Lecture 2:
PAC Learning and VC Theory I
From Adversarial Online to Statistical

• Three reasons to move from worst-case deterministic to stochastic:
  • Deal with errors. What if data not *exactly* realized by $\mathcal{H}$?
  • Avoid non-learnability due to very specific, adversarial, order of examples
  • Training on dedicated training set, then predict on “population”
The Statistical Learning Model

- **Unknown source distribution** $\mathcal{D}$ over $(x, y)$
  - Describes “reality”. What we want to classify, and what should it be classified as.
    - E.g. joint distribution over $(b, b)$
    - Can think of $\mathcal{D}$ as: distribution over $x$ and $y| x = f(x)$
      - Distribution over images we expect to see (we don’t expect to see uniformly distributed images: $\overline{\mathcal{D}}$), and what character each image represents
    - Or, as: distribution over $y$ and over $x|y$
      - Distribution over characters (‘e’ more likely then ‘&’), and for each character, over possible images of that character

- **Goal**: find predictor $h$ with small **expected error**:
  \[ L_\mathcal{D}(h) = \mathbb{P}_{(x, y) \sim \mathcal{D}}[h(x) \neq y] \]
  (also called generalization error, risk or true error)

- Based on a sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ of $m$ training points $(x_t, y_t) \sim \text{i. i. d.} \mathcal{D}$ (we can also write: $S \sim \mathcal{D}^m$)
The Statistical Learning Model

• Unknown source distribution $\mathcal{D}$ over $(x, y)$

• Goal: find predictor $h$ with small expected error:

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

• Based on sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ of $m$ training points $(x_t, y_t) \sim \text{i.i.d.} \mathcal{D}$ (i.e. $S \sim \mathcal{D}^m$)

• Statistical (batch) learning:
  1. Receive training set $S \sim \mathcal{D}^m$
  2. Learn $h = A(S)$ using learning rule $A: (X \times Y)^* \rightarrow Y^X$
  3. Use $h$ on future examples drawn from $\mathcal{D}$, suffering expected error $L_\mathcal{D}(h)$

• Main assumption:
  • i.i.d. samples
  • Samples drawn from distribution $\mathcal{D}$ we will later use the predictor on
What we care about is the expected error

\[ L_D(h) = \mathbb{P}_{(x,y) \sim D}[h(x) \neq y] \]

Why not just minimize it directly?

For a given sample \( S \) we can calculate the empirical error

\[ L_S(h) = \frac{1}{m} \sum_{t=1}^{m} [[ h(x_t) \neq y_t ]] \]

How do we use the empirical error?

Is it a good estimate for the expected error?

How good?
The Empirical Error as an Estimator for the Expected Error

- How close are the expected and empirical errors?
  \[ |L_S(h) - L_D(h)| \]

Random Variable Number: \( L_D(h) = \mathbb{P}_{(x,y) \sim D}[h(x) \neq y] \)

\[
L_S(h) = \frac{1}{m} \sum_{t=1}^{m} [h(x_t) \neq y_t] \sim \frac{1}{m} \text{Binom}(m, L_D(h))
\]

\[
\approx \mathcal{N} \left( L_D(h), \sqrt{\frac{L_D(h)(1-L_D(h))}{m}} \right)
\]

- Hoeffding Bound on trail of Binomial:
  \[
  \mathbb{P}_{Z \sim \text{Binom}(m,p)}[|Z - \mathbb{E}[Z]| > t] \leq 2e^{-t^2/m}
  \]

- Conclusion: with probability \( \geq 1 - \delta \),
  \[
  |L_D(h) - L_S(h)| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2m}}
  \]
Empirical Risk Minimization

\[ ERM_{\mathcal{H}}(S) = \hat{h} = \arg\min_{h \in \mathcal{H}} L_S(h) \]

- Can we conclude that w.p. \( \geq 1 - \delta \),

\[ |L_D(\hat{h}) - L_S(\hat{h})| \leq \sqrt{\frac{\log 2/\delta}{2m}} \]?
Uniform Convergence and the Union Bound

- For each $h$ we have:
  \[ \mathbb{P}_S(|L_S(h) - L_D(h)| \geq t) \leq 2e^{-t^2/m} \]
- And so:
  \[ \mathbb{P}_S(\exists h \in \mathcal{H} \mid |L_S(h) - L_D(h)| \geq t) \leq \sum_{h \in \mathcal{H}} \mathbb{P}_S(|L_S(h) - L_D(h)| \geq t) \leq |\mathcal{H}| \cdot 2e^{-t^2/m} \]

- **Theorem:** For any hypothesis class $\mathcal{H}$ and any $\mathcal{D}$,
  \[ \mathbb{P}_{S \sim D^m} \left( \forall h \in \mathcal{H}, |L_D(h) - L_S(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log^2/\delta}{2m}} \right) \geq 1 - \delta \]
- Another way to view the derivation:
  \[ \mathbb{P}_S \left[ |L_S(h) - L_D(h)| \geq \sqrt{\frac{\log^2/\delta}{2m}} \right] \leq \delta \overset{\text{def}}{=} \frac{\delta'}{|\mathcal{H}|} \]
  And then \( \log 2/\delta = \log 2|\mathcal{H}|/\delta' = \log |\mathcal{H}| + \log 2/\delta \)
Empirical Risk Minimization

• Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall \delta \sim_{\mathcal{D}}^m$,

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{S}}(\hat{h}) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}}$$

• Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall \delta \sim_{\mathcal{D}}^m$,

$$L_{\mathcal{D}}(\hat{h}) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2 \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}}$$

Proof: if indeed $\forall h \in \mathcal{H}$, $|L_{\mathcal{D}}(h) - L_{\mathcal{S}}(h)| \leq \sqrt{\cdots}$, then:

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{S}}(\hat{h}) + \sqrt{\cdots} \leq L_{\mathcal{S}}(h^\ast) + \sqrt{\cdots} \leq L_{\mathcal{D}}(h^\ast) + \sqrt{\cdots} + \sqrt{\cdots}$$

$\hat{h}$ minimizes of $L_{\mathcal{S}}$ and so

$L_{\mathcal{S}}(\hat{h}) \leq L_{\mathcal{S}}(h)$ for any $h$, including $h^\ast$
Post-Hoc Guarantee

• Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall S \sim \mathcal{D}^m$, 

$$L_\mathcal{D}(\hat{h}) \leq L_S(\hat{h}) + \sqrt{\frac{\log |\mathcal{H}| + \log 2/\delta}{2m}}$$

• Performance Guarantee: Without ANY assumptions about $\mathcal{D}$ (i.e. about reality), if we somehow find a predictor $h$ with low $L_S(h)$, we can be ensured, (with high probability) that it will perform well on future examples.

• Instead, use independent test set $S'$ (e.g. split available examples into a training set $S$ and test set $S'$). From Hoeffding:

$$L_\mathcal{D}(A(S)) \leq L_{S'}(A(S)) + \sqrt{\frac{\log 1/\delta}{2|S'|}}$$

Random, but depends only on $S$, independent of $S'$

• Even better using tighter Binomial tail bounds, or even better numerically with Gaussian approximation of Binomial or entropy-based bound [see homework]
A-Priori Learning Guarantee

- Theorem: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall_{S \sim D}^{\delta} m,$

$$L_D(\hat{h}) \leq \inf_{h \in \mathcal{H}} L_D(h) + 2 \sqrt{\frac{\log |\mathcal{H}| + \log^2/\delta}{2m}}$$

- If we assume, based on our expert knowledge, that there exists a good predictor $h^* \in \mathcal{H}$, then with enough examples we can learn a predictor that’s almost as good, and we know how many examples we’ll need

- For any $\delta, \epsilon > 0$, using

$$m = 2 \frac{\log|\mathcal{H}| + \log^2/\delta}{\epsilon^2}$$

samples is enough to ensure $L_D(\hat{h}) \leq L_D(h^*) + \epsilon$ w.p. $\geq 1 - \delta$
Cardinality and Learning

• We saw:
  • All finite hypothesis classes are “learnable”—if we assume there is a good predictor in the class, with enough samples we’ll be able to learn it (fairly powerful: includes all 100-line programs)
  • Sample complexity $\propto \log|\mathcal{H}|$

• Is cardinality the only thing controlling learnability and sample complexity?
• Is this sample complexity bound always tight?
• Are all classes of the same cardinality equally complex?
• Are there infinite classes that learnable?
Probably Approximately Correct (PAC)

- Definition: A hypothesis class $\mathcal{H}$ is **PAC-Learnable** (in the realizable case) if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$ s.t. $L_\mathcal{D}(h) = 0$ for some $h \in \mathcal{H}$ (i.e. $\mathcal{D}$ is realizable by $\mathcal{H}$), $\forall \delta S \sim \mathcal{D} m(\epsilon, \delta)$,
  \[
  L_\mathcal{D}(A(S)) \leq \epsilon
  \]

- Definition: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}, \forall S \sim \mathcal{D} m(\epsilon, \delta)$,
  \[
  L_\mathcal{D}(A(S)) \leq \inf_{h \in \mathcal{H}} L_\mathcal{D}(h) + \epsilon
  \]
Probably Approximately Correct (PAC)

- Definition: A hypothesis class $\mathcal{H}$ is **PAC-Learnable** (in the realizable case) if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$ s.t. $L_\mathcal{D}(h) = 0$ for some $h \in \mathcal{H}$ (i.e. $\mathcal{D}$ is realizable by $\mathcal{H}$), $\forall \delta S \sim \mathcal{D} m(\epsilon, \delta)$,

  $$L_\mathcal{D}(A(S)) \leq \epsilon$$

- Definition: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule $A$ such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall \delta S \sim \mathcal{D} m(\epsilon, \delta)$,

  $$L_\mathcal{D}(A(S)) \leq \inf_{h \in \mathcal{H}} L_\mathcal{D}(h) + \epsilon$$

- Sample complexity of a learning rule:
  $$m_{A,\mathcal{H}}(\epsilon, \delta) = \min m \text{ s.t. } \forall \mathcal{D}, \forall \delta S \sim \mathcal{D} m(\epsilon, \delta), L_\mathcal{D}(A(S)) \leq \inf_{h \in \mathcal{H}} L_\mathcal{D}(h) + \epsilon$$

- Sample complexity for learning a hypothesis class:
  $$m_{\mathcal{H}}(\epsilon, \delta) = \min_{A} m_{A,\mathcal{H}}(\epsilon, \delta)$$
Cardinality and Sample Complexity

• We saw:
  • All finite hypothesis classes are PAC learnable
  • \( m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\text{ERM},\mathcal{H}}(\varepsilon, \delta) \leq O\left(\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{\varepsilon^2}\right)\)

• Are there infinite classes that are PAC learnable?
• Is the bound on \( m_{\mathcal{H}} \) always tight? Can \( m_{\mathcal{H}} \) be smaller than \( \log|\mathcal{H}| \)?
• Are all classes of the same cardinality equally complex?

E.g.
• \( \mathcal{X} = \{1, \ldots, 100\}, \mathcal{H} = \{\pm 1\}^\mathcal{X} \)
• \( \mathcal{X} = \{1, \ldots, 2^{100} \approx 10^{30}\}, \mathcal{H} = \{ [[x \leq \theta]] \mid \theta \in 1 \ldots 2^{100} \} \)
The Growth Function

- For $C = (x_1, x_2, ..., x_m) \in X^m$:
  $\Gamma_H(C) = \left| \{ (h(x_1), h(x_2), ..., h(x_m)) \in \{\pm 1 \}^m \mid h \in H \} \right|$

- $\Gamma_H(m) = \max_{C \in X^m} \Gamma_H(C)$

E.g.

- $X = \{1, ..., 100\}, H = \{\pm 1 \}^X$
  $\Gamma(m) = \min(2^m, 2^{100})$

- $X = \{1, ..., 2^{100} \approx 10^{30}\}, H = \{ [[x \leq \theta]] \mid \theta \in 1 \ldots 2^{100} \}$
  $\Gamma(m) = \min(m + 1, 2^{100})$
Uniform Convergence using the Growth Function

• Theorem: For any hypothesis class $\mathcal{H}$ and any $\mathcal{D}$,
  \[
  \mathbb{P}_{S \sim \mathcal{D}^m} \left[ \forall h \in \mathcal{H}, |L_\mathcal{D}(h) - L_S(h)| \leq 4 \sqrt{\frac{\log|\Gamma_\mathcal{H}(2m)| + \log^2/\delta}{m}} \right] \geq 1 - \delta
  \]
  Proof: homework

• Conclusion: For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall S \sim \mathcal{D}^m$,
  \[
  L_\mathcal{D}(\hat{h}) \leq L_S(\hat{h}) + 4 \sqrt{\frac{\log|\Gamma_\mathcal{H}(2m)| + \log^2/\delta}{m}}
  \]
  and
  \[
  L_\mathcal{D}(\hat{h}) \leq \inf_{h \in \mathcal{H}} L_\mathcal{D}(h) + 8 \sqrt{\frac{\log|\Gamma_\mathcal{H}(2m)| + \log^2/\delta}{m}}
  \]
Shattering and VC Dimension

- $C = \{x_1, \ldots, x_m\}$ is **shattered** by $\mathcal{H}$ if $\Gamma_{\mathcal{H}}(C) = 2^m$, i.e. we can get all $2^m$ behaviors:

$$\forall y_1, \ldots, y_m \in \pm 1, \exists h \in \mathcal{H} \text{ s.t. } \forall i \ h(x_i) = y_i$$

- The VC-dimension of $\mathcal{H}$ is the largest number of points that can be shattered by $\mathcal{H}$:

$$\text{VCdim}(\mathcal{H}) = \max m \text{ s.t. } \Gamma_{\mathcal{H}}(m) = 2^m$$

- If $\mathcal{H}$ is infinite and $\forall m \Gamma_{\mathcal{H}}(m) = 2^m$, then $\text{VCdim}(\mathcal{H}) = \infty$
VC Dimension: Examples

- $\mathcal{X} = \{1, \ldots, 100\}, \mathcal{H} = \{\pm 1\}^\mathcal{X}$
  - $\text{VCdim} = 100$
- Discrete Threshold: $\mathcal{X} = \{1, \ldots, 2^{100} \approx 10^{30}\}, \mathcal{H} = \{ [x \leq \theta] \mid \theta \in 1 \ldots 2^{100}\}$
  - $\text{VCdim} = 1$
- Continuous Thresholds: $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{ h_\theta(x) = [x < \theta] \mid \theta \in \mathbb{R}\}$
  - Only one point can be shattered; $\text{VCdim} = 1$
- Intervals: $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{ h_{a,b}(x) = [a \leq x \leq b] \mid a, b \in \mathbb{R}\}$
  - Can shatter any two points
  - With three points, can’t realize $+ - +$
- Axis aligned rectangles
  - Can shatter 1, 2, 3 points
  - Some sets of 4 points can’t be shattered—is this a problem?
  - Some sets of 4 points can be shattered
  - Can’t shatter 5 points
Sauer-Shelah-VC Lemma

- If $VCdim(\mathcal{H}) = D$, then:

$$\Gamma_\mathcal{H}(m) \leq \sum_{i=0}^{D} \binom{m}{i} \leq \left(\frac{em}{D}\right)^D$$

for $m > D$
Sauer-Shelah-VC Lemma

- If $\text{VCdim}(\mathcal{H}) = D$, then:
  \[ \Gamma_{\mathcal{H}}(m) \leq \sum_{i=0}^{D} \binom{m}{i} \leq \left( \frac{em}{D} \right)^D \]
  for $m > D$
Conclusion: VC Learning Guarantees

• Recall:

\[
\forall_{S \sim \mathcal{D}^m}, \quad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + 4 \sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log^2/\delta}{m}}
\]

From Sauer, \( \log|\Gamma_{\mathcal{H}}(2m)| \leq \log \left( \frac{e m}{\text{VCdim}} \right)^{\text{VCdim}} \leq O(\text{VCdim} \cdot \log m) \).

We therefore have:

\[
\forall_{S \sim \mathcal{D}^m}, \quad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O \left( \sqrt{\frac{\text{VCdim}(\mathcal{H}) \log m + \log^{1/\delta}}{m}} \right)
\]

With a very complex proof, this can be improved to:

\[
\forall_{S \sim \mathcal{D}^m}, \quad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O \left( \sqrt{\frac{\text{VCdim}(\mathcal{H}) + \log^{1/\delta}}{m}} \right)
\]
VC Learning Guarantees

- **Conclusion:** If $\text{VCdim}(\mathcal{H}) < \infty$ then $\mathcal{H}$ is **agnostically PAC learnable** using $\text{ERM}_{\mathcal{H}}$

  $\forall_{S \sim \mathcal{D}}^\delta m, \quad L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O\left(\sqrt{\frac{\text{VCdim}(\mathcal{H}) + \log \frac{1}{\delta}}{m}}\right)$

  The sample complexity is bounded by:

  $$m(\epsilon, \delta) \leq O\left(\frac{\text{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right)$$

- **Finite classes are PAC-learnable,** with $m_{\text{ERM},\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\log|\mathcal{H}| + \log^1/\delta}{\epsilon^2}\right)$

- **VC classes are PAC-learnable,** with $m_{\text{ERM},\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\text{VCdim}(\mathcal{H}) + \log^1/\delta}{\epsilon^2}\right)$

- **Can a class with infinite VC-dimension be learnable?**
  Can the sample complexity be lower then the VC dimension?
VC-Dimension: More Examples

- Circles in $\mathbb{R}^2$: $\mathcal{H} = \{ h_{c,r}(x) = [||x - c|| \leq r] \mid c \in \mathbb{R}^2, r \in \mathbb{R} \}$
  - Can shatter 3 points

- Circles and their complement
  - Can shatter 4 points

- Circles around origin: $\mathcal{H} = \{ h_{c,r}(x) = [||x|| \leq r] \mid r \in \mathbb{R} \}$
  - Can shatter only 1 point

- Axis aligned ellipses:
  $$\mathcal{H} = \left\{ h_{c,r[1],r[2]}(x) = \left[ \frac{(x[1] - c[1])^2}{r[1]^2} + \frac{(x[2] - c[2])^2}{r[2]^2} \leq 1 \right] \mid c \in \mathbb{R}^2, r[1], r[2] \in \mathbb{R} \right\}$$
  - Can shatter 4 points

- General ellipses
  - Can shatter 5 points

- Upper bounds?