Machine Learning: Chenhao Tan

University of Colorado Boulder

LECTURE 26

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Logistics

- Project mid-point check-in
- HW5
- Prelim 3

Learning objectives

- Learn about basics of learning theory.
- Prove some simple bounds on errors and sample sizes.
- Gain some intuition about complexity and overfitting.

Outline

A motivating example

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Alien has a perfect alien thermometer
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have concept $c(x)$ of *nice*
- Alien wants to learn hypothesis $h(x)$

What does it mean that Alien has learned?

A motivating example

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Alien has a perfect alien thermometer
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have concept $c(x)$ of *nice*
- Alien wants to learn hypothesis $h(x)$

How many locals does he need to ask to get $h(x)$ that is 99% accurate?

A motivating example

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Alien has a perfect alien thermometer
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have concept $c(x)$ of *nice*
- Alien wants to learn hypothesis $h(x)$

How many locals does he need to ask to get $h(x)$ that is 99% accurate about 99% of the time?

PAC learnability

Assumptions:

- Data comes from distribution $\mathcal{D}$
- Concept $c : X \to Y$ comes from concept class C
- Hypothesis $h : X \to Y$ comes from hypothesis class H

Generalization Error

$$R(h) = Pr_{x\sim D}\left[h(x) \neq c(x)\right] = E_{x\sim D}[\, I[h(x) \neq c(x)]\,]$$

Goal: Given a set of data S of size m, can we learn a hypothesis h that we can say is **accurate** with high **confidence**?

PAC learnability

We say that a concept is PAC-Learnable if we can find a hypothesis that is **P**robably **A**pproximately **C**orrect using a training set S of size m where m isn't too large

$$R(h_S) \leq \epsilon$$

- Approximately correct: Accuracy is $1 - \epsilon$

PAC learnability

We say that a concept is PAC-Learnable if we can find a hypothesis that is **P**robably **A**pproximately **C**orrect using a training set S of size m where m isn't too large

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

- Approximately correct: Accuracy is $1 - \epsilon$
- Probably: Confidence in hypothesis is $1 - \delta$

PAC = Probably Approximately Correct

PAC learnability

PAC Learnability

A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$

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

for any $c \in C$ and any distribution $\mathcal{D}$ for any sample size $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$.

PAC learnability

PAC Learnability

A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$

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

for any $c \in C$ and any distribution $\mathcal{D}$ for any sample size $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$.

- S: The training set we learn from
- $\mathcal{D}$: The distribution the data comes from
- h_S: The hypothesis we learn from training set

PAC learnability

PAC Learnability

A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$

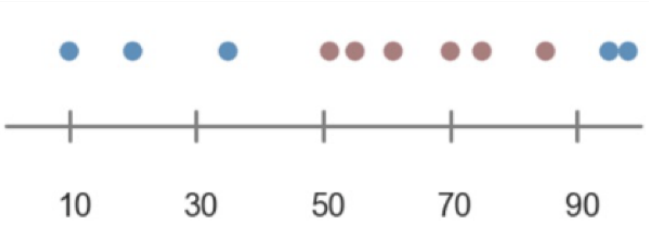
$$Pr_{S \sim \mathcal{D}^m}\left[R(h_S) \leq \epsilon\right] \geq 1 - \delta$$

for any $c \in C$ and any distribution $\mathcal{D}$ for any sample size $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$.

- $R(h_S)$: The generalization error of h_S
- $1 - \epsilon$: The accuracy of h_s
- $1 - \delta$: The confidence the accuracy $1 - \epsilon$ is realized

Outline

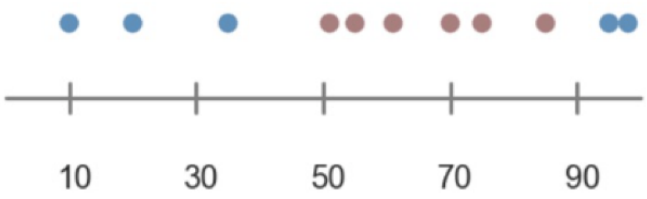
The alien example



- Concept class $C =$ Intervals on Real Line
- Hypothesis class $H =$ Intervals on Real Line

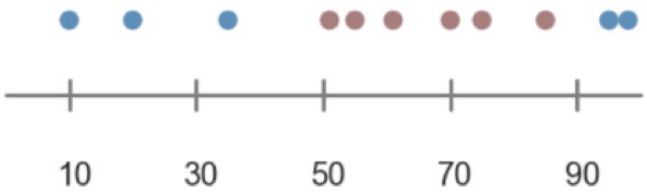
Want to obtain bound on training examples needed to satisfy PAC.

The alien example



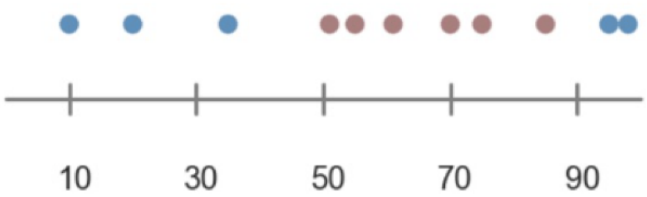
What is Algorithm $\mathcal{A}$?

The alien example



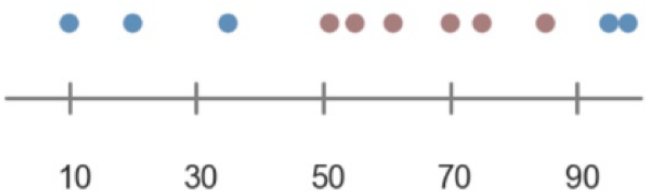
What is Algorithm $\mathcal{A}$?
Set hypothesis to smallest interval containing S: $h_s = [a, b]$.

The alien example



What is Algorithm $\mathcal{A}$?
Set hypothesis to smallest interval containing S: $h_s = [a, b]$.
Errors happen if a positive point falls outside of $h_s = [a, b]$.

The alien example



What is Algorithm $\mathcal{A}$?
Set hypothesis to smallest interval containing S: $h_s = [a, b]$.
Errors happen if a positive point falls outside of $h_s = [a, b]$.
Suppose true concept is $c = [c, d]$.

The alien example

Want to define relationship between ϵ, δ, and m such that

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

The alien example

Want to define relationship between ϵ, δ, and m such that

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

Easier to prove things about the contrapositive statement.

The alien example

Want to define relationship between ϵ, δ, and m such that

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

Easier to prove things about the contrapositive statement.

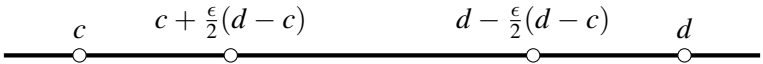
$$\begin{aligned}
&Pr_{S\sim\mathcal{D}^m}\left[R(h_S) \leq \epsilon\right] \geq 1-\delta \\
&\Leftrightarrow 1 - Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] \geq 1-\delta \\
&\Leftrightarrow -Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] \geq -\delta \\
&\Leftrightarrow Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] \leq \delta
\end{aligned}$$

So instead we'll try to prove something about

$$Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] \leq \delta.$$

The alien example

We want to bound the probability that the generalization error h_S is greater than ϵ. This is the probability that despite the fact that the true concept was $c = [c, d]$, we didn't observe any points in $[c, a]$ or $[b, d]$.



$L = [c, c + \frac{\epsilon}{2}(d - c)], R = [d - \frac{\epsilon}{2}(d - c), d]$

$$\{S|R(h_S) \leq \epsilon\} \supseteq \{\exists x_i \text{ in } L \text{ and } \exists x_i \text{ in } R\}$$
$$\{S|R(h_S) > \epsilon\} \subseteq \{\text{no } x_i \text{ in } L \text{ or no } x_i \text{ in } R\}$$

Useful Fact 1: Union Bound

$$Pr[A \cup B] \leq Pr[A] + Pr[B]$$

The alien example

$$\begin{aligned} Pr[R(h_S) > \epsilon] &\leq Pr[\text{no } x_i \text{ in } L \text{ or } R] \\ &\leq Pr[\text{no } x_i \text{ in } L] + Pr[\text{no } x_i \text{ in } R] \end{aligned}$$

$$\begin{aligned} Pr[\text{ no } x_i \text{ in } L] &= Pr[\text{all } x_i \text{ not in } L] \\ &= \prod_{i=1}^{m} \left(1 - \frac{\epsilon}{2}\right) = \left(1 - \frac{\epsilon}{2}\right)^m \end{aligned}$$

$$\begin{aligned} Pr[R(h_S) > \epsilon] &\leq \left(1 - \frac{\epsilon}{2}\right)^m + \left(1 - \frac{\epsilon}{2}\right)^m \\ &= 2\left(1 - \frac{\epsilon}{2}\right)^m \end{aligned}$$

The alien example

$$
\begin{aligned}
Pr[R(h_S) > \epsilon] &\leq \left(1 - \frac{\epsilon}{2}\right)^m + \left(1 - \frac{\epsilon}{2}\right)^m \\
&= 2\left(1 - \frac{\epsilon}{2}\right)^m
\end{aligned}
$$

Useful Fact 2: For any $z \in \mathbb{R}$, $1 + z \leq e^z$

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &\leq \left(1 - \frac{\epsilon}{2}\right)^m + \left(1 - \frac{\epsilon}{2}\right)^m \\
&= 2\left(1 - \frac{\epsilon}{2}\right)^m \\
&\leq 2e^{-\epsilon m/2}
\end{aligned}
$$

The alien example

OK, we've bounded the probability that the generalization error for h_S is greater than ϵ. Then, for a fixed δ, we have

$$2e^{-\epsilon m/2} < \delta \quad \Leftrightarrow \quad \frac{-\epsilon m}{2} < \ln \frac{\delta}{2} \quad \Leftrightarrow \quad m > \frac{2}{\epsilon} \ln \frac{2}{\delta}$$

Punchline: For any choice of $\epsilon > 0$ and $\delta > 0$, hypothesis h_S is probably approximately correct if

$$m > \frac{2}{\epsilon} \ln \frac{2}{\delta}$$

The alien example

OK, we've bounded the probability that the generalization error for h_S is greater than ϵ. Then, for a fixed δ, we have

$$2e^{-\epsilon m/2} < \delta \quad \Leftrightarrow \quad \frac{-\epsilon m}{2} < \ln \frac{\delta}{2} \quad \Leftrightarrow \quad m > \frac{2}{\epsilon} \ln \frac{2}{\delta}$$

Example: Want 99% accuracy ($\epsilon = 0.01$) with 99% confidence ($\delta = 0.01$) then need

$$m > \frac{2}{.01} \ln \frac{2}{.01} \approx 1060 \text{ training examples}$$

Important: The lower bound on m is bounded above by a polynomial in $1/\epsilon$ and $1/\delta$, thus this problem is PAC Learnable.

Outline

General cases

OK, so we saw an example proving PAC learnability for a specific problem with specific hypothesis and specific algorithm.

Can we be more general than this?

General cases

OK, so we saw an example proving PAC learnability for a specific problem with specific hypothesis and specific algorithm.

Can we be more general than this?
Yes!

- Today, H is finite
- Next time, H is infinite

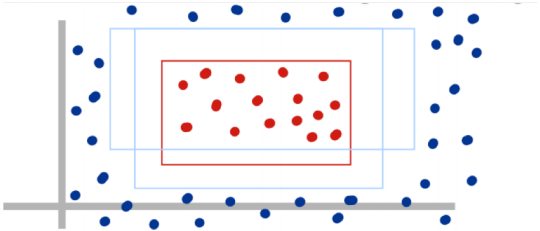
Further distinction

- H is finite and c is in H
- H is finite and c is not in H

General cases

We say that **Hypothesis Class** H is consistent if $c \in H$, that is, the concept that we're trying to learn is actually a valid hypothesis.

Example: c is the interval $[3, 7]$ and H is the consistent class of all intervals between 0 and 100 with integer endpoints.



General cases

We say that **Hypothesis Class** H is consistent if $c \in H$, that is, the concept that we're trying to learn is actually a valid hypothesis.

Example: c is the interval $[3, 7]$ and H is the consistent class of all intervals between 0 and 100 with integer endpoints.

Example: c is the interval $[3.5, 7.5]$ and H is the inconsistent class of all intervals between 0 and 100 with integer endpoints.

Question: What can you say about the training error $\hat{R}(h)$ if $h \in H$ is a consistent hypothesis?

General cases

We say that **a hypothesis** h is consistent if it admits no error on the training sample S_{train}, or in other words, $\hat{R}(h) = 0$

Example: Suppose c is the interior of an axis-aligned rectangle with integer vertices, and H is the set of all axis-aligned rectangles with integer vertices.

Finite consistent hypothesis class

Suppose our algorithm $\mathcal{A}$ can find a consistent hypothesis.
Theorem: Let H be a finite set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for an i.i.d. sample S returns a consistent hypothesis, then for any $\epsilon,\ \delta > 0$, the concept c is PAC Learnable with

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

Finite consistent hypothesis class

Example: Consider learning the concept class C_n of conjunctions of at most n Boolean literals $x_1, \ldots, x_n$.
A Boolean literal is either a variable x_i ($i \in [1, n]$) or it's negation $\bar{x}_i$.
For $n = 4$, an example of a conjunction we might try to learn is

$$x_1 \wedge \bar{x}_2 \wedge x_4$$

Positive Example: $(1, 0, 0, 1)$
Negative Example: $(1, 0, 0, 0)$

Finite consistent hypothesis class

We can now use our general error bound to find a bound on m. Note that $|H| = 3^n$ because for the i^{th} literal either x_i is present, $\bar{x}_i$ is present, or it's missing entirely. We then have for a given $\delta > 0$,

$$m \geq \frac{1}{\epsilon}\left(n \ln 3 + \ln \frac{1}{\delta}\right)$$

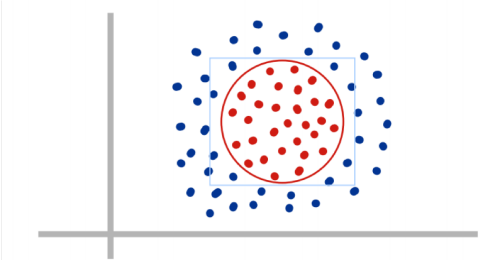
Example: If we want 90% accuracy ($\epsilon = 0.1$) with 98% confidence ($\delta = 0.02$) a length at most 10 conjunction would require $m \geq 156$ samples to learn.

Finite inconsistent hypothesis class

The more common case occurs when the true concept c does not occur in our hypothesis class H.

Finite inconsistent hypothesis class

The more common case occurs when the true concept c does not occur in our hypothesis class H.
Example: Hypothesis class H is axis aligned rectangles, but true concept is a circle.



Finite inconsistent hypothesis class

Theorem: Let H be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

Finite inconsistent hypothesis class

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

- Larger m is, better training error predicts generalization error

Finite inconsistent hypothesis class

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

- Larger m is, better training error predicts generalization error

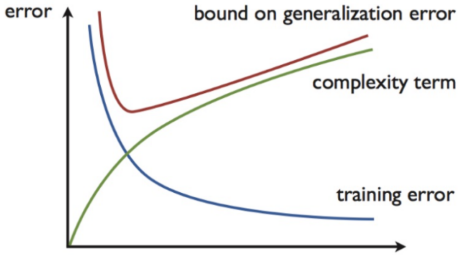
What about the case that we consider making H more complex?

- Training error would go down
- Bound term would go up ...
- Bias-variance trade-off

Finite inconsistent hypothesis class

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

- Larger m is, better training error predicts generalization error



Outline

Finite consistent hypothesis class

Suppose our algorithm $\mathcal{A}$ can find a consistent hypothesis
Theorem: Let H be a finite set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for an i.i.d. sample S returns a consistent hypothesis, then for any $\epsilon,\ \delta > 0$, the concept c is PAC Learnable with

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

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

Finite consistent hypothesis class

Proof: We want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] =$$
$$Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } ...$$
$$... \text{ or } (h_k \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)]$$

Probability of at least one of at least one of all consistent $h \in H$ having generalization error greater than ϵ

Finite consistent hypothesis class

Proof: We want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

$$
\begin{aligned}
Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] = \\
Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } ... \\
... \text{ or } (h_k \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)] \leq \\
\sum_h Pr[\ \hat{R}(h) = 0 \text{ and } R(h) > \epsilon]
\end{aligned}
$$

Using the Union Bound

Finite consistent hypothesis class

Proof: We want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

$$
\begin{aligned}
Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] = \\
Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } ... \\
... \text{ or } (h_k \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)] \leq \\
\sum_h Pr[\ \hat{R}(h) = 0 \text{ and } R(h) > \epsilon] \leq \\
\sum_h Pr[\ \hat{R}(h) = 0 \mid R(h) > \epsilon]
\end{aligned}
$$

Using the product rule and fact that $Pr[\ R(h) > \epsilon\ ] \leq 1$

Finite consistent hypothesis class

The generalization error is greater than ϵ, so we bound the probability that **no** inconsistent points in training set for a single hypothesis h as

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

Finite consistent hypothesis class

The generalization error is greater than ϵ, so we bound the probability that **no** inconsistent points in training set for a single hypothesis h as

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

But this must be true for all of the hypotheses in H, so

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \leq |H|(1 - \epsilon)^m$$

Finite consistent hypothesis class

The generalization error is greater than ϵ, so we bound the probability that **no** inconsistent points in training set for a single hypothesis h as

$$Pr[\ \hat{R}(h) = 0 \mid \text{ and } R(h) > \epsilon] \leq (1-\epsilon)^m$$

But this must be true for all of the hypotheses in H, so

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \leq |H|(1-\epsilon)^m$$

Using our exponential trick again

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \leq |H|e^{-m\epsilon}$$

Finite consistent hypothesis class

Have our bound on $Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right]$. Now for any $\delta > 0$

$$|H|e^{-m\epsilon} \leq \delta \quad \Leftrightarrow \quad \ln|H| - m\epsilon \leq \ln\delta$$

$$\Leftrightarrow \quad \ln|H| - \ln\delta \leq m\epsilon$$

$$\Leftrightarrow \quad \ln|H| + \ln\frac{1}{\delta} \leq m\epsilon$$

$$\Leftrightarrow \quad m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

Finite inconsistent hypothesis class

The more common case occurs when the true concept c does not occur in our hypothesis class H.
Example: Hypothesis class H is axis aligned rectangles, but true concept is a circle.
To handle this case we have to borrow a theorem of analysis
Theorem: Hoeffding's Inequality: Fix $\epsilon > 0$ and let S denote i.i.d. same of size m. Then, for any hypothesis $h : \mathcal{X} \to \{0, 1\}$, the following holds

$$Pr_{S \sim \mathcal{D}^m}[\ |\hat{R}(h) - R(h)| > \epsilon\ ] \leq 2\exp[-2m\epsilon^2]$$

Finite inconsistent hypothesis class

Setting $\delta = 2\exp[-2m\epsilon^2]$, solving for $\epsilon = \epsilon(\delta)$ and plugging back in yields, for a single hypothesis h

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

But this is just for a single h. We have
Theorem: Let H be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}}$$

Finite inconsistent hypothesis class

Proof: (Very similar to before). Let $h_1, \ldots, h_{|H|}$ be the elements of H. Then

$$\begin{aligned} Pr[\ \exists h \in H \text{ s.t. } |\hat{R}(h) - R(h)| > \epsilon\ ] &= \\ Pr\left[\bigvee_{h \in H} |\hat{R}(h_i) - R(h_i)| > \epsilon\right] &\leq \\ \sum_{h \in H} Pr\left[|\hat{R}(h) - R(h)| > \epsilon\right] &\leq \\ 2|H| \exp[-2m\epsilon^2] & \end{aligned}$$

Finite inconsistent hypothesis class

Proof:
If we fix $\epsilon > 0$ and set $\delta = 2|H| \exp[-2m\epsilon^2]$, we can choose m large enough such that with confidence $1 - \delta$

$$\forall h \in H \quad |\hat{R}(h) - R(h)| \leq \epsilon \leq \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

which implies that

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$