epsilon)^m$. Using the bound that $(1-x) \leq e^{-x}$, we can say that the probability of b_+ is $\leq e^{-m\epsilon}$. Using a symmetric argument, we can prove that the probability of b_- is also $\leq e^{-m\epsilon}$. Note that we either have $k_h \leq k_c$ or $k_h \geq k_c$, so h may misclassify positive examples or negative examples but not both. Thus, if neither b_+ nor b_- occur, then the probability of h misclassifying an example is $\leq \epsilon$. Because b_- and b_+ cannot both occur (they are *disjoint* events) the probability of either occuring is just the sum of the probabilities of each of the two occuring. That is, putting this all together, given m independent training examples we can say with probability at least $1 - 2e^{-m\epsilon}$ we can assert that the error of h is less than ϵ. Conversely, if we have at least $\frac{1}{\epsilon}\ln(\frac{2}{\delta})$ training examples, we can show by simple substitution that we can assert with probability at least $1-\delta$ that the error of h is less than ϵ. Because $\frac{1}{\epsilon}\ln(\frac{2}{\delta})$ is clearly bounded above by a polynomial in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$ we have shown that A satisfies the requirements for PAC learning $\mathcal{C}$ by $\mathcal{C}$. This proves that $\mathcal{C}$ is PAC learnable by $\mathcal{C}$ and completes the proof.

Thus, we have shown, for the first time, that an (admitedly simple) concept class is PAC learnable. We can interpret the bound we derived above in one further way by writing that given m training examples we can assert that the h hypothesis outputted by A has error $\leq \frac{1}{m}\ln(\frac{2}{\delta})$. This bound has much the same flavor as a confidence interval in statistics. If we want to acheive 95% ($\delta = .05$) confidence in our bound, we can assert that, with this confidence, the error of h is $\leq \frac{1}{m}\ln(40)$. However, if we want to acheive 99% confidence, then we must "widen" our confidence interval: with 99% confidence we are only able to

assert that the error is $\leq \frac{1}{m}\ln(200)$. If we are willing to only have 90% confidence, we can "tighten" our confidence interval: with 90% confidence we can assert that error is $\leq \frac{1}{m}\ln(20)$.

It is also worth noting that the error dies off linearly with the inverse of m. This is quite rapid and as the course progresses we will be interested in studying the relationship between the number of examples we are given and the error of our hypothesis; this relationship is often called the "learning curve".

3 Two More Simple Examples

In the spirit of the analysis above, we can examine two more simple examples.

3.1 Intervals

Let us once again consider $X = \mathcal{R}$. Now we consider a concept class $\mathcal{C}$ which is only slightly more sophisticated. We let $\mathcal{C}$ be the set of all closed intervals on $\mathcal{R}$. For example, we might have $c \in \mathcal{C}$ where $c = [0, 1]$ in which case $c(x) = 1$ iff $x \in [0, 1]$ and $c(x) = 0$ iff $x \notin [0, 1]$. To analyze this situation, we have to consider four "bad events" like the "bad events" b_+ and b_- we considered above. Loosely, these involve the bottom boundary of our interval being too low, the bottom boundary being too high, the top boundary being too low, and the top boundary being too high. Because the analysis here is a direct analog of the analysis in the example above, we leave the more rigorous proof as an excercise, but note that the final bound we arive at is that if we have at least $\frac{4}{\epsilon}\ln(\frac{4}{\delta})$ training examples we can assert with probability $1 - \delta$ that the error of the (correctly chosen consistent) hypothesis is less than ϵ.

3.2 Axis-Parallel Rectangles

Now we return to an example from last time. We let $X = \mathcal{R}^2$ and let $\mathcal{C}$ be the set of all axis-parallel rectangles in the real plane (as defined in last class). To put this precisely, we can view $\mathcal{C}$ as the set of all cross-producted *pairs* of closed intervals on the real line. An example is then labeled 1 if and only if its x coordinate lies inside the first closed interval and its y coordinate lies within the second interval and is labeled 0 otherwise. For an extensive analysis of this problem in the context of PAC learning, see Kearns and Varziani, Section 1.1. The intuition, however, proceeds along a similar line as the above examples.

4 A More General Theorem

We have shown, or hinted, that in a number of simple cases PAC learning is possible. However, these examples have been quite contrived and not of much general interest. What we would really like to be able to do is say something more definite and general. This is the purpose of the first truly important theorem we will prove in this course. We don't have time to prove it today, so we will save that for next time, but let us quickly state the theorem formally and in plain English.

The theorem states that if we can find a hypothesis $h \in \mathcal{H}$ that is consistent with a test sample of size m, then if $|\mathcal{H}|$ (the *cardinality* of the hypothesis space $\mathcal{H}$) is finite, we can do "the PAC jig"; that is, we can assert with a certain confidence (probability) that the error of h is less than some level ϵ. Futhermore, and more importantly, given any strictly postive

pair (ϵ, δ) we can, supposing we find a hypothesis $h \in \mathcal{H}$ that is consistent with a number of training examples m that is bounded above by a polynomial in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$, assert with probability (confidence) $1 - \delta$ that the error of the hypothesis is less than ϵ. Furthermore, we can come up with an exact expression for the number of examples required. Formally:

Theorem 2 *If we are able to find a hypothesis $h \in \mathcal{H}$ where $\mathcal{H}$ has finite cardinality $|\mathcal{H}|$ which is consistent with m independent random labeled training examples, then for any strictly positive pair (δ, ϵ) we can assert with confidence (probability) $1 - \delta$ that the error of h is less than ϵ provided that:*

$$m \geq \frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{\epsilon}$$

Next class we will prove this theorem. For now, let's try a little fun excercise.

5 (Not Quite) Just for Fun

In the statement of the theorem above, the size (cardinality) of the hypothesis space plays an important role. The larger the size of the hypothesis space, the more examples we need for a given (δ, ϵ) pair to assert with confidence $1 - \delta$ that the hypothesis we get has error less than ϵ. Thus, a large hypothesis space "hurts" us in a certain sense.

To see why this might be a case, let's try a little fun experiment. I have in my mind a "concept" for labeling every number between 1 and 20 as either + or −. For example, I might be thinking that every prime is labeled + or that every number all of whose digits appear in the first 10 places of the decimal expansion of π is labeled + (and all others in both cases are labeled −). A hypothesis about my concept is a rule (like mine) for labeling the numbers between 1 and 20. In class, every person generated one hypothesis; if you want to try to replicate this excercise at home, you have to come up with a dozen or more hypotheses on your own! Come up with your hypotheses before you look at the training set in Figure 1.

Now peek at the training set (but not the test set!) and score your various hypotheses accoring to their error on this training data. Choose the hypothesis with the least error on the training set. Imagine that the set of all of your hypotheses was your hypothesis class and that your "learning algorithm" has just output this least-error hypothesis which we will call h. How low was the error of h? In class with about 30 students, the h we selected had only 10% error. Now let's test h to see how it preforms on test data that is labeled according to the same rule. The test data is the numbers 11 through 20. Figure 2 shows how the concept I had in mind labeled the test set.

How did h do this time? Not quite as well? This isn't a big surprise. My "concept" was just randomly flipping a coin. Yet how did h manage to get such low error on the training set? The basic point is that there were *so many* hypotheses that one was bound to do very well; however, it did not *generalize* well to the test set because, of course, the test set is totally random, so it couldn't hope to generalize! This shows us that we should expect preformance to degrade as hypothesis classes grow larger; good preformance of a hypothesis drawn from a large class on a training set may not tell us very much at all about how well it will generalize, unless we compensate for this larger class with correspondingly more data. This idea, which is the basic justification for the "Occam's Razor" principle which prefers simple (small) hypothesis classes, is a fundamental concept in machine learning.

example	label
train	
1	+
2	−
3	−
4	+
5	−
6	−
7	+
8	+
9	−
10	+

Figure 1: Training Set

example	label
test	
11	−
12	+
13	−
14	+
15	+
16	−
17	−
18	−
19	+
20	−

Figure 2: Test Set

A Appendix: Probably Approximately Correct Learning and Expectation Learning

One question that comes to mind is whether PAC learning is equivelent to what we might call "Expectation learning". Expectation learning demands that we be able to acheive an arbitrarily low *expected* error of the hypothesis , as opposed to insisting (as PAC does) that with arbitrarily high confidence we can get the error arbitarily low. Formally, let us define "Expectation learning" (E) and then prove that it is equivelent to PAC learning. We will say that a concept class $\mathcal{C}$ is "E learnable" by a hypothesis class $\mathcal{H}$ if and only if there exists an algorithm A which can preform the following task: given any target concept $c \in \mathcal{C}$, target distribution D over the set of all possible examples X, and strictly positive real number γ (note: any γ we will consider will be less than 1), A takes as input a set (called the training set) m of indepdently drawn random labeled examples $(x_i, c(x_i))$, where $x_i \in X$ is drawn according to D and m is bounded above by a polynomial in $\frac{1}{\gamma}$, and outputs a hypothesis $h \in \mathcal{H}$ whose *expected* error (defined above) is at worst γ (where the expectation is taken over all possible choices of the training set).

Theorem 3 *E learning is equivelent to PAC learning. That is, a concept class $\mathcal{C}$ is E learnable by $\mathcal{H}$ if and only if it is PAC learnable by $\mathcal{H}$.*

Proof: Our strategy is to show that any algorithm A which satisfies the requirements of PAC learning also satsifies the requirements of E learning (that E learning *reduces* to PAC learning) and, conversely, that any algorith A' which satisfies the requirements of E learning also satsifies those of PAC learning (that PAC learning *reduces* to E learning). First we prove that E learning reduces to PAC learning.

Suppose we are given an algorithm A which satisfies the conditions of PAC learning (given above). We must show that (for all concpets, distributions, and strictly positive values of γ) A can take a training sample of size m that is bounded above by a polynomial in $\frac{1}{\gamma}$ and output a hypothesis $h \in \mathcal{H}$ whose *expected* error is less than γ. A can do this (given any concept, distribution and strictly postive value of γ) in the following manner. It generates a hypothesis $h \in \mathcal{H}$ with the property that $\Pr\left[\text{err}(h) > \frac{\gamma}{2}\right] \leq \frac{\gamma}{2}$, which it can do

because we have assumed it satsifies the conditions of PAC learning. We then have:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{err}(h)\right] &= \mathrm{E}\left[\mathrm{err}(h)|\mathrm{err}(h) \leq \frac{\gamma}{2}\right] \Pr\left[\mathrm{err}(h) \leq \frac{\gamma}{2}\right] + \\
&\quad \mathrm{E}\left[\mathrm{err}(h)|\mathrm{err}(h) > \frac{\gamma}{2}\right] \Pr\left[\mathrm{err}(h) > \frac{\gamma}{2}\right] \\
&\leq \frac{\gamma}{2}\Pr\left[\mathrm{err}(h) \leq \frac{\gamma}{2}\right] + \mathrm{E}\left[\mathrm{err}(h)|\mathrm{err}(h) > \frac{\gamma}{2}\right] \frac{\gamma}{2} \\
&\leq \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma
\end{aligned}$$

The first equality follows by conditioning on whether the error is $>$ or $\leq \frac{\gamma}{2}$. The first inequality follows from the fact that $\mathrm{E}\left[\mathrm{err}(h)|\mathrm{err}(h) \leq \frac{\gamma}{2}\right] \leq \frac{\gamma}{2}$ and our construction from the PAC properties. The second inequality follows from the fact that all probabilities are bounded above by 1.

Thus, h satisfies $\mathrm{E}\left[\mathrm{err}(h)\right] \leq \gamma$. Additionally, the number of examples needed to generate h, m, is bounded above by a polynomial in $\frac{2}{\gamma}$ and $\frac{2}{\gamma}$ because A satisfies the requirements of PAC learning and if m is bounded above by a polynomial in $\frac{2}{\gamma}$ and $\frac{2}{\gamma}$ then it is clearly also bounded above by a polynomial in $\frac{1}{\gamma}$. This shows that A satisfies the requirements of E learning and therefore proves that E learning reduces to PAC learning. Next we prove that PAC learning reduces to E learning.

Suppose we are given an algorithm A which satisfies the conditions of E learning (given above). We must show that (for all concpets, distributions, and strictly positive values pairs (δ, ϵ)), A can take a training sample of size that is bounded above by a polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ and output a hypothesis $h \in \mathcal{H}$ about which we can say, with confidence greater than $1 - \delta$ that the error of the hypothesis is $\leq \epsilon$. A can do this (given any concept, distribution and value of γ) in the following manner. It generates a hypothesis $h \in \mathcal{H}$ with the property that $\mathrm{E}\left[\mathrm{err}(h)\right] \leq \delta\epsilon$, which it can do because we have assumed it satsifies the conditions of E learning. We then suppose that $\Pr\left[\mathrm{err}(h) > \epsilon\right] > \delta$. Then we have

$$\begin{aligned}
\mathrm{E}\left[\mathrm{err}(h)\right] &= \mathrm{E}\left[\mathrm{err}(h)|\mathrm{err}(h) > \epsilon\right] \Pr\left[\mathrm{err}(h) > \epsilon\right] + \\
&\quad \mathrm{E}\left[\mathrm{err}(h)|\mathrm{err}(h) \leq \epsilon\right] \Pr\left[\mathrm{err}(h) \leq \epsilon\right] \\
&> \epsilon\delta
\end{aligned}$$

The equality is again by conditioning and the inequality follows by our supposition and that all probabilities must be non-negative. This gives us $\mathrm{E}\left[\mathrm{err}(h)\right] > \delta\epsilon$ which contradicts that $\mathrm{E}\left[\mathrm{err}(h)\right] < \delta\epsilon$. Thus we must have $\Pr\left[\mathrm{err}(h) > \epsilon\right] < \delta$. Furthermore, the number of examples needed to generate h, m, is bounded above by a polynomial in $\frac{1}{\delta\epsilon}$, which is a polynomial in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$, and because a polynomial of a polynomial is itself a polynomial we have that m is bounded above by a polynomial in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$. This shows that A satisfies the requirement of PAC learning and therefore that PAC learning reduces to E learning. This completes the proof of the equivelence of E and PAC learning.