Classification: The PAC Learning Framework

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder

LECTURE 5

Slides adapted from Eli Upfal
What does it mean to learn something?

- What are the things that we’re learning?
- What does it mean to be learnable?
- Provides a framework for reasoning about what we can *theoretically* learn
What does it mean to learn something?

• What are the things that we’re learning?
• What does it mean to be learnable?
• Provides a framework for reasoning about what we can theoretically learn
  ◦ Sometime theoretically learnable things are very difficult
  ◦ Sometimes things that should be hard actually work
Example

- Californian just moved to Colorado
- When is it “nice” outside?
- Has a perfect thermometer, but natives call 50F (10C) “nice”
Example

- Californian just moved to Colorado
- When is it “nice” outside?
- Has a perfect thermometer, but natives call 50°F (10°C) “nice”
- Each temperature is an observation $x$
- Coloradan concept of “nice” $c(x)$
- Californian wants to learn hypothesis $h(x)$ close to $c(x)$
Example

- Californian just moved to Colorado
- When is it “nice” outside?
- Has a perfect thermometer, but natives call 50F (10C) “nice”
- Each temperature is an observation \( x \)
- Coloradan concept of “nice” \( c(x) \)
- Californian wants to learn hypothesis \( h(x) \) close to \( c(x) \)

Generalization error

\[
R(h) = \Pr_{x \sim D} [h(x) \neq c(x)] = \mathbb{E}_{x \sim D} [\mathbf{1} [h(x) \neq c(x)]] \tag{1}
\]
Example

• Californian just moved to Colorado
• When is it “nice” outside?
• Has a perfect thermometer, but natives call 50F (10C) “nice”
• Each temperature is an observation \( x \)
• Coloradan concept of “nice” \( c(x) \)
• Californian wants to learn hypothesis \( h(x) \) close to \( c(x) \)

Generalization error

\[
R(h) = \Pr_{x \sim D} [h(x) \neq c(x)] = \mathbb{E}_{X \sim D} [\mathbb{1} [h(x) \neq c(x)]]
\]  (1)
The Californian gets \( n \) random examples.
The Californian gets \( n \) random examples.
The Californian gets $n$ random examples.
The best rule that conforms with the examples is \([a, b]\).
Let \([c, d]\) be the correct (unknown) rule. Let \(\Delta\) be the gap between. The probability of being wrong is the probability that \(n\) samples missed \(\Delta\).
PAC-learning definition

**Definition**

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $\mathcal{A}$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m} [ R(h_S) \leq \epsilon ] \geq 1 - \delta$$

(2)

for any sample size $m \geq f \left( \frac{1}{\epsilon}, \frac{1}{\delta}, n, |c| \right)$
PAC-learning definition

**Definition**

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $\mathcal{A}$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$

(2)

for any sample size $m \geq f\left(\frac{1}{\epsilon}, \frac{1}{\delta}, n, |c|\right)$

The sample we learn from
PAC-learning definition

**Definition**

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $A$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$

(2)

for any sample size $m \geq f\left(\frac{1}{\epsilon}, \frac{1}{\delta}, n, |c|\right)$

The data distribution the sample comes from
PAC-learning definition

**Definition**

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $\mathcal{A}$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$

(2)

for any sample size $m \geq f\left(\frac{1}{\epsilon}, \frac{1}{\delta}, n, |c|\right)$

The hypothesis we learn
PAC-learning definition

**Definition**

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $\mathcal{A}$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$

(2)

for any sample size $m \geq f \left( \frac{1}{\epsilon}, \frac{1}{\delta}, n, |c| \right)$

Generalization error
PAC-learning definition

**Definition**

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $\mathcal{A}$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$  \hspace{1cm} (2)

for any sample size $m \geq f \left( \frac{1}{\epsilon}, \frac{1}{\delta}, n, |c| \right)$

Our bound on the generalization error (e.g., we want it to be better than 0.1)
PAC-learning definition

Definition

**PAC-learnable** A concept $C$ is PAC-learnable if $\exists$ algorithm $\mathcal{A}$ and a polynomial function $f$ such that for any $\epsilon$ and $\delta$, $\forall D(X)$ and $c \in C$

$$\Pr_{S \sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

(2)

for any sample size $m \geq f \left(\frac{1}{\epsilon, \frac{1}{\delta}}, n, |c|\right)$

The probability of learning a hypothesis with error greater than $\epsilon$ (e.g., 0.05)
Is a Californian learning temperature PAC learnable?

- The only way for the bad event to happen is if a point lands in $\Delta$

$$\Pr\left[ x_1 \notin \Delta \land \cdots \land x_m \notin \Delta \right] = \prod_{i}^{m} \Pr\left[ x_i \notin \Delta \right]$$  \hspace{1cm} (3)

- We want the probability of a point landing there to be less than $\varepsilon$

$$\Pr\left[ x_1 \notin \Delta \land \cdots \land x_m \notin \Delta \right] = (1 - \varepsilon)^m \leq e^{-\varepsilon m}$$  \hspace{1cm} (4)
Is a Californian learning temperature PAC learnable?

- The only way for the bad event to happen is if a point lands in $\Delta$

$$\Pr [x_1 \not\in \Delta \land \cdots \land x_m \not\in \Delta] = \prod_{i}^{m} \Pr [x_i \not\in \Delta]$$  \hspace{1cm} (3)

**Independence!**

- We want the probability of a point landing there to be less than $\epsilon$

$$\Pr [x_1 \not\in \Delta \land \cdots \land x_m \not\in \Delta] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$  \hspace{1cm} (4)
Is a Californian learning temperature PAC learnable?

• The only way for the bad event to happen is if a point lands in $\Delta$

$$\Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = \prod_{i}^{m} \Pr [x_i \notin \Delta]$$

(3)

• We want the probability of a point landing there to be less than $\epsilon$

$$\Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$

(4)

Useful inequality: $1 + x \leq e^x$

Graph for $1+x$, $e^x$
Is a Californian learning temperature PAC learnable?

• The only way for the bad event to happen is if a point lands in \( \Delta \)

\[
\Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = \prod_{i}^{m} \Pr [x_i \notin \Delta] \tag{3}
\]

• We want the probability of a point landing there to be less than \( \varepsilon \)

\[
\Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = (1 - \varepsilon)^m \leq e^{-\varepsilon m} \tag{4}
\]

• We want the generalization to violate \( \varepsilon \) less than \( \delta \), solving for \( m \):

\[
\Pr [R(h) \geq \varepsilon] \leq 1 - \delta \tag{5}
\]

\[
e^{-\varepsilon m} \leq \delta \tag{6}
\]

\[
-\varepsilon m \leq \ln \delta \tag{7}
\]

\[
\frac{1}{\varepsilon} \frac{1}{-\ln \delta} \leq m \tag{8}
\]
Is a Californian learning temperature PAC learnable?

- The only way for the bad event to happen is if a point lands in $\Delta$
  \[ \Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = \prod_{i}^{m} \Pr [x_i \notin \Delta] \]  
  (3)

- We want the probability of a point landing there to be less than $\epsilon$
  \[ \Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = (1 - \epsilon)^m \leq e^{-\epsilon m} \]  
  (4)

- We want the generalization to violate $\epsilon$ less than $\delta$, solving for $m$:
  \[ \Pr [R(h) \geq \epsilon] \leq 1 - \delta \]  
  (5)
  \[ e^{-\epsilon m} \leq \delta \]  
  (6)
  \[ -\epsilon m \leq \ln \delta \]  
  (7)
  \[ \frac{1}{\epsilon} \ln \frac{1}{\delta} \leq m \]  
  (8)
Is a Californian learning temperature PAC learnable?

• The only way for the bad event to happen is if a point lands in $\Delta$

$$\Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = \prod_{i=1}^{m} \Pr [x_i \notin \Delta]$$

(3)

• We want the probability of a point landing there to be less than $\varepsilon$

$$\Pr [x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = (1 - \varepsilon)^m \leq e^{-\varepsilon m}$$

(4)

• We want the generalization to violate $\varepsilon$ less than $\delta$, solving for $m$:

$$\Pr [R(h) \geq \varepsilon] \leq 1 - \delta$$

(5)

$$e^{-\varepsilon m} \leq \delta$$

(6)  

Take log of both sides

$$-\varepsilon m \leq \ln \delta$$

(7)

$$\frac{1}{\varepsilon} \frac{1}{\delta} \leq m$$

(8)
Is a Californian learning temperature PAC learnable?

- The only way for the bad event to happen is if a point lands in $\Delta$

$$\Pr[x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = \prod_{i}^{m} \Pr[x_i \notin \Delta]$$  \hspace{1cm} (3)

- We want the probability of a point landing there to be less than $\epsilon$

$$\Pr[x_1 \notin \Delta \land \cdots \land x_m \notin \Delta] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$  \hspace{1cm} (4)

- We want the generalization to violate $\epsilon$ less than $\delta$, solving for $m$:

$$\Pr[R(h) \geq \epsilon] \leq 1 - \delta$$ \hspace{1cm} (5)

$$e^{-\epsilon m} \leq \delta$$ \hspace{1cm} (6) Direction of inequality flips when you divide by $-m$

$$-\epsilon m \leq \ln \delta$$ \hspace{1cm} (7)

$$\frac{1}{\epsilon} \frac{1}{-\ln \delta} \leq m$$ \hspace{1cm} (8)
Consistent Hypotheses, Finite Spaces

- Possible to prove that specific problems are learnable (and we will!)
- Can we do something more general?
- Yes, for \textbf{finite} hypothesis spaces \( c \in H \)
- That are also consistent with training data

**Theorem**

\textit{Learning bounds for finite }\( H \), consistent \ Let \( H \) be a finite set of functions mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). Let \( \mathcal{A} \) be an algorithm that for a iid sample \( S \) returns a consistent hypothesis (training error \( \hat{R}(h) = 0 \)), then for any \( \epsilon, \delta > 0 \), the concept is PAC learnable with samples

\[
m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right)
\]  

(9)
Proof: Setup

We want to bound the probability that some \( h \in H \) is consistent and has error more than \( \varepsilon \).

\[
\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \varepsilon \right] \\
= \Pr \left[ \left( h_1 \in H \land \hat{R}(h_1) = 0 \land R(h_1) > \varepsilon \right) \lor \cdots \lor \left( h_i \in H \land \hat{R}(h_i) = 0 \land R(h_i) > \varepsilon \right) \right] \\
\leq \sum_h \Pr \left[ \hat{R}(h) = 0 \land R(h) > \varepsilon \right] \\
\leq \sum_h \Pr \left[ \hat{R}(h) = 0 \mid R(h) > \varepsilon \right]
\]
Proof: Setup

We want to bound the probability that some $h \in H$ is consistent and has error more than $\epsilon$.

$$\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] = \Pr \left[ \left( h_1 \in H \land \hat{R}(h_1) = 0 \land R(h_1) > \epsilon \right) \lor \cdots \lor \left( h_i \in H \land \hat{R}(h_i) = 0 \land R(h_i) > \epsilon \right) \right] \leq \sum_h \Pr \left[ \hat{R}(h) = 0 \land R(h) > \epsilon \right] \leq \sum_h \Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right]$$
Proof: Setup

We want to bound the probability that some $h \in H$ is consistent and has error more than $\epsilon$.

\[
\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] \tag{10}
\]
\[
= \Pr \left[ \left( h_1 \in H \land \hat{R}(h_1) = 0 \land R(h_1) > \epsilon \right) \lor \cdots \lor \left( h_i \in H \land \hat{R}(h_i) = 0 \land R(h_i) > \epsilon \right) \right]
\]
\[
\leq \sum_h \Pr \left[ \hat{R}(h) = 0 \land R(h) > \epsilon \right] \tag{11}
\]
\[
\leq \sum_h \Pr \left[ \hat{R}(h) = 0 | R(h) > \epsilon \right] \tag{12}
\]

Union bound
Proof: Setup

We want to bound the probability that some $h \in H$ is consistent and has error more than $\epsilon$.

\[
\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] = \Pr \left[ \left( h_1 \in H \land \hat{R}(h_1) = 0 \land R(h_1) > \epsilon \right) \lor \cdots \lor \left( h_i \in H \land \hat{R}(h_i) = 0 \land R(h_i) > \epsilon \right) \right] 
\leq \sum_h \Pr \left[ \hat{R}(h) = 0 \land R(h) > \epsilon \right] 
\leq \sum_h \Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right]
\]

Definition of conditional probability
The generalization error is greater than $\epsilon$, so we bound probability of no inconsistent points in training for a single hypothesis $h$.

$$\Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right] \leq (1 - \epsilon)^m \quad (13)$$
Proof: Connection back to interval learning

The generalization error is greater than $\varepsilon$, so we bound probability of no inconsistent points in training for a single hypothesis $h$.

$$\Pr \left[ \hat{R}(h) = 0 \mid R(h) > \varepsilon \right] \leq (1 - \varepsilon)^m \quad (13)$$

but this must be true of all of the hypotheses in $H$,

$$\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \varepsilon \right] \leq |H|(1 - \varepsilon)^m \quad (14)$$
Proof: Connection back to interval learning

The generalization error is greater than $\epsilon$, so we bound probability of no inconsistent points in training for a single hypothesis $h$.

$$\Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right] \leq (1 - \epsilon)^m \quad (13)$$

but this must be true of all of the hypotheses in $H$,

$$\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] \leq |H|(1 - \epsilon)^m \quad (14)$$

$$|H|(1 - \epsilon)^m \leq |H|e^{-m\epsilon} = \delta \quad \text{we set the RHS to be equal to } \delta$$
Proof: Connection back to interval learning

The generalization error is greater than $\epsilon$, so we bound probability of no inconsistent points in training for a single hypothesis $h$.

$$\Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right] \leq (1 - \epsilon)^m \quad (13)$$

but this must be true of all of the hypotheses in $H$,

$$\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] \leq |H|(1 - \epsilon)^m \quad (14)$$

$$|H|(1 - \epsilon)^m \leq |H|e^{-m\epsilon} = \delta$$

$$\ln \delta = \ln |H| - m\epsilon$$

apply log to both sides
Proof: Connection back to interval learning

The generalization error is greater than \( \epsilon \), so we bound probability of no inconsistent points in training for a single hypothesis \( h \).

\[
\Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right] \leq (1 - \epsilon)^m \tag{13}
\]

but this must be true of all of the hypotheses in \( H \),

\[
\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] \leq |H|(1 - \epsilon)^m \tag{14}
\]

\[
|H|(1 - \epsilon)^m \leq |H|e^{-m\epsilon} = \delta
\]

move \( \ln |H| \) to the other side, and rewrite \( \ln \delta = -0 - (-\ln \delta) = -1(\ln 1 - \ln \delta) = -\ln \left( \frac{1}{\delta} \right) \):

\[
\ln \delta = \ln |H| - m\epsilon
\]

\[
-\ln \frac{1}{\delta} - \ln |H| = -m\epsilon
\]
Proof: Connection back to interval learning

The generalization error is greater than $\epsilon$, so we bound probability of no inconsistent points in training for a single hypothesis $h$.

$$\Pr \left[ \hat{R}(h) = 0 \mid R(h) > \epsilon \right] \leq (1 - \epsilon)^m \quad (13)$$

but this must be true of all of the hypotheses in $H$,

$$\Pr \left[ \exists h \in H : \hat{R}(h) = 0 \land R(h) > \epsilon \right] \leq |H|(1 - \epsilon)^m \quad (14)$$

$$|H|(1 - \epsilon)^m \leq |H|e^{-me} = \delta$$

$$\ln \delta = \ln |H| - me$$

$$- \ln \frac{1}{\delta} - \ln |H| = - me$$

Divide by $-\epsilon$

$$- \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right) = m$$
But what does it all mean?

\[ m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]  \hspace{1cm} (15)

- Confidence
- Complexity
But what does it all mean?

\[ m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]  

(15)

- **Confidence**: More certainty means more training data
- **Complexity**
But what does it all mean?

\[ m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]  

(15)

- **Confidence**: More certainty means more training data
- **Complexity**: More complicated hypotheses need more training data
But what does it all mean?

\[
m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right)
\]  

(15)

- **Confidence**: More certainty means more training data
- **Complexity**: More complicated hypotheses need more training data

**Scary Question**

What’s $|H|$ for logistic regression?
What’s next . . .

- In class: examples of PAC learnability
- Next time: how to deal with infinite hypothesis spaces
What’s next . . .

- In class: examples of PAC learnability
- Next time: how to deal with infinite hypothesis spaces
- Takeaway
  - Even though we can’t prove anything about logistic regression, it still works
  - However, using the theory will lead us to a better classification technique: support vector machines