TTIC 31250 An Introduction to the Theory of Machine Learning

Rademacher Bounds

Avrim Blum 04/27/20

1

Uniform Convergence (VC)

- Theorem 1': For any class H, distribution D over $X \times \{-1,1\}$, if $m = |S| \ge \frac{8}{\epsilon^2} \left[\ln(H[2m]) + \ln\left(\frac{2}{\delta}\right) \right]$, then with prob 1- δ , all $h \in H$ have $|err_D(h) - err_S(h)| \le \epsilon$.
- Proof: same as for Thm 1 except def of B*:
- $B^*_{S',SS'}$ = event that $\exists h \in H$ with $|err_{S'}(h) err_{S}(h)| \ge \frac{\epsilon}{2}$.
- To show: for any |S''| = 2m, $\Pr_{S'',S,S'} \left[B^*_{S'',S,S'} \right] \le \delta/2$.
- Fix $h \in H[S^n]$, pairing $(a_1, b_1), \dots, (a_m, b_m)$. 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 $|\#heads \#tails| \ge \epsilon m/2$ in $m' \le m$ flips.
- View as ratio being off from expectation by $\geq \left(\frac{\epsilon m}{4m'}\right)$ and apply Hoeffding.

3

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: 2α + (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



2

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*].
- 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

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".

Some preliminaries

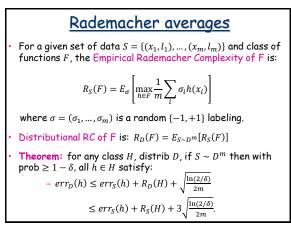
- Rather than writing m as a function of ε, write ε as function of m. E.g., would write Theorem 1' as:
- For any class H and distribution D, whp all h in H satisfy

eı

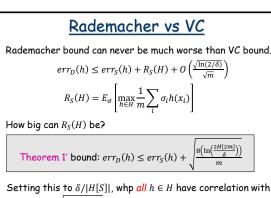
$$r_D(h) \le err_S(h) + \sqrt{\frac{8\left(\ln\left(\frac{2H[2m]}{\delta}\right)\right)}{m}}$$

(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



9



 σ at most $2\sqrt{\frac{\ln(|H[S]|/\delta)}{2m}}$. So, $R_S(H)$ can't be much larger.

<u>Rademacher vs VC</u>

Rademacher averages

For a given set of data $S = \{(x_1, l_1), \dots, (x_m, l_m)\}$ and class of

functions F, the Empirical Rademacher Complexity of F is:

 $R_{S}(F) = E_{\sigma} \left[\max_{h \in F} \frac{1}{m} \sum_{i} \sigma_{i} h(x_{i}) \right]$

I.e., if you pick a random labeling σ of S, on average how well

Note: $h: X \to \{-1, 1\}$ so $\sigma_i h(x_i) = 1$ if agree, -1 if disagree. Note "correlation" = agreement - disagreement, so error

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

correlated is the most-correlated $h \in F$?

45% means correlation of 10%.

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

$$err_{D}(h) \leq err_{S}(h) + R_{S}(H) + O\left(\frac{\sqrt{\ln(1/\delta)}}{\sqrt{m}}\right)$$
$$R_{S}(H) = E_{\sigma}\left[\max_{h \in H} \frac{1}{m} \sum_{i} \sigma_{i}h(x_{i})\right]$$

• How big can R_S(H) be?

- Class H produces labelings $h_1, ..., h_{|H[S]|}$ 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

