15-859(B) Machine Learning Theory

Lecture 02/13/06, Avrim Blum

- PAC model & Occam recap
- Chernoff and Hoeffding bounds, uniform convergence
- MB $\Rightarrow$ PAC
- MB $\Rightarrow$ PAC II
- greedy set cover

PAC model recap

- Examples drawn from unknown probability distribution D over instance space X.
- Labeled by unknown target function

$$c : X \rightarrow \{0,1\}$$

- For hypothesis h,

$$err(h) = \Pr_{x \leftarrow D}[h(x) \neq c(x)]$$

- Algorithm PAC-learns C by H if for any $c \in C$, any distrib D, any given $\varepsilon > 0$, $\delta > 0$, with probability $\geq 1 - \delta$ the algorithm produces $h \in H$ with $err(h) < \varepsilon$.
- Want algorithm to be efficient in running time and number of examples too.

Basic sample-complexity bound

- After

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$

  examples, with probability $\geq 1 - \delta$, all $h \in H$ with $err(h) \geq \varepsilon$ have $\widehat{err}(h) > 0$. [$\widehat{err}(h)$ = empirical error on sample]
- Argument: fix bad h. Prob of consistency $\leq (1-\varepsilon)^m \leq \delta/|H|$. Now use union bound.
- "If not too many rules to choose from, then unlikely some bad one will fool you just by chance."
- So, if the target concept is in H, and we have an algorithm for the consistency problem, then we only need this many examples to achieve the PAC guarantee.

Gives an answer to the question: when does the data justify a hypothesis?

Occam's razor

A nice way of looking at this bound, in terms of number of bits needed to describe the hypotheses produced.

- Say we have some description language.
- Say "simple" = "short description".
- At most 2^s hypotheses are $< s$ bits long.
- If number of examples seen satisfies

$$m \geq \frac{1}{\varepsilon}\left[s \ln 2 + \ln\left(\frac{1}{\delta}\right)\right].$$

  then it's unlikely a bad simple hypothesis will fool you just by chance.

This holds no matter what your description language is.

Of course, there's no guarantee that there *will* be a simple explanation consistent with data. That depends on your representation.

Tail inequalities

Tail inequality: bound on probability mass in tail of distribution.

- Consider a hypothesis with true error p and let $q = 1 - p$.
- If we see m examples, then the expected fraction of mistakes is p. The *standard deviation* σ of this quantity is $\sqrt{pq/m}$.
- A convenient rule for iid Bernoulli trials, in our terminology, is:

$$\Pr[|\text{observed error} - \text{true error}| > 1.96\sigma] < 0.05.$$

- E.g., if we want with 95% confidence for our true and observed errors to differ by only ε, then we need to see only $(1.96)^2 pq/\varepsilon^2 < 1/\varepsilon^2$ examples. [worst case is when $p = 1/2$]

Chernoff and Hoeffding bounds extend to case where we want to show something is really unlikely, so can rule out lots of hypotheses.

Typical use of these bounds

Theorem 1 *After m examples, with probability $\geq 1 - \delta$, all $h \in H$ have $|err(h) - \widehat{err}(h)| < \varepsilon$, for*

$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right].$$

Proof: Just apply Hoeffding.

- Chance of failure at most $2|H|e^{-2m\varepsilon^2}$.
- Set to δ.
- Solve.

So, with prob $1 - \delta$, best on sample is ε-best over D.

Note: this is worse than previous bound ($\frac{1}{\varepsilon}$ has become $\frac{1}{\varepsilon^2}$), because we are asking for something stronger. Can also get bounds "between" these two.

Uniform Convergence

Our basic result only bounds the chance that a bad hypothesis looks perfect on the data.

What if there is no perfect $h \in H$?

- Another kind of bound is to show that after m examples, with probability $\geq 1 - \delta$, all $h \in H$ have $|err(h) - \widehat{err}(h)| < \varepsilon$.
- Called "uniform convergence".
- Gives justification for optimizing on the training data more generally.

To prove bounds like this, we need some good tail inequalities: Chernoff and Hoeffding bounds.

Chernoff and Hoeffding bounds

Consider coin of bias p flipped m times. Let S be the observed # heads. Let $\varepsilon \in [0, 1]$.

Hoeffding bounds:

- $\Pr[\frac{S}{m} > p + \varepsilon] \leq e^{-2m\varepsilon^2}$, and
- $\Pr[\frac{S}{m} < p - \varepsilon] \leq e^{-2m\varepsilon^2}$.

Chernoff bounds:

- $\Pr[\frac{S}{m} > p(1 + \varepsilon)] \leq e^{-mp\varepsilon^2/3}$, and
- $\Pr[\frac{S}{m} < p(1 - \varepsilon)] \leq e^{-mp\varepsilon^2/2}$.

E.g., $\Pr[S < (expectation)/2] \leq e^{-(expectation)/8}$.

E.g., $\Pr[S > 2(expectation)] \leq e^{-(expectation)/3}$.

Relating PAC and MB models

- The PAC model should be easier than the MB model since we are restricting examples to be coming from a distribution.
- Can make this formal: show how to convert any MB alg to a PAC alg.
- Will give two conversion methods.
  - First is simpler. Gives sample-size bound of $O\left(\frac{M}{\epsilon}\log\left(\frac{M}{\delta}\right)\right)$.
  - Second is more complicated (and uses Chernoff). Gives better bound of $O\left(\frac{1}{\epsilon}[M + \log(1/\delta)]\right)$.

MB $\Rightarrow$ PAC (better bound)

Theorem 4 *We can actually get a better bound of* $O\left(\frac{1}{\epsilon}[M + \log(1/\delta)]\right)$.

To do this, we will split data into a "training set" S_1 of size max $\left[\frac{4M}{\epsilon}, \frac{16}{\epsilon}\ln\frac{1}{\delta}\right]$ and a "test set" S_2 of size $\frac{32}{\epsilon}\ln\frac{M}{\delta}$. We will run alg on S_1 and test all hyps produced on S_2.

Claim 1: w.h.p., at least one hyp produced on S_1 has error $< \epsilon/2$.
Proof: (tricky!!)

- If all are $\geq \epsilon/2$ then expected number of mistakes is $\geq 2M$.
- By Chernoff, $\Pr[\leq M] \leq e^{(-expect)/8} \leq \delta$.
- View as game: after M mistakes, alg forced to reveal target. If alg keeps giving bad hyps, then whp will be forced to do it.

Claim 2: W.h.p., best one on S_2 has error $< \epsilon$.

Proof. Suffices to show that good one is likely to look better than $3\epsilon/4$ and all with true error $> \epsilon$ are likely to look worse than $3\epsilon/4$. Just apply Chernoff again to the set of M hypotheses....

Typical use of these bounds (II)

Theorem 2 *After m examples, with probability $\geq 1 - \delta$, all $h \in H$ of $err(h) > 2\varepsilon$ have $\widehat{err}(h) > \varepsilon$, and all $h \in H$ of $err(h) < \varepsilon/2$ have $\widehat{err}(h) < \varepsilon$, for*

$$m \geq \frac{6}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right].$$

So this is useful if belief is that optimal function in H is good but not perfect. (If optimal has true error $< \varepsilon/2$ then whp the best on the sample has true error $< 2\varepsilon$.)

MB $\Rightarrow$ PAC (simpler version)

Theorem 3 *If we can learn C with mistake-bound M, then we can learn in the PAC model using a training set of size* $O\left(\frac{M}{\epsilon}\log\left(\frac{M}{\delta}\right)\right)$.

Proof:

- Assume MB alg is "conservative".
- Look at sequence of hypotheses produced: $h_1, h_2, \ldots$.
- For each one, if consistent with the next $\frac{1}{\epsilon}\log\frac{M}{\delta}$ examples, then stop.
- If h_i has error $> \epsilon$, the chance we stopped was at most δ/M. So there's at most a δ chance we are fooled by any of the hypotheses.

Learning an OR function revisited

Alternative greedy-set-cover approach to learning OR function:

- Pick literal that captures the most positive examples, without capturing any negatives.
- Cross of examples covered and repeat.

If there exists an OR function of size r, then:

- If continue until totally consistent, this will find one of size $O(r \log m)$, where $m =$ size of training set.
- If continue until training error $\leq \epsilon/2$ then find one of size $O(r \log \frac{1}{\epsilon})$.

Using our Occam bound, sample-size is $O\left(\frac{1}{\epsilon}\left[\left(r \log \frac{1}{\epsilon}\right) \log(n) + \ln \frac{1}{\delta}\right]\right)$.

This is slightly worse than Winnow (by $\log \frac{1}{\epsilon}$).