TTIC 31250
An Introduction to the Theory of Machine Learning

Rademacher Bounds

Avrim Blum
04/27/20

1

Last time we ended with…

2

Uniform Convergence (VC)

- Theorem 1': For any class H, distribution D over $X \times \{-1,1\}$, 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 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.

3

Motivation and Plan

These bounds are nice but have two drawbacks we'd like to address:

1. **Computability/estimability:** say we have a hypothesis class H that we don't understand well. It might be hard to compute or estimate $H[m]$.
2. **Tightness:** Our bounds have two sources of loss. One is we did a union bound over labelings of the double-sample $S"$, which is overly pessimistic if many are very similar to each other. A second is that we did worst-case over $S"$, whereas we would rather do expected case, or even have a bound that depends on our *actual training set*.

We will be able to address both, at least in the uniform convergence case.

4

In particular, we will show:

- Suppose you replaced all true labels of S with random labels and found the $h \in H$ of lowest "empirical error" for these.
- Say E[lowest "empirical error"] $= \frac{1}{2} - \alpha$.
- Clearly, in this experiment, we are overfitting by α since $err_D(h)$ for a random target function is exactly $\frac{1}{2}$.
- Claim: $2\alpha + \text{(low order)}$ is an upper bound on the amount of overfitting we get for the true target function.

Bounding overfitting of target by 2x amount of overfitting of random noise

5

Example where need the factor 2

- Suppose the target is all negative. Hypothesis class H is all Boolean functions over large domain X.
- For random labels, E[lowest "empirical error"] $= \frac{1}{2} - \alpha$, for $\alpha = \frac{1}{2}$ since can fit anything.
- For true target, can overfit even worse using $h =$ "if $x \in S$ predict negative, else predict positive".

6

Some preliminaries

- Rather than writing m as a function of ϵ, write ϵ as function of m. E.g., would write Theorem 1' as:
- For any class H and distribution D, whp all h in H satisfy

$$err_D(h) \le err_S(h) + \sqrt{\frac{8\left(\ln\left(\frac{2H[2m]}{\delta}\right)\right)}{m}}$$

(And we bound in the other direction as well, but let's just focus on this direction - i.e., how much we overfit the sample).

7

Rademacher averages

- For a given set of data $S = \{(x_1, l_1), \ldots, (x_m, l_m)\}$ and class of functions F, the Empirical Rademacher Complexity of F is:

$$R_S(F) = E_\sigma\left[\max_{h\in F}\frac{1}{m}\sum_i \sigma_i h(x_i)\right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ is a random $\{-1,+1\}$ labeling.

- I.e., if you pick a random labeling σ of S, on average how well correlated is the most-correlated $h \in F$?
- Note: $h: X \to \{-1,1\}$ so $\sigma_i h(x_i) = 1$ if agree, -1 if disagree.
- Note "correlation" = agreement – disagreement, so error 45% means correlation of 10%.

8

Rademacher averages

- For a given set of data $S = \{(x_1, l_1), \ldots, (x_m, l_m)\}$ and class of functions F, the Empirical Rademacher Complexity of F is:

$$R_S(F) = E_\sigma\left[\max_{h\in F}\frac{1}{m}\sum_i \sigma_i h(x_i)\right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ is a random $\{-1,+1\}$ labeling.

- Distributional RC of F is: $R_D(F) = E_{S\sim D^m}[R_S(F)]$
- **Theorem:** for any class H, distrib D, if $S \sim D^m$ then with prob $\ge 1-\delta$, all $h \in H$ satisfy:
  - $err_D(h) \le err_S(h) + R_D(H) + \sqrt{\frac{\ln(2/\delta)}{2m}}$

$$\le err_S(h) + R_S(H) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

9

Rademacher vs VC

- Rademacher bound can never be much worse than VC bound.

$$err_D(h) \le err_S(h) + R_S(H) + O\left(\frac{\sqrt{\ln(1/\delta)}}{\sqrt{m}}\right)$$

$$R_S(H) = E_\sigma\left[\max_{h\in H}\frac{1}{m}\sum_i \sigma_i h(x_i)\right]$$

- How big can $R_S(H)$ be?
- Class H produces labelings $h_1, \ldots, h_{|H[S]|}$ of S. For each such labeling h_i, the probability that its correlation with σ is more than 2ϵ is at most $e^{-2m\epsilon^2}$ by Hoeffding bounds.
- Setting this to $\delta/|H[S]|$, whp *all* $h \in H$ have correlation with σ at most $2\sqrt{\frac{\ln(|H[S]|/\delta)}{2m}}$. So, $R_S(H)$ can't be much larger.

10

Rademacher vs VC

- Rademacher bound can never be much worse than VC bound.

$$err_D(h) \le err_S(h) + R_S(H) + O\left(\frac{\sqrt{\ln(2/\delta)}}{\sqrt{m}}\right)$$

$$R_S(H) = E_\sigma\left[\max_{h\in H}\frac{1}{m}\sum_i \sigma_i h(x_i)\right]$$

- How big can $R_S(H)$ be?

Theorem 1' bound: $err_D(h) \le err_S(h) + \sqrt{\frac{8\left(\ln\left(\frac{2H[2m]}{\delta}\right)\right)}{m}}$

- Setting this to $\delta/|H[S]|$, whp *all* $h \in H$ have correlation with σ at most $2\sqrt{\frac{\ln(|H[S]|/\delta)}{2m}}$. So, $R_S(H)$ can't be much larger.

11

On to the proof.

For this, we need to introduce another tail inequality…

12

McDiarmid's inequality

Say $x_1, \ldots, x_m$ are independent RVs, and $\phi(x_1, \ldots, x_m)$ is some real-valued function.

Assume ϕ satisfies the Lipschitz condition that changing x_i can change ϕ by at most c_i.

Then:

$$\Pr_x[\phi(x) > \mathrm{E}[\phi(x)] + \epsilon] \le e^{-2\epsilon^2/(\sum_i c_i^2)}$$

- E.g., if $x_i \in [0,1]$ and $\phi(x) = \frac{x_1+\cdots+x_m}{m}$, then $c_i = \frac{1}{m}$, and we get $e^{-2\epsilon^2 m}$ (we recover Hoeffding).

13

Rademacher proof

- **Theorem:** for any class H, distrib D, if $S \sim D^m$ then with prob $\ge 1-\delta$, all $h \in H$ satisfy:
  - $err_D(h) \le err_S(h) + R_D(H) + \sqrt{\frac{\ln(2/\delta)}{2m}} \le err_S(h) + R_S(H) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$.

  where $R_S(H) = E_\sigma\left[\max_{h\in H} \frac{1}{m}\sum_i \sigma_i h(x_i)\right]$, $R_D(H) = E_S[R_S(H)]$.

- **Step 1:** Define $MAXGAP(S) = \max_{h\in H}[err_D(h) - err_S(h)]$. We want to show that with prob $\ge 1-\delta$, $MAXGAP(S) \le R_D(H) + \sqrt{\frac{\ln(2/\delta)}{2m}}$.

  Claim 1: with prob $\ge 1-\delta/2$, $MAXGAP(S) \le E_S[MAXGAP(S)] + \sqrt{\frac{\ln(2/\delta)}{2m}}$.

  Proof: Think of $MAXGAP(S)$ as ϕ in McDiarmid. Examples are iid RVs. $MAXGAP$ can change by at most $\frac{1}{m}$ if any (x_i, y_i) changes. Claim 1 follows.

  So, suffices to show $E_S[MAXGAP(S)] \le R_D(H)$.

14

Rademacher proof

- **Theorem:** for any class H, distrib D, if $S \sim D^m$ then with prob $\ge 1-\delta$, all $h \in H$ satisfy:
  - $err_D(h) \le err_S(h) + R_D(H) + \sqrt{\frac{\ln(2/\delta)}{2m}} \le err_S(h) + R_S(H) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$.

  where $R_S(H) = E_\sigma\left[\max_{h\in H} \frac{1}{m}\sum_i \sigma_i h(x_i)\right]$, $R_D(H) = E_S[R_S(H)]$.

- **Step 1:** Define $MAXGAP(S) = \max_{h\in H}[err_D(h) - err_S(h)]$. We want to show that with prob $\ge 1-\delta$, $MAXGAP(S) \le R_D(H) + \sqrt{\frac{\ln(2/\delta)}{2m}}$.

  Claim 1: with prob $\ge 1-\delta/2$, $MAXGAP(S) \le E_S[MAXGAP(S)] + \sqrt{\frac{\ln(2/\delta)}{2m}}$.

  Claim 2: with prob $\ge 1-\delta/2$, $R_S(H)$ is within $2\sqrt{\frac{\ln(2/\delta)}{2m}}$ of $R_D(H)$. Proof: apply McDiarmid to $R_S(H)$.

  So, suffices to show $E_S[MAXGAP(S)] \le R_D(H)$.

15

Rademacher proof

$MAXGAP(S) = \max_{h\in H}[err_D(h) - err_S(h)]$

- **Step 2:** show $E_S[MAXGAP(S)] \le R_D(H)$.
- **Proof (uses a ghost sample argument):**
  - Let's rewrite $err_D(h)$ as $E_{S'}[err_{S'}(h)]$ where S' is "ghost sample".

    $$E_S[MAXGAP(S)] = E_S\left[\max_{h\in H}[E_{S'}[err_{S'}(h)] - err_S(h)]\right]$$

    $$\le E_{S,S'}\left[\max_{h\in H}[err_{S'}(h) - err_S(h)]\right]$$

    (you get to pick h after seeing both S and S')

  - Say $S = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, $S' = \{(x_1',y_1'),\ldots,(x_m',y_m')\}$. Can rewrite as:

    $$E_{S,S'}\left[\max_{h\in H}\frac{\left[\sum_i err_{x_i'}(h) - err_{x_i}(h)\right]}{m}\right]$$

    $err_{x_i}(h) = \mathbf{1}_{h(x_i)\ne y_i}$

16

Rademacher proof

$MAXGAP(S) = \max_{h\in H}[err_D(h) - err_S(h)]$

- **Step 2:** show $E_S[MAXGAP(S)] \le R_D(H)$.
- **Proof (uses a ghost sample argument):**
  - Now, like in the VCdim proof, let's flip a coin σ_i for each i to decide whether or not to swap (x_i,y_i) and (x_i',y_i') before taking the max.

    $$E_{S,S',\sigma}\left[\max_{h\in H}\frac{\left[\sum_i \sigma_i(err_{x_i'}(h) - err_{x_i}(h))\right]}{m}\right]$$

  - Say $S = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, $S' = \{(x_1',y_1'),\ldots,(x_m',y_m')\}$. Can rewrite as:

    $$E_{S,S'}\left[\max_{h\in H}\frac{\left[\sum_i err_{x_i'}(h) - err_{x_i}(h)\right]}{m}\right]$$

    $err_{x_i}(h) = \mathbf{1}_{h(x_i)\ne y_i}$

17

Rademacher proof

$MAXGAP(S) = \max_{h\in H}[err_D(h) - err_S(h)]$

- **Step 2:** show $E_S[MAXGAP(S)] \le R_D(H)$.
- **Proof (uses a ghost sample argument):**
  - Now, like in the VCdim proof, let's flip a coin σ_i for each i to decide whether or not to swap (x_i,y_i) and (x_i',y_i') before taking the max.

    $$E_{S,S',\sigma}\left[\max_{h\in H}\frac{\left[\sum_i \sigma_i(err_{x_i'}(h) - err_{x_i}(h))\right]}{m}\right]$$

    $$\le E_{S',\sigma}\left[\max_{h\in H}\frac{\left[\sum_i \sigma_i\, err_{x_i'}(h)\right]}{m}\right] - E_{S,\sigma}\left[\min_{h\in H}\frac{\left[\sum_i \sigma_i\, err_{x_i}(h)\right]}{m}\right]$$

    (gap is only larger if we allow the h's to differ)

18

Rademacher proof

- **Step 2:** show $E_S[MAXGAP(S)] \leq R_D(H)$.

$MAXGAP(S) = \max_{h \in H}[err_D(h) - err_S(h)]$

- **Proof (uses a ghost sample argument):**
  - Now, like in the VCdim proof, let's flip a coin σ_i for each i to decide whether or not to swap (x_i, y_i) and (x_i', y_i') before taking the max.

$$E_{S,S',\sigma}\left[\max_{h \in H} \frac{[\sum_i \sigma_i (err_{x_i'}(h) - err_{x_i}(h))]}{m}\right]$$

$$\leq E_{S',\sigma}\left[\max_{h \in H} \frac{[\sum_i \sigma_i\, err_{x_i'}(h)]}{m}\right] - E_{S,\sigma}\left[\min_{h \in H} \frac{[\sum_i \sigma_i\, err_{x_i}(h)]}{m}\right]$$

$$= 2\, E_{S,\sigma}\left[\max_{h \in H} \frac{[\sum_i \sigma_i\, err_{x_i}(h)]}{m}\right]$$

19

Rademacher proof

- **Step 2:** show $E_S[MAXGAP(S)] \leq R_D(H)$.

$MAXGAP(S) = \max_{h \in H}[err_D(h) - err_S(h)]$

- **Proof (uses a ghost sample argument):**
  - Almost done: this looks very close to definition of $R_D(H)$.
  - There's an extra factor of 2.

$R_D(H) = E_{S,\sigma}\left[\max_{h \in H} \frac{1}{m}\sum_i \sigma_i h(x_i)\right]$

  - We are looking at the correlation of the losses of h with σ, rather than the correlation of h with σ.
  - To fix these, suppose we cheated by changing the def of $R_D(H)$ so that σ is a random $\{-1,1\}$ multiplier applied to the true labels rather than a random $\{-1,1\}$ labeling. Is that cheating?

$$= 2\, E_{S,\sigma}\left[\max_{h \in H} \frac{[\sum_i \sigma_i\, err_{x_i}(h)]}{m}\right]$$

20

Rademacher proof

- **Step 2:** show $E_S[MAXGAP(S)] \leq R_D(H)$.

$MAXGAP(S) = \max_{h \in H}[err_D(h) - err_S(h)]$

$$R_D(H) = E_S E_\sigma\left[\max_{h \in H} \frac{1}{m}\sum_i \sigma_i y_i h(x_i)\right] = E_S E_\sigma\left[\max_{h \in H} \frac{1}{m}\sum_i \sigma_i \left(1 - 2err_{x_i}(h)\right)\right]$$

$$= E_S E_\sigma\left[\frac{1}{m}\sum_i \sigma_i + \max_{h \in H} \frac{1}{m}\sum_i \left(-2\sigma_i err_{x_i}(h)\right)\right]$$

- To fix these, suppose we cheated by changing the def of $R_D(H)$ so that σ is a random $\{-1,1\}$ multiplier applied to the true labels rather than a random $\{-1,1\}$ labeling. Is that cheating?

$$= 2\, E_{S,\sigma}\left[\max_{h \in H} \frac{[\sum_i \sigma_i\, err_{x_i}(h)]}{m}\right]$$

21

Rademacher proof

- **Step 2:** show $E_S[MAXGAP(S)] \leq R_D(H)$.

$MAXGAP(S) = \max_{h \in H}[err_D(h) - err_S(h)]$

$$R_D(H) = E_S E_\sigma\left[\max_{h \in H} \frac{1}{m}\sum_i \sigma_i y_i h(x_i)\right] = E_S E_\sigma\left[\max_{h \in H} \frac{1}{m}\sum_i \sigma_i \left(1 - 2err_{x_i}(h)\right)\right]$$

$$= E_S E_\sigma\left[\frac{1}{m}\sum_i \sigma_i + \max_{h \in H} \frac{1}{m}\sum_i \left(-2\sigma_i err_{x_i}(h)\right)\right]$$

- Now we're done. First term is 0. Second term is 2 times the correlation with $-\sigma$, which is distributed exactly the same as σ.
- Proved by {Bartlett, Boucheron, Lugosi, Mendelson} 2000-2002.

-Done-

Possible project: understand/explain extensions to online case "sequential Rademacher complexity"

22