TTIC 31250
An Introduction to the Theory of Machine Learning

Uniform convergence, tail inequalities, VC-dimension I

Avrim Blum

04/20/20

1

Today: back to distributional setting

- We are given sample $S = \{(x_i, y_i)\}$.
  - Assume x's come from some fixed probability distribution D over instance space.
  - View labels y as being produced by some target function. [Or can think of distrib over pairs (x_i, y_i).]
- Alg does optimization over S to produce some hypothesis h. Want h to do well on new examples also from D.
- How big does S have to be to get this kind of guarantee?

2

Basic sample complexity bound recap

- If $|S| \geq \frac{1}{\epsilon}\left[\ln|H| + \ln\frac{1}{\delta}\right]$, then with probability $\geq 1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

- Argument: fix bad h. Prob of consistency at most $(1-\varepsilon)^{|S|}$. Set to $\delta/|H|$ and use union bound.

- So, if the target concept is in H, and we have an algorithm that can find consistent functions, then we only need this many examples to achieve the PAC guarantee.

3

Today: two issues

- If $|S| \geq \frac{1}{\epsilon}\left[\ln|H| + \ln\frac{1}{\delta}\right]$, then with probability $\geq 1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

1. Look at more general notions of "uniform convergence".
2. Replace ln(|H|) with better measures of complexity.

4

Uniform Convergence

- Our basic result only bounds the chance that a bad hypothesis looks perfect on the data. What if there is no perfect $h \in H$?
- Without making any assumptions about the target function, can we say that whp all $h \in H$ satisfy $|err_D(h) - err_S(h)| \leq \varepsilon$?
  - Called "uniform convergence".
  - Motivates optimizing over S, even if we can't find a perfect function.
- To prove bounds like this, need some good tail inequalities.

5

Tail inequalities

Tail inequality: bound probability mass in tail of distribution.

- Consider a hypothesis h with true error p.
- If we see m examples, then the expected fraction of mistakes is p, and the standard deviation σ is $(p(1-p)/m)^{1/2}$.
- A convenient rule for iid Bernoulli trials, in our notation, is: $\Pr[|err_D(h) - err_S(h)| > 1.96\sigma] < 0.05$.
  - If we want 95% confidence that true and observed errors differ by only ε, only need $(1.96)^2 p(1-p)/\varepsilon^2 < 1/\varepsilon^2$ 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.

6

Chernoff and Hoeffding bounds

Consider m flips of a coin of bias p. Let N_{heads} be the observed # heads. Let $\varepsilon, \alpha \in [0,1]$.

Hoeffding bounds:

- $\Pr[N_{heads}/m > p + \varepsilon] \le e^{-2m\varepsilon^2}$, and
- $\Pr[N_{heads}/m < p - \varepsilon] \le e^{-2m\varepsilon^2}$.

Chernoff bounds:

- $\Pr[N_{heads}/m > p(1+\alpha)] \le e^{-mp\alpha^2/3}$, and
- $\Pr[N_{heads}/m < p(1-\alpha)] \le e^{-mp\alpha^2/2}$.

E.g,

- $\Pr[N_{heads} > 2(\text{expectation})] \le e^{-(\text{expectation})/3}$.
- $\Pr[N_{heads} < (\text{expectation})/2] \le e^{-(\text{expectation})/8}$.

7

Typical use of bounds

Thm: If $|S| \geq \frac{1}{2\epsilon^2}\left[\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)\right]$, then with prob $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

- Proof: Just apply Hoeffding + union bound.
  - Chance of failure at most $2|H|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.

Hoeffding bounds:

- $\Pr[N_{heads}/m > p + \varepsilon] \le e^{-2m\varepsilon^2}$
- $\Pr[N_{heads}/m < p - \varepsilon] \le e^{-2m\varepsilon^2}$

8

Typical use of bounds

Thm: If $|S| \geq \frac{1}{2\epsilon^2}\left[\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)\right]$, then with prob $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

- Proof: Just apply Hoeffding + union bound.
  - Chance of failure at most $2|H|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.
- So, whp, best on sample is ε-best over D.
  - Note: this is worse than previous bound ($1/\varepsilon$ has become $1/\varepsilon^2$), because we are asking for something stronger.
  - Can also get bounds "between" these two.

9

Typical use of bounds

Thm: If $|S| \geq \frac{6}{\epsilon}\left[\ln|H| + \ln\frac{1}{\delta}\right]$, then with prob $\geq 1-\delta$, all $h \in H$ with $err_D(h) > 2\varepsilon$ have $err_S(h) > \varepsilon$, and all $h \in H$ with $err_D(h) < \varepsilon/2$ have $err_S(h) < \varepsilon$.

Proof: apply Chernoff…

Chernoff bounds:

- $\Pr[N_{heads}/m > p(1+\alpha)] \le e^{-mp\alpha^2/3}$
- $\Pr[N_{heads}/m < p(1-\alpha)] \le e^{-mp\alpha^2/2}$

E.g,

- $\Pr[N_{heads} > 2(\text{expectation})] \le e^{-(\text{expectation})/3}$.
- $\Pr[N_{heads} < (\text{expectation})/2] \le e^{-(\text{expectation})/8}$.

10

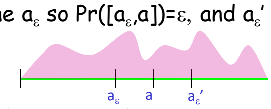
Next topic: improving the |H|

- For convenience, let's go back to the question: how big does S have to be so that whp, $err_S(h) = 0 \Rightarrow err_D(h) \le \varepsilon$.

11

VC-dimension and effective size of H

- If many hypotheses in H are very similar, we shouldn't have to pay so much
- E.g., consider the class $H = \{[0,a]: 0 \le a \le 1\}$.
  - Define a_ε so $\Pr([a_\varepsilon, a]) = \varepsilon$, and a_ε' so $\Pr([a, a_\varepsilon']) = \varepsilon$.



  - Enough to get at least one example in each interval. Just need $(1-\varepsilon)^{|S|} \le \delta/2$.
  - $(1/\varepsilon)\ln(2/\delta)$ examples.
- How can we generalize this notion?

12

Effective number of hypotheses

Define: H[S] = set of all different ways to label points in S using concepts in H.

Define H[m] = maximum |H[S]| over datasets S of m points.

What is H[m] for "initial intervals"?

13

Effective number of hypotheses

Define: H[S] = set of all different ways to label points in S using concepts in H.

Define H[m] = maximum |H[S]| over datasets S of m points.

What is H[m] for linear separators in R^2?

14

Effective number of hypotheses

Define: H[S] = set of all different ways to label points in S using concepts in H.

Define H[m] = maximum |H[S]| over datasets S of m points.

Thm: For any class H, distribution D, if

$$|S| = m > \frac{2}{\epsilon}\left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right)\right],$$

then with prob. 1-δ, all $h \in H$ with error > ε are inconsistent with data. [Will prove next class]

I.e., can roughly replace "|H|" with "H[2m]".

15

Effective number of hypotheses

Define: H[S] = set of all different ways to label points in S using concepts in H.

Define H[m] = maximum |H[S]| over datasets S of m points.

- H[m] is sometimes hard to calculate exactly, but can get a good bound using "VC-dimension".
- VC-dimension is roughly the point at which H stops looking like it contains all functions.

16

Shattering

- Defn: A set of points S is shattered by H if there are concepts in H that label S in all of the $2^{|S|}$ possible ways.
  - In other words, all possible ways of classifying points in S are achievable using concepts in H.
- E.g., any 3 non-collinear points in $\mathbb{R}^2$ can be shattered by linear threshold functions, but no set of 4 points can be.

17

VC-dimension

- The VC-dimension of a hypothesis class H is the size of the largest set of points that can be shattered by H.
- So, if the VC-dimension is d, that means there exists a set of d points that can be shattered, but no set of d+1 points can be shattered.
- E.g., VC-dim(linear threshold fns in 2-D) = 3.
  - Will later show VC-dim(LTFs in R^n) = n+1.
  - What is the VC-dim of intervals on the real line? 2
  - How about C = {all 0/1 functions on $\{0,1\}^n$}? 2^n

18

Upper and lower bound theorems

- Theorem 1: For any class H, distribution D, if $m = |S| > \frac{2}{\epsilon}\left[\log_2(2H[2m]) + \log_2 \frac{1}{\delta}\right]$, then with prob. 1-δ, all $h \in H$ with error > ε are inconsistent with data.
- Theorem 2 (Sauer's lemma):

$$H[m] \leq \sum_{i=0}^{VCdim(H)} \binom{m}{i} = O(m^{VCdim(H)}).$$

- Corollary 3: can replace bound in Thm 1 with

$$O\left(\frac{1}{\epsilon}\left[VCdim(H)\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

- Theorem 4: For any alg A, class H, exists distrib D and target in H such that if $|S| < \frac{VCdim(H)-1}{8\epsilon}$ then $E[err_D(A)] \geq \epsilon$.

19

Upper and lower bound theorems

- Theorem 4: For any alg A, class H, exists distrib D, $f \in H$ s.t. if $|S| < \frac{VCdim(H)-1}{8\epsilon}$ then $E[err_D(A)] \geq \epsilon$.
- Proof:
  - Consider d = VC-dim(H) shattered points. Define distrib D with prob $1 - 4\epsilon$ on one point and prob $\frac{4\epsilon}{d-1}$ on the rest.
  - Pick a random labeling of the d points as the target.
  - $E[err_D(A)] = \Pr[mistake\ on\ test\ point] \geq \frac{1}{2}\Pr[test\ point\ not\ in\ S] \geq \frac{1}{2}(4\epsilon)\left(1-\frac{4\epsilon}{d-1}\right)^{|S|} \geq (2\epsilon)\left(1-\frac{|S|4\epsilon}{d-1}\right) = (2\epsilon)\left(1-\frac{1}{2}\right) = \epsilon.$

20

Upper and lower bound theorems

- Theorem 2 (Sauer's lemma): $H[m] \leq \binom{m}{\leq d}$ = ways of choosing d=VCdim(H) or fewer items out of m.
- Proof:
  - First, note that $\binom{m}{\leq d} = \binom{m-1}{\leq d} + \binom{m-1}{\leq d-1}$. See why?

21

Upper and lower bound theorems

- Theorem 2 (Sauer's lemma): $H[m] \leq \binom{m}{\leq d}$ = ways of choosing d=VCdim(H) or fewer items out of m.
- Proof:
  - First, note that $\binom{m}{\leq d} = \binom{m-1}{\leq d} + \binom{m-1}{\leq d-1}$. See why?
  - Say we have a set S of m examples. Look at $H[S]$.
  - Pick an $x \in S$. Call $h, h' \in H[S]$ "twins" if differ only on x.
  - We know $H[S \setminus \{x\}]$ has $\leq \binom{m-1}{\leq d}$ labelings by induction.
  - How much larger is $H[S]$ compared to $H[S \setminus \{x\}]$? Just the number of twins. Let $H' = \{h \in H[S]$ that labels x negative but has a twin that labels x positive$\}$.
  - VCdim$(H') \leq d-1$. (Since VCdim$(H) \geq$ VCdim$(H')+1$.)
  - Proof follows.

22