PAC learning
+ uniform convergence

CPSC 532D: Modern Statistical Learning Theory
14 September 2022
cs.ubc.ca/~dsuth/532D/22w1/
Admin

• Everyone should be registered now; if not, talk to me
  • If you want to audit, email me a form

• A1 is up
  • Work in pairs if you want
  • Cite any sources you use other than the course books (SSBD, MRT, Tel)
    • Including talking to people not in your group: say so + what extent
  • Gradescope link to submit will be up soon

• UBC is closed next Monday for the Queen’s funeral
  • So, class is canceled again…sorry
  • Assignment deadline likely to become Tuesday – will update on Piazza

• Final is scheduled: Wednesday Dec 14, 2-4:30pm, ICCS 246
  • Let me know if there’s a serious problem and we can maybe adapt
Last time: definitions

• \((x, y) \sim \mathcal{D}\), a distribution over \(\mathcal{E} = \mathcal{X} \times \mathcal{Y}\)
• Training “set” \(S = (z_1, \ldots, z_n) = ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n\)
• Loss function \(\ell : \mathcal{H} \times \mathcal{E} \to \mathbb{R}\), e.g. \(\ell_{0-1}(h, (x, y)) = \mathbb{I}(h(x) \neq y)\)
• Want to find \(h\) minimizing \(L_\mathcal{D}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]\), e.g. error rate = 1-accuracy for 0-1
  • name \(\in \{\text{“true”, “population”}\} \times \{\text{“risk”, “loss”}\}\)
  
  Have \(L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i); \quad \text{name} \in \{\text{“empirical”, “training”}\} \times \{\text{“risk”, “loss”}\}\)

• Empirical risk minimization (ERM): choose \(h\) minimizing \(L_S(h)\) from a \textit{hypothesis class} \(\mathcal{H}\) of functions \(h : \mathcal{X} \to \mathcal{Y}\)
• To start with something simple, assume \textit{realizability} for a nonnegative loss: 
  \[\text{there is an } h^* \in \mathcal{H} \text{ with } L_\mathcal{D}(h^*) = 0\]
  • Implies (a.s.) that \(L_S(h^*) = 0\)
Realizable, finite $\mathcal{H}$

- Assume $0 \leq \ell(h, z) \leq 1$ for all $h, z$; also assume realizability
- $\hat{h}_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$
  - Realizable means that $L_S(\hat{h}_S) = 0$, but maybe $L_\mathcal{D}(\hat{h}_S) > 0$
- Would like to show $\Pr_S \left( L_\mathcal{D}(\hat{h}_S) \leq \varepsilon \right) \geq 1 - \delta$, i.e. $\Pr(L_\mathcal{D}(h_S) > \varepsilon) < \delta$
- Call $\mathcal{H}_\varepsilon$ the set of “bad” hypotheses, $\left\{ h \in \mathcal{H} : L_\mathcal{D}(h) > \varepsilon \right\}$
- If ERM failed, $S$ must be consistent with a bad hypothesis:

\[
\Pr(L_\mathcal{D}(\hat{h}_S) > \varepsilon) \leq \Pr \left( S \in \bigcup_{h \in \mathcal{H}_\varepsilon} \{ S : L_S(h) = 0 \} \right) \leq \sum_{h \in \mathcal{H}_\varepsilon} \Pr_{S \sim \mathcal{D}^n}(L_S(h) = 0)
\]
Realizable, finite $\mathcal{H}$

- $\Pr(L_{\mathcal{D}}(\hat{h}_S) > \varepsilon) \leq \sum_{h \in \mathcal{H}_\varepsilon} \Pr(L_S(h) = 0)$

- $\Pr(L_S(h) = 0) = \Pr(\forall i \in [n] \cdot \ell(h, z_i) = 0)$

Because $S$ is iid, this is just $\prod_{i=1}^{n} \Pr(\ell(h, z_i) = 0) = p_0(h)^n$

where $p_0(h) = \Pr(\ell(z, h) = 0)$

- Know that $L_{\mathcal{D}}(h) = p_0(h) \times 0 + (1 - p_0(h)) \times \mathbb{E}_z[\ell(z, h) \mid \ell(z, h) > 0]$
  - So, if $L_{\mathcal{D}}(h) > \varepsilon$, then must have $1 - p_0(h) > \varepsilon$, i.e. $p_0(h) < 1 - \varepsilon$

- $\Pr(L_{\mathcal{D}}(\hat{h}_S) > \varepsilon) < \sum_{h \in \mathcal{H}_\varepsilon} (1 - \varepsilon)^n$
  
  $= |\mathcal{H}_\varepsilon|(1 - \varepsilon)^n < |\mathcal{H}|(1 - \varepsilon)^n \leq |\mathcal{H}|e^{-\varepsilon n}$

If a hypothesis is bad, we’re likely to sample at least one data point where it’s wrong

$1 - \varepsilon \leq e^{-\varepsilon}$

Not too likely to get unlucky with any bad hypothesis
Finite $\mathcal{H}$ are (realizable) PAC-learnable

- We showed that $\Pr \left( L_{\mathcal{D}}(\hat{h}_S) < \epsilon \right) \geq 1 - |\mathcal{H}| e^{-\epsilon n}$
- Or: if we have $n \geq \frac{1}{\epsilon} \left( \log |\mathcal{H}| + \log \frac{1}{\delta} \right)$, $L_{\mathcal{D}}(h) \leq \epsilon$ with prob. at least $1 - \delta$.
- Or: error is at most $\frac{1}{n} \left( \log |\mathcal{H}| + \log \frac{1}{\delta} \right)$ with probability at least $1 - \delta$

$\mathcal{H}$ is **PAC learnable** if there is a function $n_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ and a learning alg. s.t.:
- For every $\epsilon, \delta \in (0,1)$, for every $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ which is realizable by $\mathcal{H}$,
- then running the algorithm on $n \geq n_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples from $\mathcal{D}$
- will return a hypothesis $h$ with $L_{\mathcal{D}}(h) \leq \epsilon$
- with probability at least $1 - \delta$ over the choice of examples $S$
Example: Boolean conjunctions

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

\( \mathcal{H} \): conjunctions of the form 
\[
a \land \bar{c} \land f
\]

Algorithm:
- Start with \( a \land \bar{a} \land \cdots \land f \land \bar{f} \)
- Cross out bits inconsistent with the positives
Example: Boolean conjunctions

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

$\mathcal{H}$: conjunctions of the form

$$a \land \bar{c} \land f$$

Algorithm:

- Start with $a \land \bar{a} \land \cdots \land f \land \bar{f}$
- Cross out bits inconsistent with the positives
Example: Boolean conjunctions

$\mathcal{H}$: conjunctions of the form $a \land \bar{c} \land f$

Algorithm:
- Start with $a \land \bar{a} \land \cdots \land f \land \bar{f}$
- Cross out bits inconsistent with the positives
Example: Boolean conjunctions

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
</tbody>
</table>

$\mathcal{H}$: conjunctions of the form

$$a \land \bar{c} \land f$$

Algorithm:
- Start with $a \land \bar{a} \land \cdots \land f \land \bar{f}$
- Cross out bits inconsistent with the positives
Example: Boolean conjunctions

\[ c \land \overline{d} \land f \]

\[ |\mathcal{H}| = 3^d: \left\lfloor \frac{1}{\varepsilon} \left( d \log(3) + \log \frac{1}{\delta} \right) \right\rfloor \text{ samples enough} \]

\[ \mathcal{H}: \text{conjunctions of the form } a \land \overline{c} \land f \]

Algorithm:
- Start with \( a \land \overline{a} \land \cdots \land f \land \overline{f} \)
- Cross out bits inconsistent with the positives

Assuming realizability, this gives an ERM
- Algorithm makes every + example a +
- True function f is only “less specific” than h: h(x) = - for anything truly -
So, are we done with the course?

- Every practical $\mathcal{H}$ is finite if you put it on a computer
- Total size of weights in a big deep network is typically up to $\sim 1$GB
- Say 100MB, $8 \times 100 \times 2^{20}$ bits, so there are $2^{25} \cdot 2^{25}$ possible networks
  - $\log \left( 2^{25} \cdot 2^{25} \right) = 25 \cdot 2^{25} \log(2) \approx 252$ million
- If we want, say, $\varepsilon = 0.1$ (90% accuracy): 2.5 billion training points
- (Plus, we don’t actually do ERM with realizable, fixed hypothesis classes...)
PAC learnability and computational efficiency

• Valiant (1984)’s formulation required the algorithm to run in polynomial time
• We’re going to mostly not care about runtime (call poly version “efficient PAC learning”), but be aware many authors keep that in the definition

• Independent(?), closely related development by Vapnik and Chervonenkis in the USSR; much more on their work soon
PAC learnability and computational efficiency

• A class that can be PAC-learned but **not in polynomial time** (assuming \( P = \text{BPP} \) and \( P \neq \text{NP} \)):

• 3-DNF: 3-term clauses in *disjunctive normal form*
  \[ T_1 \lor T_2 \lor T_3 \]
  terms are conjunctions: \( T_1 = a \land \bar{c} \land \cdots \)

• Graph 3-coloring reduces to learning 3-DNFs

• But: 3-DNF \( \subset \) 3-CNF, \( \bigwedge (a \lor b \lor c) \),
  \[ T_1 \lor T_2 \lor T_3 = \bigwedge_{u \in T_1, v \in T_2, w \in T_3} (u \lor v \lor w) \]

• and 3-CNF **can** be efficiently PAC-learned
(pause)
Non-realizable (agnostic) learning

• What if we don’t know that \( \mathcal{H} \) can realize \( \mathcal{D} \)?
  • (Does the class of ResNet-101s realize ImageNet? 🤷‍♂️)

• What if we know that \( \mathcal{H} \) can’t realize \( \mathcal{D} \)?
  • If one \( x \) can have two possible \( y \)s, no function can get zero loss*
    • *if there’s a positive probability of getting such an \( x \)
Agnostic PAC

- \( \mathcal{H} \) is **agnostically PAC learnable** for a set \( \mathcal{X} \) and loss \( \ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R} \) if there is a function \( n_\mathcal{H} : (0,1)^2 \to \mathbb{N} \) and a learning algorithm such that:
  
  For every \( \varepsilon, \delta \in (0,1) \) and every distribution \( \mathcal{D} \) over \( \mathcal{X} \),
  
  then running the algorithm on \( n \geq n_\mathcal{H}(\varepsilon, \delta) \) i.i.d. examples from \( \mathcal{D} \)
  
  will return a hypothesis \( h \in \mathcal{H} \) with \( L_\mathcal{D}(h) \leq \inf_{h' \in \mathcal{H}} L_\mathcal{D}(h') + \varepsilon \)
  
  with probability at least \( 1 - \delta \) over the choice of examples

- We don’t (necessarily) get error arbitrarily close to 0 anymore!
  
  - Realizable means \( \inf_{h' \in \mathcal{H}} L_\mathcal{D}(h') = 0 \): then, this is same as realizable PAC
  
  - Otherwise, \( \inf_{h' \in \mathcal{H}} L_\mathcal{D}(h') \) is the best loss achievable in \( \mathcal{H} \)
Improper Agnostic PAC

- $\mathcal{H}$ is improperly agnostically PAC learnable in $\mathcal{H}'$ for $\mathcal{L}$, loss $\ell : \mathcal{H}' \times \mathcal{L} \rightarrow \mathbb{R}$ if there is a function $n_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm such that:
  - For every $\varepsilon, \delta \in (0,1)$ and every distribution $\mathcal{D}$ over $\mathcal{L}$, then running the algorithm on $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples from $\mathcal{D}$ will return a hypothesis $h \in \mathcal{H}' \supset \mathcal{H}$ with $L_{\mathcal{D}}(h) \leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$
  - with probability at least $1 - \delta$ over the choice of examples

- e.g.: learn a polynomial classifier almost as good as the best linear classifier, or learn a 3-DNF function with a 3-CNF

- Shai+Shai: “there is nothing improper about representation-independent learning”
Bayes error rate

• What can we say about $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$?

• It’s at least as big as the **Bayes error**: error of the Bayes-optimal predictor
  
  e.g. for 0-1 loss, $f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \Pr(y = 1 \mid x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

• The best predictor in $\mathcal{H}$ might be as good as this, or it might be worse

• Gap between Bayes error and $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ called **approximation error**
ERM on finite classes, agnostic edition

- Want $\hat{h}_S$ to compete with best predictor in $\mathcal{H}$ with high probability

- First step: “good” $S$ are $\varepsilon$-representative, $|L_S(h) - L_\mathcal{D}(h)| \leq \varepsilon$ for all $h$
  - The generalization gap is small, for all $h$
  - Lemma: If $S$ is $\varepsilon$-representative, then for any comparator $h' \in \mathcal{H}$,
    \[ L_\mathcal{D}(\hat{h}_S) \leq L_S(\hat{h}_S) + \varepsilon \leq L_S(h') + \varepsilon \leq L_\mathcal{D}(h') + 2\varepsilon \]
    and so $L_\mathcal{D}(\hat{h}_S) \leq \inf_{h \in \mathcal{H}} L_\mathcal{D}(h) + 2\varepsilon$

- $\mathcal{H}$ has the uniform convergence property w.r.t. $\mathcal{F}$ and $\ell$ if, with $n \geq n_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ samples from any distribution $\mathcal{D}$ over $\mathcal{F}$,
  \[ S \sim \mathcal{D}^n \] is $\varepsilon$ representative with probability at least $1 - \delta$

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

\[
\Pr\left( \exists h \in \mathcal{H} \mid |L_S(h) - L_\mathcal{D}(h)| > \varepsilon \right) (\text{we want to show it's } < \delta)
\]

\[
= \Pr\left( S \in \bigcup_{h \in \mathcal{H}} \{ S : |L_S(h) - L_\mathcal{D}(h)| > \varepsilon \} \right) \leq \sum_{h \in \mathcal{H}} \Pr_{S \sim \mathcal{D}^n} \left( |L_S(h) - L_\mathcal{D}(h)| > \varepsilon \right)
\]

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

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

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

\[
\Pr \left( \left| \frac{1}{n} \sum X_i - \mu \right| > \varepsilon \right) \leq 2 \exp \left( \frac{-2n\varepsilon^2}{(b - a)^2} \right)
\]
Finite $\mathcal{H}$ have the uniform convergence property

$$\Pr_S \left( \exists h \in \mathcal{H} . |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon \right)$$

(we want to show it’s < $\delta$)

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

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

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

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

ERM agnostically PAC-learns $\mathcal{H}$ with $n > \frac{2B^2}{\varepsilon^2} \left[ \log(2|\mathcal{H}|) + \log \frac{1}{\delta} \right]$ samples
Finite $\mathcal{H}$ have the uniform convergence property

$$\Pr_S \left( \exists h \in \mathcal{H} . \ |L_S(h) - L_\mathcal{D}(h)| > \varepsilon \right) \leq \sum_{h \in \mathcal{H}} \Pr_{S \sim \mathcal{D}^n} \left( |L_S(h) - L_\mathcal{D}(h)| > \varepsilon \right)$$

(we want to show it’s $< \delta$)

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

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

Equivalently: error of ERM over $\mathcal{H}$ is at most

$$\sqrt{\frac{2B^2}{n} \left[ \log(2|\mathcal{H}|) + \log \frac{1}{\delta} \right]}$$

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

- ERM for finite hypothesis classes, $n$ to get excess error $\varepsilon$ w/ prob. $1 - \delta$, for a loss bounded in $[0,1]$:
  - Realizable: $n \geq \frac{1}{\varepsilon} \left( \log |\mathcal{H}| + \log \frac{1}{\delta} \right)$ “$\frac{1}{n}$ rate”
  - Agnostic: $n > \frac{2}{\varepsilon^2} \left[ \log |\mathcal{H}| + \log \frac{2}{\delta} \right]$ “$\frac{1}{\sqrt{n}}$ rate”

- Late in the course, we’ll (probably) see “optimistic rates”: interpolate between the two regimes based on $\inf_{h \in \mathcal{H}} L_\mathcal{D}(h)$
Summary

- PAC learnability: realizable, agnostic, improper
- Finite classes are PAC learnable, both in realizable and agnostic settings
  - but rate is different

- **Uniform convergence** of $L_S(h)$ to $L_\mathcal{D}(h)$ over $\mathcal{H}$
  - Key tool: Hoeffding bound (a concentration inequality)

- Next time: choosing $\mathcal{H}$; what about infinite hypothesis classes?