TTIC 31250
An Introduction to the Theory of Machine Learning

Uniform convergence, tail inequalities, VC-dimension I

Avrim Blum

04/11/18

Today: back to distributional setting

- We are given sample S = $\{(x_i, l_i)\}$.
  - Assume x's come from some fixed probability distribution D over instance space.
  - View labels l as being produced by some target function. [Or can think of distrib over pairs (x_i, l_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?

Basic sample complexity bound recap

- If $|S| \geq (1/\varepsilon)[\ln(|C|) + \ln(1/\delta)]$, then with probability $\geq 1 - \delta$, all $h \in C$ 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 $\delta/|C|$ and use union bound.
- So, if the target concept is in C, and we have an algorithm that can find consistent functions, then we only need this many examples to achieve the PAC guarantee.

Today: two issues

- If $|S| \geq (1/\varepsilon)[\ln(|C|) + \ln(1/\delta)]$, then with probability $\geq 1 - \delta$, all $h \in C$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

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

Uniform Convergence

- Our basic result only bounds the chance that a bad hypothesis looks perfect on the data. What if there is no perfect $h \in C$?
- Without making any assumptions about the target function, can we say that whp all $h \in C$ 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.

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.

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:

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

Chernoff bounds:

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

E.g,

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

Typical use of bounds

Thm: If $|S| \geq (1/(2\varepsilon^2))[\ln(2|C|) + \ln(1/\delta)]$, then with probability $\geq 1-\delta$, all $h \in C$ have $|err_D(h) - err_S(h)| < \varepsilon$.

- Proof: Just apply Hoeffding.
  - Chance of failure at most $2|C|e^{-2|S|\varepsilon^2}$.
  - 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.

Typical use of bounds

Thm: If $|S| \geq (6/\varepsilon)[\ln(|C|) + \ln(1/\delta)]$, then w.p. $\geq 1-\delta$, all $h \in C$ with $err_D(h) > 2\varepsilon$ have $err_S(h) > \varepsilon$, and all $h \in C$ with $err_D(h) < \varepsilon/2$ have $err_S(h) < \varepsilon$.

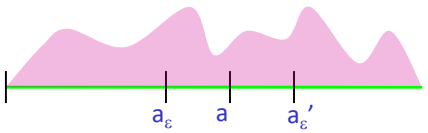
- Proof: apply Chernoff.

Next topic: improving the |C|

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

VC-dimension and effective size of C

- If many hypotheses in C are very similar, we shouldn't have to pay so much
- E.g., consider the class $C = \{[0,a]: 0 \leq a \leq 1\}$.
  - 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 $(1-\varepsilon)^{|S|} \leq \delta/2$.
  - $(1/\varepsilon)\ln(2/\delta)$ examples.
- How can we generalize this notion?

Effective number of hypotheses

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

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

Effective number of hypotheses

Define: $C[m]$ = maximum number of ways to label m points using concepts in C.

- What is $C[m]$ for "initial intervals"?
- How about linear separators in R^2?

- Thm: For any class C, distribution D, if $|S| = m > (2/\varepsilon)[\log_2(2C[2m]) + \log_2(1/\delta)]$, then with prob. $1-\delta$, all $h \in C$ with error $> \varepsilon$ are inconsistent with data. [Will prove soon]
- I.e., can roughly replace "$|C|$" with "$C[2m]$".

Effective number of hypotheses

Define: $C[m]$ = maximum number of ways to label m points using concepts in C.

- What is $C[m]$ for "initial intervals"?
- How about linear separators in R^2?

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

Shattering

- Defn: A set of points S is shattered by C if there are concepts in C 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 C.
- E.g., any 3 non-collinear points can be shattered by linear threshold functions in 2-D.
- But no set of 4 points in R^2 can be shattered by LTFs.

VC-dimension

- The VC-dimension of a concept class C is the size of the largest set of points that can be shattered by C.
- So, if the VC-dimension is d, that means there exists a set of d points that can be shattered, but there is no set of d+1 points that 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?
  - How about C = {all 0/1 functions on $\{0,1\}^n$}?

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if $m=|S| > (2/\varepsilon)[\log_2(2C[2m]) + \log_2(1/\delta)]$, then with prob. $1-\delta$, all $h \in C$ with error $> \varepsilon$ are inconsistent with data.
- Theorem 2 (Sauer's lemma):
$$C[m] \leq \sum_{i=0}^{VCdim(C)} \binom{m}{i} = O(m^{VCdim(C)})$$
- Corollary 3: can replace bound in Thm 1 with
$$O\left(\frac{1}{\epsilon}[VCdim(C)\log(1/\epsilon) + \log(1/\delta)]\right)$$
- Theorem 4: For any alg A, class C, exists distrib D and target in C such that $|S| < (VCdim(C)-1)/(8\varepsilon) \Rightarrow E[err_D(A)] \geq \varepsilon$.

Upper and lower bound theorems

- Theorem 4: For any alg A, class C, exists distrib D, $f \in C$ s.t. $|S|<(VCdim(C)-1)/(8\varepsilon) \Rightarrow E[err_D(A)] \geq \varepsilon$.
- Proof:
  - Consider d = VC-dim(C) shattered points. Define distrib D with $1-4\epsilon$ prob on one point and prob $\frac{4\epsilon}{d-1}$ on the rest.
  - Pick a random labeling of the d points as the target.
  - $E[err_D(A)] \geq \frac{1}{2}\Pr[unseen] \geq \frac{1}{2}(4\epsilon)\left(1-\frac{4\epsilon}{d-1}\right)^{|S|} \geq$
$$(2\epsilon)\left(1-\frac{|S|4\epsilon}{d-1}\right) = (2\epsilon)\left(1-\frac{1}{2}\right) = \epsilon.$$

Upper and lower bound theorems

- Theorem 2 (Sauer's lemma): $C[m] \leq \binom{m}{\leq d}$ = ways of choosing d=VCdim(C) 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 $C[S]$.
  - Pick an $x \in S$. Call $h, h' \in C[S]$ "twins" if differ only on x.
  - We know $C[S \setminus \{x\}]$ has $\leq \binom{m-1}{\leq d}$ labelings by induction.
  - How much larger is $C[S]$ compared to $C[S \setminus \{x\}]$? Just the number of twins. Let $C' = \{h \in C[S]$ that labels x negative but has a twin that labels x positive$\}$.
  - Proof follows from showing that VCdim$(C') \leq d - 1$.