Recap & Today

- Notion of batch learning
- Identically, independently, distributed (i.i.d) samples from $\mathcal{D}$
- Probably Approximately Correct learning
- PAC learnability with *finite* hypothesis classes
- Agnostic PAC learnability
- Agnostic learning of finite hypothesis classes
- Infinite hypothesis classes
PAC Learning

- Accuracy, $\varepsilon$, and confidence, $\delta$, parameters
- Training data, $S$, of $m(\varepsilon, \delta) = |S|$ i.i.d samples from an unknown distribution $\mathcal{D}$
- Find an hypothesis $h$ s.t.

$$\mathcal{L}_\mathcal{D}(h) \leq \varepsilon$$
PAC Learning

- Accuracy, $\varepsilon$, and confidence, $\delta$, parameters
- Training data, $S$, of $m(\varepsilon, \delta) = |S|$ i.i.d samples from an unknown distribution $\mathcal{D}$
- Find an hypothesis $h$ s.t.

\[
\mathcal{L}_\mathcal{D}(h) \leq \varepsilon \approx \]

Q.1 What candidate hypotheses for $h$ to consider?
Q.2 How to assess $\mathcal{L}_\mathcal{D}(h)$?
PAC Learning

- Accuracy, $\varepsilon$, and confidence, $\delta$, parameters

- Training data, $S$, of $m(\varepsilon, \delta) = |S|$ i.i.d samples from an unknown distribution $\mathcal{D}$

- Find an hypothesis $h$ s.t.

$$\mathcal{L}_\mathcal{D}(h) \leq \varepsilon \quad \text{w.p. } 1 - \delta$$
PAC Learning

- Accuracy, $\varepsilon$, and confidence, $\delta$, parameters
- Training data, $S$, of $m(\varepsilon, \delta) = |S|$ i.i.d samples from an unknown distribution $\mathcal{D}$
- Find an hypothesis $h$ s.t.

$$\mathcal{L}_\mathcal{D}(h) \leq \varepsilon \approx w.p. \ 1 - \delta$$

- **Q.1** What candidate hypotheses for $h$ to consider?
- **Q.2** How to assess $\mathcal{L}_\mathcal{D}(h)$
Perils of Lack of (Prior) Knowledge

- Suppose $|\mathcal{X}|$ is infinite
- Pick an arbitrarily large $m$
- $R$ is a random set of examples of size $2m$
- Define $\mathcal{D}(x) = \frac{1}{2m}$ if $x \in S$ and 0 o.w.
- Set $S$ to $m$ random samples from $R$ according to $\mathcal{D}$
- Number of unique instances in $S$ is at most $m$
- Suppose $\mathcal{H}$ consists of all functions from $\mathcal{X}$ to $\{-1, +1\}$
- Any learning algorithm can only guess the labels of $R - S$
- Since $|R - S| \geq m$ error of predicted hypothesis would have an error rate of about $1/4$ (in expectation)
- Need to constrain the hypothesis class $\mathcal{H}$
Finite Hypothesis Classes

• Assume that $\mathcal{H}$ has finite number of hypotheses
  • $\mathcal{X} = \{-1, 1\}^n$, $Y = \{0, +1\}$, and $\mathcal{H}$ is all truth tables
  • Linear thresholds of the form $\text{sign}(\mathbf{w} \cdot \mathbf{x})$ with $w_j = \frac{i}{j}$, $i,j \in [k]$
  • All Python functions that take at most $b_1$ bytes and with memory of $b_2$ bytes (very large but finite) with inputs over $\{0, 1\}^{32}$

• Distinguish between the following cases:
  • Realizable: $h^* \in \mathcal{H}$ such that for all $(\mathbf{x}, y)$, $h^*(\mathbf{x}) = y$
  • Agnostic: not realizable, but either $D(+1|\mathbf{x}) = 1$ or $D(-1|\mathbf{x}) = 1$
  • Stochastic: not agnostic, $0 < D(+1|\mathbf{x}) < 1$ for "many" $\mathbf{x}$
Empirical Risk Minimization

- **Input:** training set $S = \{(x^i, y^i)\}_{i=1}^m$

- **Realizable case:**
  - Output: $h \in \mathcal{H}$ s.t. $\forall i, y^i = h(x^i)$

- **Unrealizable case:**
  - **Empirical risk:**
    \[
    \mathcal{L}_S(h) = \frac{1}{m} \left| \{ i : h(x^i) \neq y^i \} \right|
    \]
  - Output:
    \[
    h = \arg \min_{h \in \mathcal{H}} \mathcal{L}_S(h)
    \]
  - Can use same ERM procedure
  - $\mathcal{L}_S(h) = 0$ in realizable case
  - Why distinguish between the two settings?
ERM in Realizable Settings

View ERM as a function that takes $\mathcal{H}$ and $S$ as inputs and returns $h \in \mathcal{H}$ such that $\mathcal{L}_S(h) = 0$

**Theorem (Relizable PAC)**

Fix $\varepsilon, \delta$ and assume realizability. If the number of examples

$$m \geq \frac{\log(|\mathcal{H}|) + \log(1/\delta)}{\varepsilon}$$

then for every $\mathcal{D}$, with probability of at least $1 - \delta$ (over the choice of $S$ of size $m$),

$$\mathcal{L}_D(ERM(S, \mathcal{H})) \leq \varepsilon.$$
Proof

• Let $L_D(h)$ be the loss of $h$ on (unknown) $D$
• Note that $S$ is a random set determined by $D$
• We need to prove that the probability mass of $S$ for which ERM returns inaccurate hypothesis is at most $\delta$

$$D(\{S : L_D(ERM(S, \mathcal{H})) > \varepsilon\}) \leq \delta$$

• Let $\mathcal{H}_B$ be the set of “inaccurate” hypotheses,

$$\mathcal{H}_B = \{ h \in \mathcal{H} : L_D(h) > \varepsilon \}$$

• Let $M$ be the set of “ill-guiding” samples (set of sets),

$$M = \{ S : \exists h \in \mathcal{H}_B, L_S(h) = 0 \}$$

$$= \bigcup_{h \in \mathcal{H}_B} \{ S : L_S(h) = 0 \}$$

• First, note that

$$\{ S : L_D(ERM(S, \mathcal{H})) > \varepsilon \} \subseteq M = \bigcup_{h \in \mathcal{H}_B} \{ S : L_S(h) = 0 \}$$
Next we use the **Union Bound**: for $\forall A, B$ distribution $\mathcal{D}$

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$
Proof (Cont.)

Next we use the Union Bound: for $\forall A, B$ distribution $\mathcal{D}$

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$

Therefore, using the union bound

$$\mathcal{D}([S : \mathcal{L}_D(\text{ERM}(S, \mathcal{H})) > \varepsilon])$$

$$\leq \sum_{h \in \mathcal{H}_B} \mathcal{D}([S : \mathcal{L}_S(h) = 0])$$

$$\leq |\mathcal{H}_B| \max_{h \in \mathcal{H}_B} \mathcal{D}([S : \mathcal{L}_S(h) = 0])$$
Proof (Cont.)

- Next, we use, $D(\{S : \mathcal{L}_S(h) = 0\}) = (1 - \mathcal{L}_D(h))^m$

- If $h \in \mathcal{H}_B$ then $\mathcal{L}_D(h) > \varepsilon$ and therefore

  $$D(\{S : \mathcal{L}_S(h) = 0\}) < (1 - \varepsilon)^m$$

- We showed that,

  $$D(\{S : \mathcal{L}_D(\text{ERM}(S, \mathcal{H}) > \varepsilon}\}) < |\mathcal{H}_B| (1 - \varepsilon)^m$$

- Finally, using $1 - \varepsilon \leq e^{-\varepsilon}$ and $|\mathcal{H}_B| \leq |\mathcal{H}|$ we get,

  $$D(\{S : \mathcal{L}_D(\text{ERM}(S, \mathcal{H}) > \varepsilon}\}) < |\mathcal{H}| e^{-\varepsilon m}$$

- The right-hand side would be $\leq \delta$ if

  $$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$$
Hypothesis class $\mathcal{H}$ is PAC learnable using algorithm $A$ if for all $m \geq m_\mathcal{H}(\epsilon, \delta)$, any distribution $\mathcal{D}$ over $\mathcal{X}$, then $L_\mathcal{D}(h) \leq \epsilon$ with probability $1 - \delta$ where $h = A(S, \mathcal{H})$. 
PAC Learnability

Hypothesis class $\mathcal{H}$ is PAC learnable using algorithm $\mathcal{A}$ if for all $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, any distribution $\mathcal{D}$ over $\mathcal{X}$, then $\mathcal{L}_D(h) \leq \epsilon$ with probability $1 - \delta$ where $h = \mathcal{A}(S, \mathcal{H})$.

$m_{\mathcal{H}}$ is termed the sample complexity of learning $\mathcal{H}$
Agnostic PAC Learning

- So far, assumed labels are generated by $h^* \in \mathcal{H}$
- Assumption is often unrealistic
- Instead of $\mathcal{D}$ over $\mathcal{X}$ let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$
- Replace $\exists h^*$ with conditional distribution $\mathcal{D}(y|x)$
- Define risk as:

$$L_{\mathcal{D}}(h) \overset{\text{def}}{=} \mathbb{P}_{(x,y) \sim \mathcal{D}}[h(x) \neq y]$$

- Relax notion of “approximately correct”

$$L_{\mathcal{D}}(\mathcal{A}(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \leq \varepsilon$$
Realizable vs. Agnostic PAC

<table>
<thead>
<tr>
<th></th>
<th>PAC</th>
<th>Agnostic PAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist</td>
<td>$\mathcal{D}$ over $\mathcal{X}$</td>
<td>$\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$</td>
</tr>
<tr>
<td>Truth</td>
<td>$h^* \in \mathcal{H}$</td>
<td>not in class, may not exist</td>
</tr>
<tr>
<td>Risk</td>
<td>$L_{\mathcal{D}}(h) = \mathcal{D}({x : h(x) \neq h^*(x)})$</td>
<td>$L_{\mathcal{D}}(h) = \mathcal{D}({(x, y) : h(x) \neq y})$</td>
</tr>
<tr>
<td>Input</td>
<td>${x^i}_i \sim \mathcal{D}^m$</td>
<td>${(x^i, y^i)}_i \sim \mathcal{D}^m$</td>
</tr>
<tr>
<td></td>
<td>$\forall i, y_i = h^*(x_i)$</td>
<td></td>
</tr>
<tr>
<td>Goal</td>
<td>$\mathcal{L}_{\mathcal{D}}(\mathcal{A}(S)) \leq \varepsilon$</td>
<td>$\mathcal{L}<em>{\mathcal{D}}(\mathcal{A}(S)) \leq \min</em>{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) + \varepsilon$</td>
</tr>
</tbody>
</table>
Agnostic PAC

Require that for every $\varepsilon, \delta \in (0, 1)$, $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, and
distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

$$\mathcal{D} \left( \left\{ S \in (\mathcal{X} \times \mathcal{Y})^m : \mathcal{L}_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) + \varepsilon \right\} \right) \geq 1 - \delta$$
Representative Sample

A training set $S$ is called $\varepsilon$-representative if

$$\forall h \in \mathcal{H}, \ |\mathcal{L}_S(h) - \mathcal{L}_D(h)| \leq \varepsilon$$
Representative Sample

A training set \( S \) is called \( \varepsilon \)-representative if

\[
\forall h \in \mathcal{H}, \quad |\mathcal{L}_S(h) - \mathcal{L}_D(h)| \leq \varepsilon
\]

Lemma Assume that a training set \( S \) is \( \varepsilon \)-representative. Then, the output of \( \text{ERM}_\mathcal{H}(S) \),

\[
\hat{h} \in \arg \min_{h \in \mathcal{H}} \mathcal{L}_S(h)
\]

satisfies

\[
\mathcal{L}_D(\hat{h}) \leq \min_{h \in \mathcal{H}} \mathcal{L}_D(h) + 2\varepsilon.
\]
For every $h \in \mathcal{H}$,

$$\mathcal{L}_D(\hat{h}) \leq \mathcal{L}_S(\hat{h}) + \varepsilon$$

$$\leq \mathcal{L}_S(h) + \varepsilon$$

$$\leq \mathcal{L}_D(h) + \varepsilon + \varepsilon$$

$$= \mathcal{L}_D(h) + \varepsilon$$
Agnostic PAC for Finite Classes

Assume $\mathcal{H}$ is finite. Then, $\mathcal{H}$ is agnostically PAC learnable using ERM with sample complexity

$$\left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$
Proof (cont.)

• We need to show

\[ \mathcal{D}(\{ S : \exists h \in \mathcal{H}, |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \}) < \delta \]
Proof (cont.)

- We need to show

\[ \mathcal{D} \left( \{ S : \exists h \in \mathcal{H}, |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \} \right) < \delta \]

- Using the union bound,

\[
\begin{align*}
\mathcal{D} \left( \{ S : \exists h \in \mathcal{H}, |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \} \right) \\
= \mathcal{D} (\bigcup_{h \in \mathcal{H}} \{ S : |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \}) \\
\leq \sum_{h \in \mathcal{H}} \mathcal{D} (\{ S : |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \})
\end{align*}
\]
Hoeffding’s inequality

Let $z_1, \ldots, z_m$ be a sequence of i.i.d. $\sim B(\theta)$. Denote by 
\[ \hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} z_i \] 
their empirical average. Then, for any $\varepsilon > 0$, 
\[ P[|\hat{\theta} - \theta| > \varepsilon] \leq 2 e^{-2m\varepsilon^2} \]
Hoeffding’s inequality

Let \( z_1, \ldots, z_m \) be a sequence of i.i.d. \( \sim B(\theta) \). Denote by

\[ \hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} z_i \]

their empirical average. Then, for any \( \varepsilon > 0 \)

\[ \mathbb{P} [ |\hat{\theta} - \theta| > \varepsilon ] \leq 2 e^{-2m \varepsilon^2} \]

This implies:

\[ \mathcal{D}( \{ S : |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \} ) \leq 2 \exp \left( -2m \varepsilon^2 \right) . \]
Concluding

We showed

$$\mathcal{D}(\{S : \exists h \in \mathcal{H}, |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon\}) \leq 2|\mathcal{H}| e^{-2m\varepsilon^2}$$
Concluding

We showed

\[ D(\{ S : \exists h \in \mathcal{H}, |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon \}) \leq 2|\mathcal{H}| e^{-2m\varepsilon^2} \]

We want \( 2|\mathcal{H}| e^{-2m\varepsilon^2} \leq \delta \) and therefore,

\[ m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \]
Infinite Classes Made Finite

- $\mathcal{H}$ is “parameterized” by $n$ numbers
- Assume it’s sufficient to use floating points
- Then $|\mathcal{H}| \leq 2^{32n}$,

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{64n + 2 \log(2/\delta)}{\varepsilon^2} \right\rceil$$

- Sample complexity of $\tilde{O}\left(\frac{n}{\varepsilon^2}\right)$ is not too shabby
- However, ERM would take exponential time in the dimension
- In reasonably small ML applications $n \approx 10^5$ ...