Finish agnostic finite PAC
+ no free lunch
+ start of VC dimension

CPSC 532S: Modern Statistical Learning Theory

19 January 2022

cs.ubc.ca/~dsuth/532S/22/

Admin

- We're now under the cap
  - If you aren't officially registered but want to be, email me your form ASAP
  - If you want to audit, email me your form ASAP
    - To audit: come to at least 75% of classes
      or a brief writeup at end of term (details TBD but it'll be short)

- A1 due tomorrow night
  - It's maybe harder than I intended – will do a calibration poll afterwards
  - Future assignments will allow groups
    - Might be shorter / longer to work on them / more hints available
  - For 1b in particular, Exercise 2.3 or Example 6.1 might give good inspiration
  - Office hours immediately after class today (until 3:55) and tomorrow 4-5

Last time: ERM with uniform convergence

- Want h_S to compete with best predictor in $\mathscr{H}$ with high probability

- First step: "good" S are **ε-representative**, $|L_S(h) - L_{\mathscr{D}}(h)| \leq \varepsilon$ for **all** h
  - The **generalization gap** is small, for all h
- Lemma: If S is $\varepsilon/2$-representative, then for *any* $h \in \mathscr{H}$,

  $L_{\mathscr{D}}(h_S) \leq L_S(h_S) + \frac{1}{2}\varepsilon \leq L_S(h) + \frac{1}{2}\varepsilon \leq L_{\mathscr{D}}(h) + \varepsilon$ and so $L_{\mathscr{D}}(h_S) \leq \inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) + \varepsilon$

- $\mathscr{H}$ has the **uniform convergence property** w.r.t. $\mathscr{Z}$ and ℓ if,
  with $n \geq n_{\mathscr{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ samples from *any* distribution $\mathscr{D}$ over $\mathscr{Z}$,
  $S \sim \mathscr{D}^n$ is ε representative with probability at least $1 - \delta$

- So: sufficient to show that finite $\mathscr{H}$ have the uniform convergence property

Last time: Finite $\mathscr{H}$ have the uniform convergence property

$$\Pr_S\left(\exists h \in \mathscr{H} \,.\, |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\right) \qquad \text{(we want to show it's } < \delta\text{)}$$

$$= \mathscr{D}^n\left(\bigcup_{h\in\mathscr{H}} \{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathscr{H}} \mathscr{D}^n\left(\{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right)$$

assume $A \leq \ell(h, z) \leq A + B$

$$\leq \sum_{h\in\mathscr{H}} 2\exp\left(-\frac{2}{B^2}n\varepsilon^2\right) = 2|\mathscr{H}|\exp\left(-\frac{2}{B^2}n\varepsilon^2\right)$$

Hoeffding Bound (1963)

Wassily Hoeffding

If $X_1, \ldots, X_n \in \mathbb{R}$ independent, $\mathbb{E}[X_i] = \mu$, $\Pr(a \leq X_i \leq b) = 1$,

then $\Pr\left(\left|\frac{1}{n}\sum X_i - \mu\right| > \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$

Last time: Finite $\mathscr{H}$ have the uniform convergence property

$$\Pr_S\left(\exists h \in \mathscr{H} \,.\, |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\right) \qquad \text{(we want to show it's } < \delta\text{)}$$

$$= \mathscr{D}^n\left(\bigcup_{h\in\mathscr{H}} \{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathscr{H}} \mathscr{D}^n\left(\{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right)$$

assume $A \leq \ell(h,z) \leq A + B$

$$\leq \sum_{h\in\mathscr{H}} 2\exp\left(-\frac{2}{B^2}n\varepsilon^2\right) = 2|\mathscr{H}|\exp\left(-\frac{2}{B^2}n\varepsilon^2\right)$$

$$2|\mathscr{H}|\exp\left(-\frac{2}{B^2}n\varepsilon^2\right) < \delta \text{ iff } -\frac{2}{B^2}n\varepsilon^2 < \log\frac{\delta}{2|\mathscr{H}|} \text{ iff } n > \frac{B^2}{2\varepsilon^2}\left[\log(2|\mathscr{H}|) + \log\frac{1}{\delta}\right]$$

ERM agnostically PAC-learns $\mathscr{H}$ with $n > \frac{2B^2}{\varepsilon^2}\left[\log(2|\mathscr{H}|) + \log\frac{1}{\delta}\right]$ samples

Last time: Finite $\mathscr{H}$ have the uniform convergence property

$$\Pr_S\left(\exists h \in \mathscr{H} \,.\, |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\right) \qquad \text{(we want to show it's } < \delta\text{)}$$

$$= \mathscr{D}^n\left(\bigcup_{h\in\mathscr{H}} \{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathscr{H}} \mathscr{D}^n\left(\{S : |L_S(h) - L_{\mathscr{D}}(h)| > \varepsilon\}\right)$$

assume $A \leq \ell(h, z) \leq A + B$

$$\leq \sum_{h\in\mathscr{H}} 2\exp\left(-\frac{2}{B^2}n\varepsilon^2\right) = 2|\mathscr{H}|\exp\left(-\frac{2}{B^2}n\varepsilon^2\right)$$

Equivalently: error of ERM over $\mathscr{H}$ is at most $\sqrt{\frac{2B^2}{n}\left[\log(2|\mathscr{H}|) + \log\frac{1}{\delta}\right]}$

ERM agnostically PAC-learns $\mathscr{H}$ with $n > \frac{2B^2}{\varepsilon^2}\left[\log(2|\mathscr{H}|) + \log\frac{1}{\delta}\right]$ samples

We need *uniform* convergence

A tempting,
but **wrong** argument:

“
Just apply a Hoeffding bound to h_S and $h^* = \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$

– then we’d only have to union two bounds instead of $|\mathcal{H}|$ in

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{1}{2}\varepsilon \leq L_S(h^*) + \frac{1}{2}\varepsilon \leq L_{\mathcal{D}}(h^*) + \varepsilon$$

and would get PAC learning with $n > \frac{2B^2}{\varepsilon^2} \log \frac{4}{\delta}$ – no $|\mathcal{H}|$!
”

The $\ell(h_S, z_i)$ terms here are *not* independent:

the identity of h_S depends on those same terms!

Realizable vs agnostic case

- High probability bounds on error for finite $\mathscr{H}$, 0-1 loss:
  - Realizable case: $\frac{1}{n}\left(\log|\mathscr{H}| + \log\frac{1}{\delta}\right)$
  - Agnostic case: $\sqrt{\frac{1}{2n}\left(\log|\mathscr{H}| + \log\frac{2}{\delta}\right)}$
- Possible to interpolate between them: "optimistic rates"
  - (maybe on assignment 2…)

Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression

Lijia Zhou* *Department of Statistics, University of Chicago* ZLJ@UCHICAGO.EDU

Frederic Koehler* *Simons Institute, University of California at Berkeley* FKOEHLER@BERKELEY.EDU

Danica J. Sutherland *University of British Columbia; Alberta Machine Intelligence Institute* DSUTH@CS.UBC.CA

Nathan Srebro *Toyota Technological Institute at Chicago* NATI@TTIC.EDU

Collaboration on the Theoretical Foundations of Deep Learning (deepfoundations.ai)

(pause)

No free lunch

The Lack of A Priori Distinctions Between Learning Algorithms (1996)

David H. Wolpert
The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, 87501, USA

- **Theorem:** For any learning algorithm A for binary classification (0-1 loss) on $\mathcal{X}$
  - Let $n < \frac{1}{2}|\mathcal{X}|$ be a training set size
  - Then there exists a $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:
    - There exists an $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$
    - With probability at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^n$, $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$

- **Corollary:** If $|\mathcal{X}| = \infty$, the set of all functions from $\mathcal{X}$ to $\{0,1\}$ is not PAC learnable.

No free lunch: basic proof idea

- Let $C \subseteq \mathcal{X}$ with $|C| = 2n$
- If we only care about C, there are 2^{2n} functions $f : C \to \{0,1\}$
  - Call them $f_1, f_2, \ldots, f_T$; let $\mathcal{D}_i$ have x uniform over C and $y = f_i(x)$
- Seeing n samples from C, there are at least n points in C we haven't seen
- The algorithm needs to pick one of the f_i, but it's just as likely to be wrong as right
- We need prior information to learn anything

So how do we pick an $\mathscr{H}$?

Key decomposition: $L_{\mathscr{D}}(h_S) = \varepsilon_{\text{approx}} + \varepsilon_{\text{est}}$ Bigger $\mathscr{H}$: smaller $\varepsilon_{\text{approx}}$, bigger ε_{est}

$\varepsilon_{\text{approx}} = \inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h)$ Approximation error: how good is $\mathscr{H}$

at the concept we're trying to learn?

Sometimes $\varepsilon_{\text{approx}}$ is defined as $\inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - \varepsilon_{\text{Bayes}}$

$\varepsilon_{\text{est}} = L_{\mathscr{D}}(h_S) - \varepsilon_{\text{approx}}$ Estimation error: how good are we at learning in $\mathscr{H}$?

Note: what 340 called "approximation error" is **something different** – $L_{\mathscr{D}}(h) - L_S(h)$, the generalization gap (more like estimation error)!

(pause)

Infinite classes

- So far (in class) we've only proved PAC learning for finite $|\mathscr{H}|$
- But homework problem 1b has infinitely many hypotheses, and it PAC learns
- Another example: threshold functions $h_a(x) = \mathbb{I}_{[x<a]}$ on $\mathbb{R}$ (Example 6.1)
- So: finite $|\mathscr{H}|$ is sufficient, but not necessary

Shattering

- No-free-lunch theorem relied on being able to choose *any* function on C
- So, to dodge it, we need to make sure that $\mathscr{H}$ can't do *everything* on C
- **Restriction of $\mathscr{H}$ to C** is $\mathscr{H}_C = \left\{ \left(h(c_1), \ldots, h(c_{|C|}) \right) : h \in \mathscr{H} \right\}$
- Say $\mathscr{H}$ **shatters** $C \subseteq \mathscr{X}$ if $\mathscr{H}_C$ contains all functions from C to $\{0,1\}$
  - Equivalent: $|\mathscr{H}_C| = 2^{|C|}$
- **Corollary** to no free lunch: if there is a $C \subseteq \mathscr{X}$ of size $2n$ shattered by $\mathscr{H}$, then there is a $\mathscr{D}$ over $\mathscr{X} \times \{0,1\}$ where there is a perfect predictor, but any learning algorithm A has probability at least 1/7 of error at least 1/8

VC dimension

- The **VC dimension** of $\mathscr{H}$, is the size of the largest set that $\mathscr{H}$ can shatter (or ∞ if it can shatter arbitrarily large sets)

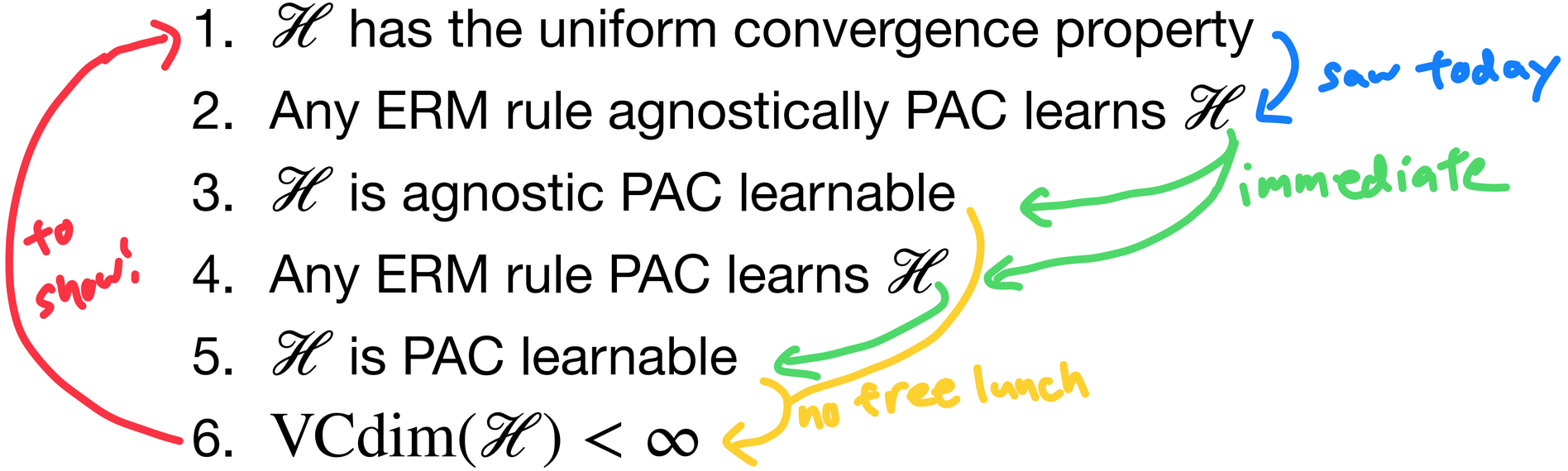
- Doesn't need that *all* sets of size VCdim can be shattered – it's **worst-case**
  - There is a C with $|C| = \mathrm{VCdim}$ that can be shattered
  - There is **no** C with $|C| = \mathrm{VCdim} + 1$ that can be shattered
- We'll cover some non-worst-case analyses soon

- **Corollary** of no-free-lunch: if VCdim($\mathscr{H}$) = ∞, $\mathscr{H}$ is not PAC learnable

“Fundamental Theorem” of Learning

For binary classification with 0-1 loss, these are all equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. Any ERM rule PAC learns $\mathscr{H}$
5. $\mathscr{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathscr{H}) < \infty$



And with some numbers:

- For a binary classification problem with 0-1 loss, if $\mathrm{VCdim}(\mathscr{H}) = d$:
  - $\mathscr{H}$ has uniform convergence property, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathscr{H}}^{UC} \leq \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is agnostic PAC learnable, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathscr{H}} \leq \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is PAC learnable, $\frac{C_1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathscr{H}} \leq \frac{C_2}{\varepsilon}\left[d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$