TTIC 31250
An Introduction to the Theory of Machine Learning

VC-dimension II

Avrim Blum

04/18/22

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 $|S| \geq \frac{1}{2\epsilon^2}\left[\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)\right]$, 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:

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

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.

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

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.

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. I.e., largest d s.t. $H[d] = 2^d$.

- 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.

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):**

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

- **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 $E[err_D(A)] \geq \varepsilon$.

Upper and lower bound theorems

- Theorem 1: For any class H, distribution D, if m=|S| > (2/ε)[$\log_2$(H[2m]) + $\log_2$(2/δ)], then with prob. 1-δ, all h∈H 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 $h \in H$ 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 $h \in H$ with $err_{S'}(h) \geq \frac{\epsilon}{2}$ but $err_S(h) = 0$. **Claim:** $\Pr_{S,S' \sim D^m}[B_{S,S'}] \geq \frac{1}{2} \Pr_{S \sim D^m}[A_S]$.
  - **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, $\Pr[B] \geq 1/2 * \Pr[A]$.

Upper and lower bound theorems

- Theorem 1: For any class H, distribution D, if m=|S| > (2/ε)[$\log_2$(H[2m]) + $\log_2$(2/δ)], then with prob. 1-δ, all h∈H 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 $h \in H$ 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 $h \in H$ with $err_{S'}(h) \geq \frac{\epsilon}{2}$ but $err_S(h) = 0$. **Claim:** $\Pr_{S,S' \sim D^m}[B_{S,S'}] \geq \frac{1}{2} \Pr_{S \sim D^m}[A_S]$.
  - So suffices to show $\Pr[B] \leq \delta/2$.

Upper and lower bound theorems

- Theorem 1: For any class H, distribution D, if m=|S| > (2/ε)[log₂(H[2m]) + log₂(2/δ)], then with prob. 1-δ, all h∈H 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 $h \in H$ 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 $|S''| = 2\mathrm{m}$, $\Pr_{S,S'}\left[B^*_{S'',S,S'}\right] \leq \delta/2$.

Upper and lower bound theorems

- Theorem 1: For any class H, distribution D, if m=|S| > (2/ε)[log$_2$(H[2m]) + log$_2$(2/δ)], then with prob. 1-δ, all h∈H 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 $H[2m]$ 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 $h \in H[S"]$. 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 H, distribution D, if m=|S| > (2/ε)[$\log_2$(H[2m]) + $\log_2$(2/δ)], then with prob. 1-δ, all h∈H 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, $a_i \rightarrow S, b_i \rightarrow S'$; if tails, $a_i \rightarrow S', b_i \rightarrow S$.

Upper and lower bound theorems

- Theorem 1: For any class H, distribution D, if m=|S| > (2/ε)[log₂(H[2m]) + log₂(2/δ)], then with prob. 1-δ, all h∈H 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 $\leq H[2m]2^{-\epsilon m/2} \leq \frac{\delta}{2}$.

Upper and lower bound theorems

- Theorem 1′: For any class H, distribution D, if $m = |S| \geq \frac{8}{\epsilon^2}\left[\ln(H[2m] + \ln\left(\frac{2}{\delta}\right)\right]$, then with prob 1-δ, all $h \in \mathrm{H}$ 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 $\exists h \in H$ 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 $h \in H[S"]$, pairing $(a_1, b_1), \ldots, (a_m, b_m)$. Say there are 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.