TTIC 31250
An Introduction to the Theory of Machine Learning

VC-dimension II

Avrim Blum

04/16/18

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.

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 datapoints.

- E.g., linear separators in the plane. $C[3]=8$, $C[4]=14$.

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. I.e., largest d s.t. $C[d] = 2^d$.
- 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 $m=|S| > (2/\varepsilon)[\log_2(C[2m]) + \log_2(2/\delta)]$, then with prob. 1-$\delta$, all $h \in C$ with error > ε 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 unif random labeling of the d points as the target.
  - $E[err_D(A)] \geq \frac{1}{2}\Pr[unseen] \geq \frac{1}{2}(4\epsilon)\left(1-\frac{4\epsilon}{d-1}\right)^{|S|} \geq$
$$(2\epsilon)\left(1-\frac{|S|4\epsilon}{d-1}\right) = (2\epsilon)\left(1-\frac{1}{2}\right) = \epsilon.$$

Intuition: Given $\frac{d-1}{8\epsilon}$ training points, expect only $\frac{d-1}{2}$ to be in the "rest". So, expect $\geq 2\epsilon$ probability mass to be unseen. Since labels are random, expect error $\geq \epsilon$.

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$\}$.
  - VCdim$(C') \leq d-1$. (Since VCdim$(C) \geq$ VCdim$(C')+1$.)
  - Proof follows.

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if m=|S| > $(2/\varepsilon)[\log_2(C[2m]) + \log_2(2/\delta)]$, then with prob. 1-$\delta$, all h∈C with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.
- Proof (Step 1):
  - Given a set S of m examples, define A_S = event that exists $h \in C$ with $err_D(h) \geq \epsilon$ but $err_S(h) = 0$. Want to show $\Pr_{S \sim D^m}[A_S] \leq \delta$.
  - Now, consider drawing two sets S, S' of m examples each. Let $B_{S,S'}$ = event that exists $h \in C$ with $err_{S'}(h) \geq \frac{\epsilon}{2}$ but $err_S(h) = 0$. **Claim:** $\Pr_{S \sim D^m}[A_S] \leq 2 \Pr_{S,S' \sim D^m}[B_{S,S'}]$.
  - **Proof:** $\Pr[B] \geq \Pr[A] * \Pr[B|A]$. $\Pr[B|A] \geq \frac{1}{2}$ by Chernoff so long as $m \geq \frac{8}{\epsilon}$. So, $\Pr[A] \leq 2 * \Pr[B]$.

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if m=|S| > (2/ε)[log$_2$(C[2m]) + log$_2$(2/δ)], then with prob. 1-δ, all h∈C with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.
- Proof (Step 1):
  - Given a set S of m examples, define A_S = event that exists $h \in C$ with $err_D(h) \geq \epsilon$ but $err_S(h) = 0$. Want to show $\Pr_{S\sim D^m}[A_S] \leq \delta$.
  - Now, consider drawing two sets S, S' of m examples each. Let $B_{S,S'}$ = event that exists $h \in C$ with $err_{S'}(h) \geq \frac{\epsilon}{2}$ but $err_S(h) = 0$. **Claim:** $\Pr_{S\sim D^m}[A_S] \leq 2 \Pr_{S,S'\sim D^m}[B_{S,S'}]$.
  - So suffices to show $\Pr[B] \leq \delta/2$.

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if m=|S| > (2/ε)[log$_2$(C[2m]) + log$_2$(2/δ)], then with prob. 1-δ, all h∈C with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.
- Proof (Step 2):
  - Now, consider a 3rd experiment. Draw a set S'' of $2m$ examples, then randomly partition into S, S' of m each.
  - Let $B^*_{S'',S,S'}$ = event that exists $h \in C$ with $err_{S'}(h) \geq \frac{\epsilon}{2}$ but $err_S(h) = 0$. **Claim:** $\Pr_{S''\sim D^{2m},S,S'}\left[B^*_{S'',S,S'}\right] = \Pr_{S,S'\sim D^m}[B_{S,S'}]$. (think of examples as sealed envelopes)
  - So, it suffices to show $\Pr_{S''\sim D^{2m},S,S'}\left[B^*_{S'',S,S'}\right] \leq \delta/2$.
  - Will actually prove: for any $|S''| = 2m$, $\Pr_{S,S'}\left[B^*_{S'',S,S'}\right] \leq \delta/2$.

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if m=|S| > (2/ε)[log$_2$(C[2m]) + log$_2$(2/δ)], then with prob. 1-δ, all h∈C with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.
- To show: for any S'' of $2m$ examples, $\Pr_{S,S'}\left[B^*_{S'',S,S'}\right] \leq \delta/2$.
  - Key idea: Now that S'' is fixed, at most $C[2m]$ labelings to worry about. For each one, show that its chance of being perfect on S but error $\geq \epsilon/2$ on S' is low (over the random partition into S, S'). Then apply union bound.
  - So, fix some labeling $h \in C[S'']$. Can assume h makes at least $\epsilon m/2$ mistakes in S'' (else prob of bad event is 0).
  - When we split S'' into S, S', what's the chance all these mistakes go into S'?

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if m=|S| > (2/ε)[log$_2$(C[2m]) + log$_2$(2/δ)], then with prob. 1-δ, all h∈C with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.
- To show: for any S'' of $2m$ examples, $\Pr_{S,S'}\left[B^*_{S'',S,S'}\right] \leq \delta/2$.
  - h makes at least $\epsilon m/2$ mistakes in S''. What's the chance all these mistakes go into S'?
  - Let's partition S'' by first randomly pairing the points together $(a_1, b_1), \ldots, (a_m, b_m)$. Then for each pair i, flip a coin: if heads, $a_i \to S, b_i \to S'$; if tails, $a_i \to S', b_i \to S$.
  - If there is any i s.t. h makes mistakes on both a_i and b_i then the chance is 0; else the chance (over the random coin flips) is at most $2^{-\epsilon m/2}$.
  - Overall failure prob $\leq C[2m]2^{-\epsilon m/2} \leq \frac{\delta}{2}$.

Upper and lower bound theorems

- Theorem 1': For any class C, distribution D, if $m = |S| \geq \frac{8}{\epsilon^2}\left[\ln(C[2m] + \ln\left(\frac{2}{\delta}\right)\right]$, then with prob 1-$\delta$, all $h \in C$ have $|err_D(h) - err_S(h)| \leq \epsilon$.
- Proof: same as for Thm 1 except def of B^*:
  - $B^*_{S",S,S'}$ = event that $\exists h \in C$ with $|err_{S'}(h) - err_S(h)| \geq \frac{\epsilon}{2}$.
  - To show: for any $|S"| = 2m$, $\Pr_{S,S'}\left[B^*_{S",S,S'}\right] \leq \delta/2$.
  - Fix $h \in C[S"]$, pairing $(a_1, b_1), \ldots, (a_m, b_m)$. Say m' indices i s.t. only one of $h(a_i), h(b_i)$ is a mistake.
  - Prob that h is bad over coin-flip experiment is prob that get $|\#heads - \#tails| \geq \epsilon m/2$ in $m' \leq m$ flips.
  - View as ratio being off from expectation by $\geq \left(\frac{\epsilon m}{4m'}\right)$ and apply Hoeffding.