University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chris Ketelsen

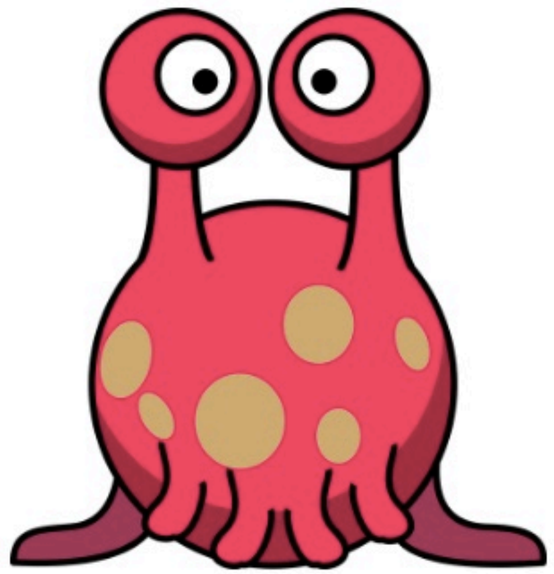
Lecture 12: Learning Theory Part 1
The PAC Framework

Learning Objectives

- [ ] Learn about what we can possibly learn
- [ ] Prove some simple bounds on errors and sample sizes
- [ ] Gain some intuition about complexity and overfitting

Simple 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 $S = \{(x_i, y_i)\}_{i=1}^{m}$
- 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?

Simple Example

Assumptions:

- Data comes from distribution D
- Concept c comes from concept class C
- Hypothesis h comes from hypothesis class H

Def: **Generalization Error**

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

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*?

Simple Example

Assumptions:

- Data comes from distribution D
- Concept c comes from concept class C
- Hypothesis h comes from hypothesis class H

Def: Generalization Error

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

Def: Training Error

$$\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} I[h(x_i) \neq c(x_i)]$$

Simple Example

Def: **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 $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$

Simple Example

Def: **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 $\delta > 0$

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

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

- S: The training set we learn from

Simple Example

Def: **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 $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$

- $\mathcal{D}$: The distribution the data comes from

Simple Example

Def: **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 $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$

- h_S: The hypothesis we learn from training set

Simple Example

Def: **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

Simple Example

Def: **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 $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$

- $1 - \epsilon$: The accuracy of h_s

Simple Example

Def: **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 $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$

- $1 - \delta$: The confidence the accuracy $1 - \epsilon$ is realized

Simple Example

Def: **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 $\delta > 0$

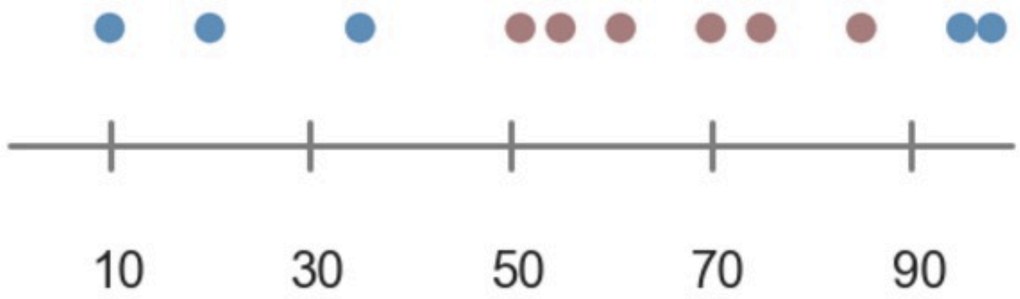
$$Pr_{S \sim \mathcal{D}^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$

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

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

PAC = Probably Approximately Correct

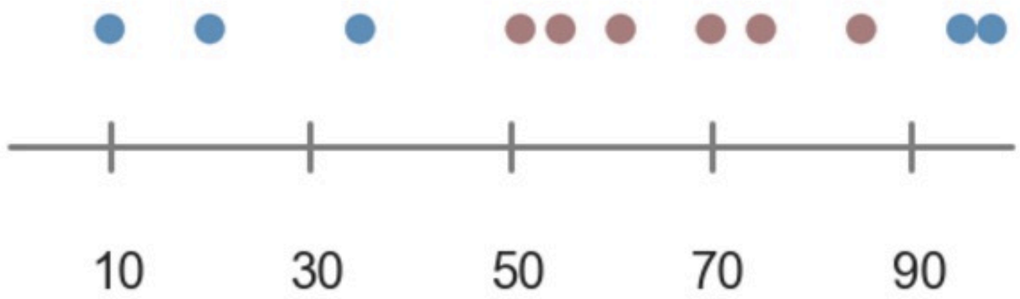
Simple Example



- 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

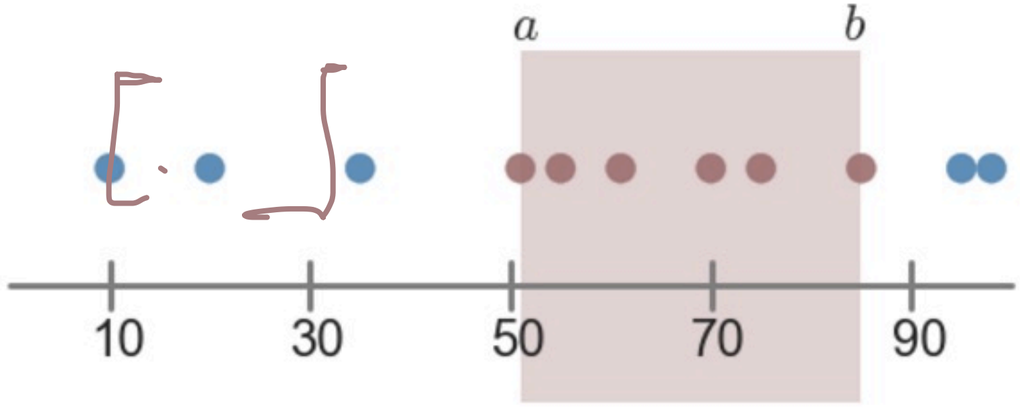
Simple Example



- What is Algorithm $\mathcal{A}$?

Set hypothesis to smallest interval containing S

Simple Example

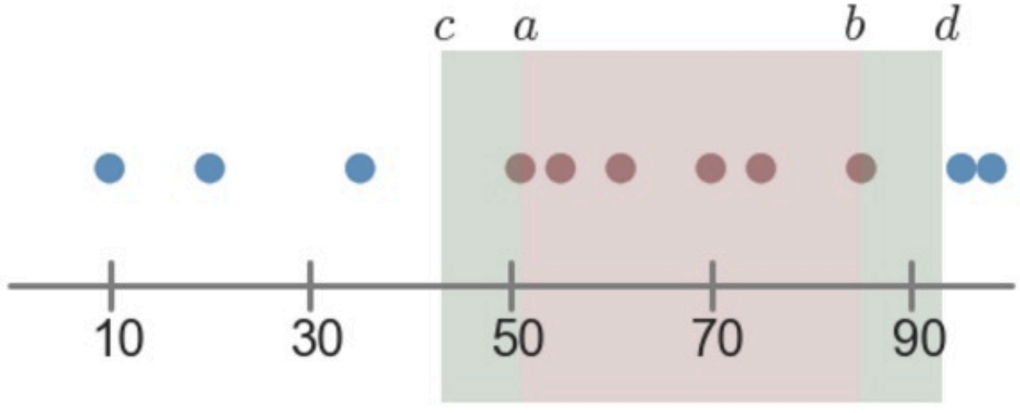


- What is Algorithm $\mathcal{A}$?

Set hypothesis to smallest interval containing S: $h_s = [a, b]$

Errors happen if a positive point falls outside of $h_s = [a, b]$

Simple Example



- What is Algorithm $\mathcal{A}$?

Set hypothesis to smallest interval containing S: $h_s = [a, b]$

Errors happen if a positive point false outside of $h_s = [a, b]$

Suppose true concept is $c = [c, d]$

Simple 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

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

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

So instead we'll try to prove something about

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

Simple Example

We want 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 $[b, d]$.

WLOG assume that probability of a point x from $\mathcal{D}$ landing in either missed interval is $\epsilon/2$

Useful Fact 1: Union Bound

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

Call $Pr_{S \sim \mathcal{D}^m}\ [R(h_S) > \epsilon]$ simplified $Pr[h_s \text{ is bad}]$

Simple Example

$$\begin{aligned} Pr[h_s \text{ is bad}] &= Pr[\text{no } x_i \text{ in } [c,a] \text{ or } [b,d]\ ] \\ &\leq Pr[\text{no } x_i \text{ in } [c,a]\ ] + Pr[\text{no } x_i \text{ in } [b,d]\ ] \end{aligned}$$

$$Pr[\text{ no } x_i \text{ in } [c,a]\ ] = Pr[\text{all } x_i \text{ not in } [c,a]\ ]$$

$$= \prod_{i=1}^{m} \left(1 - \frac{\epsilon}{2}\right) = \left(1 - \frac{\epsilon}{2}\right)^m$$

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

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

Simple Example

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &= Pr[\text{no } x_i \text{ in } [c,a] \text{ or } [b,d]\ ] \\
&\leq Pr[\text{no } x_i \text{ in } [c,a]\ ] + Pr[\text{no } x_i \text{ in } [b,d]\ ]
\end{aligned}
$$

$$
Pr[\text{ no } x_i \text{ in } [c,a]\ ] = Pr[\text{all } x_i \text{ not in } [c,a]\ ]
$$

$$
= \prod_{i=1}^{m} \left(1 - \frac{\epsilon}{2}\right) = \left(1 - \frac{\epsilon}{2}\right)^m
$$

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

Simple 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

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

Simple 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.

General Case, Finite Hypothesis Class

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?

The answer is Yes!

- Today, the case when H is finite
- Next Time, the case when H is infinite

Distinction:

- H is finite and $c \in H$
- H is finite and $c \notin H$

General Case, Finite Consistent Hypothesis Class

Def: 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.

Note: Might be multiple consistent h

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 Case, Finite Consistent Hypothesis Class

Def: 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

Note: Might be multiple consistent h

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?

Answer: The training error $\hat{R}(h) = 0$

General Case, 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

General Case, 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

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

$$Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } \ldots$$

$$\ldots \text{ or } (h_k \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)]$$

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

General Case, 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

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

$$Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } \ldots$$

$$\ldots \text{ or } (h_k \in H \text{ and } R(h_k) = 0 \text{ and } R(h_k) > \epsilon)] \quad \leq$$

$$\sum_h Pr[\ \hat{R}(h) = 0 \text{ and } R(h) > \epsilon]$$

Using the Union Bound

General Case, 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

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

$$Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } \ldots$$

$$\ldots \text{ or } (h_k \in H \text{ and } R(h_1) = 0 \text{ and } R(h_1) > \epsilon)] \quad \leq$$

$$\sum_h Pr[\ \hat{R}(h) = 0 \text{ and } R(h) > \epsilon] \quad \leq$$

$$\sum_h Pr[\ \hat{R}(h) = 0 \mid \text{and } R(h) > \epsilon]$$

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

General Case, 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

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

General Case, 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

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

General Case, 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

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

General Case, Finite Consistent Hypothesis Class

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

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

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

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

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

General Case, 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: $(1, 0, 0, 0)$

General Case, Finite Consistent Hypothesis Class

Question: What is an algorithm that will return a consistent hypothesis h given training data S?

General Case, Finite Consistent Hypothesis Class

Question: What is an algorithm that will return a consistent hypothesis h given training data S?

Answer: Loop over all positive examples $(b_1, b_2, \ldots, b_n)$. If $b_i = 1$ throw out $\bar{x}_i$. If $b_i = 0$, throw out x_i

x_1	x_2	x_3	x_4	x_5	x_6	y
0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	-
0	1	1	1	1	1	+
1	0	0	1	1	0	-
0	1	0	0	1	1	+
0	1	?	?	1	1	

Algorithm returns $\bar{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$

General Case, 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

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

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

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

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

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

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

Finite Inconsistent Hypothesis Class

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

Could talk about this one for an hour

- Larger m is, better training error predicts gen. error

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

- Training error would go down
- Bound term would go up ...

Finite Inconsistent Hypothesis Class

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

Could talk about this one for an hour

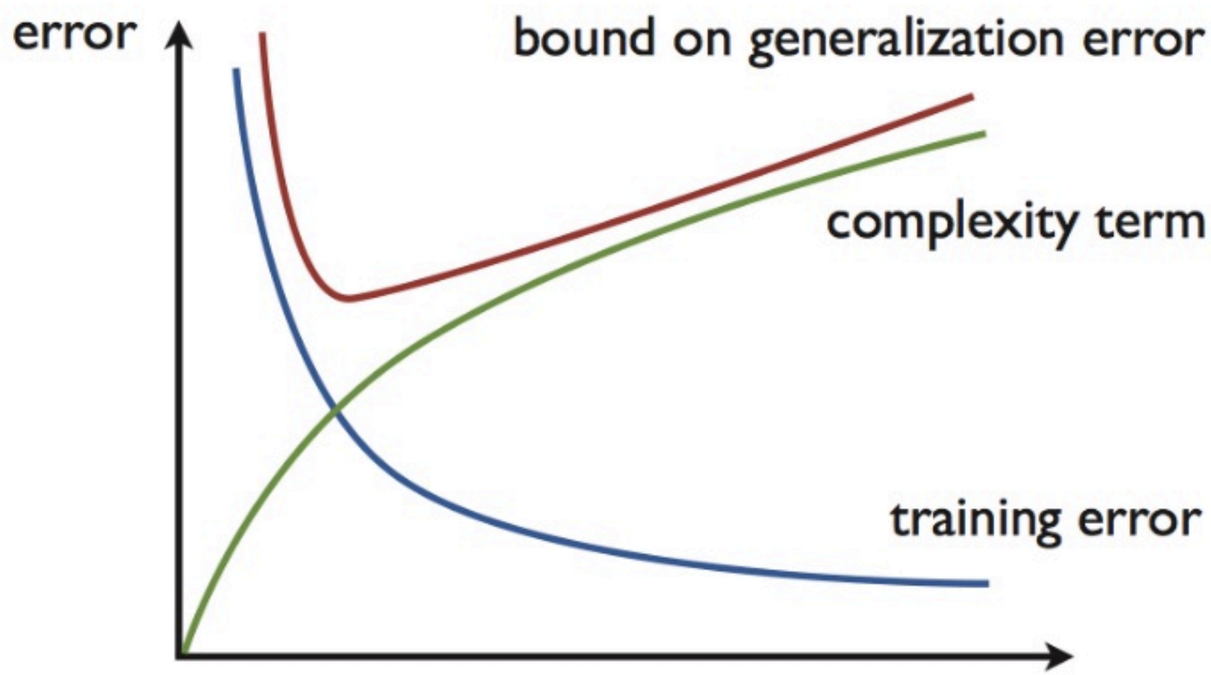
- Larger m is, better training error predicts gen. error

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

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



In Class

- **We Are In Class!**

Next Time

- See what happens when H is infinite
- Talk about the VC dimension
- Relate all this to SVMs

In Class

In Class

In Class

In Class