

# Machine Learning 10-601

Tom M. Mitchell  
Machine Learning Department  
Carnegie Mellon University

October 11, 2012

## Today:

- Computational Learning Theory
- Probably Approximately Correct (PAC) learning theorem
- Vapnik-Chervonenkis (VC) dimension

## Recommended reading:

- Mitchell: Ch. 7
- suggested exercises: 7.1, 7.2, 7.7

## Computational Learning Theory

- What general laws constrain inductive learning?
- Want theory to relate
  - Number of training examples
  - Complexity of hypothesis space
  - Accuracy to which target function is approximated
  - Manner in which training examples are presented
  - Probability of successful learning

\* See annual Conference on Computational Learning Theory

## Sample Complexity

How many training examples suffice to learn target concept

1. If learner proposes instances as queries to teacher?
  - learner proposes  $x$ , teacher provides  $f(x)$
2. If teacher (who knows  $f(x)$ ) proposes training examples?
  - teacher proposes sequence  $\langle x^1, f(x^1) \rangle, \dots \langle x^n, f(x^n) \rangle$
3. If some random process (e.g., nature) proposes instances, and teacher labels them?
  - instances drawn according to  $P(X)$

## Sample Complexity 3

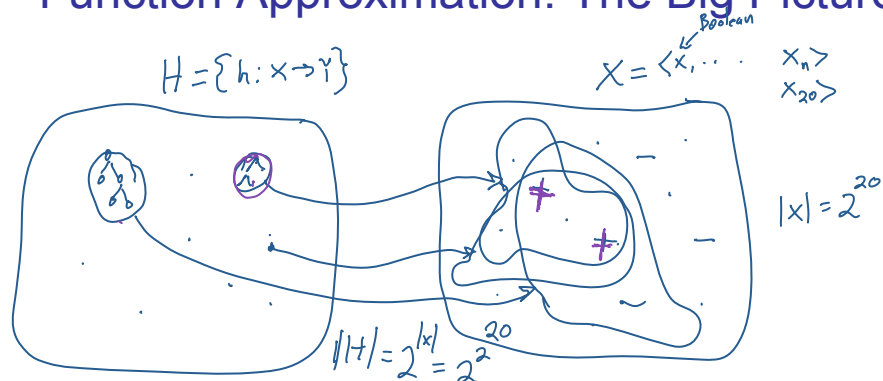
Problem setting:

- Set of instances  $X$
- Set of hypotheses  $H = \{h : X \rightarrow \{0, 1\}\}$
- Set of possible target functions  $C = \{c : X \rightarrow \{0, 1\}\}$
- Sequence of training instances drawn at random from  $P(X)$   
teacher provides noise-free label  $c(x)$

Learner outputs a hypothesis  $h \in H$  such that

$$h = \arg \min_{h \in H} \text{error}_{\text{train}}(h)$$

## Function Approximation: The Big Picture



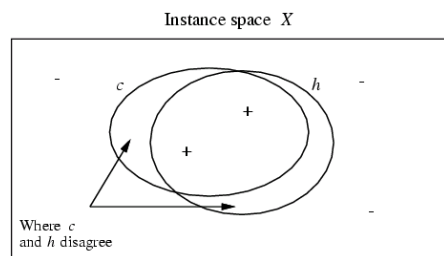
How many labeled examples are needed in order to determine which of the  $2^{20}$  hypotheses is the correct one?

All  $2^{20}$  instances in  $X$  must be labeled!

There is no free lunch!

Inductive inference - generalizing beyond the training data is impossible unless we add more assumptions (e.g. priors over  $h$ )

## True Error of a Hypothesis



The *true error* of  $h$  is the probability that it will misclassify an example drawn at random from  $P(X)$

$$\text{error}_{\text{true}}(h) \equiv \Pr_{x \sim P(X)}[h(x) \neq c(x)]$$

## Two Notions of Error

*Training error* of hypothesis  $h$  with respect to target concept  $c$

- How often  $h(x) \neq c(x)$  over training instances  $D$

$$error_{train} \equiv \Pr_{x \in D} [h(x) \neq c(x)] = \frac{1}{|D|} \sum_{x \in D} \frac{\delta(h(x) \neq c(x))}{|D|}$$

*True error* of hypothesis  $h$  with respect to  $c$

- How often  $h(x) \neq c(x)$  over future instances drawn at random from  $D$

$$error_{true}(h) \equiv \Pr_{x \sim P(X)} [h(x) \neq c(x)]$$

training examples  $D$

Probability distribution  $P(X)$

## Overfitting

Consider a hypothesis  $h$  and its

- Error rate over training data:  $error_{train}(h)$
- True error rate over all data:  $error_{true}(h)$

We say  $h$  overfits the training data if

$$error_{true}(h) > error_{train}(h)$$

Amount of overfitting =

$$error_{true}(h) - error_{train}(h)$$

## Overfitting

Consider a hypothesis  $h$  and its

- Error rate over training data:  $error_{train}(h)$
- True error rate over all data:  $error_{true}(h)$

We say  $h$  overfits the training data if

$$error_{true}(h) > error_{train}(h)$$

Amount of overfitting =

$$error_{true}(h) - error_{train}(h)$$

Can we bound  $error_{true}(h)$   
in terms of  $error_{train}(h)$  ??

$$error_{train} \equiv \Pr_{x \in D} [h(x) \neq c(x)] = \frac{1}{|D|} \sum_{x \in D} \frac{\delta(h(x) \neq c(x))}{|D|}$$

training  
examples

$$error_{true}(h) \equiv \Pr_{x \sim P(X)} [h(x) \neq c(x)]$$

Probability  
distribution  $P(x)$

if  $D$  was a set of examples drawn from  $P(X)$  and ***independent*** of  $h$ , then we could use standard statistical confidence intervals to determine that with 95% probability  $error_{true}(h)$  lies in the interval:

$$error_D(h) \pm 1.96 \sqrt{\frac{error_D(h)(1 - error_D(h))}{n}}$$

but  $D$  is the ***training data*** for  $h$  ....

## Version Spaces

$$c: X \rightarrow \{0,1\}$$

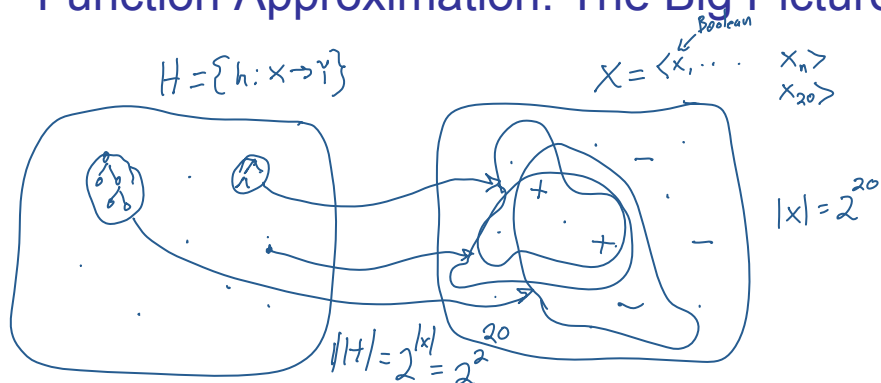
A hypothesis  $h$  is **consistent** with a set of training examples  $D$  of target concept  $c$  if and only if  $h(x) = c(x)$  for each training example  $\langle x, c(x) \rangle$  in  $D$ .

$$\text{Consistent}(h, D) \equiv (\forall \langle x, c(x) \rangle \in D) h(x) = c(x)$$

The **version space**,  $VS_{H,D}$ , with respect to hypothesis space  $H$  and training examples  $D$ , is the subset of hypotheses from  $H$  consistent with all training examples in  $D$ .

$$VS_{H,D} \equiv \{h \in H \mid \text{Consistent}(h, D)\}$$

## Function Approximation: The Big Picture



How many labeled examples are needed in order to determine which of the  $2^{2^{20}}$  hypotheses is the correct one?

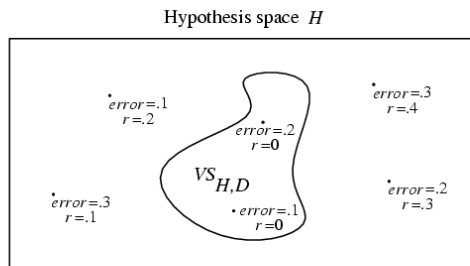
All  $2^{20}$  instances in  $X$  must be labeled!

There is no free lunch!

Inductive inference - generalizing beyond the training data is impossible unless we add more assumptions (e.g. priors over  $H$ )

## Exhausting the Version Space

---



( $r$  = training error,  $error$  = true error)

**Definition:** The version space  $VS_{H,D}$  with respect to training data  $D$  is said to be  $\epsilon$ -**exhausted** if every hypothesis  $h$  in  $VS_{H,D}$  has true error less than  $\epsilon$ .

$$(\forall h \in VS_{H,D}) \text{error}_{true}(h) < \epsilon$$

How many examples will  $\epsilon$ -exhaust the VS?

---

**Theorem:** [Haussler, 1988].

If the hypothesis space  $H$  is finite, and  $D$  is a sequence of  $m \geq 1$  independent random examples of some target concept  $c$ , then for any  $0 \leq \epsilon \leq 1$ , the probability that the version space with respect to  $H$  and  $D$  is not  $\epsilon$ -exhausted (with respect to  $c$ ) is less than

$$|H|e^{-\epsilon m}$$

How many examples will  $\epsilon$ -exhaust the VS?

---

**Theorem:** [Haussler, 1988].

If the hypothesis space  $H$  is finite, and  $D$  is a sequence of  $m \geq 1$  independent random examples of some target concept  $c$ , then for any  $0 \leq \epsilon \leq 1$ , the probability that the version space with respect to  $H$  and  $D$  is not  $\epsilon$ -exhausted (with respect to  $c$ ) is less than

$$|H|e^{-\epsilon m}$$

Interesting! This bounds the probability that any consistent learner will output a hypothesis  $h$  with  $error(h) \geq \epsilon$

Any(!) learner that outputs a hypothesis consistent with all training examples (i.e., an  $h$  contained in  $VS_{H,D}$ )



## What it means

[Haussler, 1988]: probability that the version space is not  $\epsilon$ -exhausted after  $m$  training examples is at most  $|H|e^{-\epsilon m}$

$$\Pr[(\exists h \in H) \text{ s.t. } (error_{train}(h) = 0) \wedge (error_{true}(h) > \epsilon)] \leq |H|e^{-\epsilon m}$$

↑

Suppose we want this probability to be at most  $\delta$

1. How many training examples suffice?

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

2. If  $error_{train}(h) = 0$  then with probability at least  $(1-\delta)$ :

$$error_{true}(h) \leq \frac{1}{m} (\ln |H| + \ln(1/\delta))$$

## Example: H is Conjunction of up to N Boolean Literals

Consider classification problem  $f: X \rightarrow Y$ :  $m \geq \frac{1}{\epsilon} (\ln |H| + \ln(1/\delta))$

- instances:  $X = \langle X_1 X_2 X_3 X_4 \rangle$  where each  $X_i$  is boolean
- Each hypothesis in  $H$  is a rule of the form:
  - IF  $\langle X_1 X_2 X_3 X_4 \rangle = \langle 0, ?, 1, ? \rangle$ , THEN  $Y=1$ , ELSE  $Y=0$
  - i.e., rules constrain any subset of the  $X_i$

How many training examples  $m$  suffice to assure that with probability at least 0.99, *any* consistent learner using  $H$  will output a hypothesis with true error at most 0.05?

## Example: $H$ is Decision Tree with depth=2

Consider classification problem  $f: X \rightarrow Y$ :

- instances:  $X = \langle X_1 \dots X_N \rangle$  where each  $X_i$  is boolean
- learned hypotheses are decision trees of depth 2, using only two variables

How many training examples  $m$  suffice to assure that with probability at least 0.99, *any* consistent learner will output a hypothesis with true error at most 0.05?

## PAC Learning

---

Consider a class  $C$  of possible target concepts defined over a set of instances  $X$  of length  $n$ , and a learner  $L$  using hypothesis space  $H$ .

*Definition:*  $C$  is **PAC-learnable** by  $L$  using  $H$  if for all  $c \in C$ , distributions  $\mathcal{D}$  over  $X$ ,  $\epsilon$  such that  $0 < \epsilon < 1/2$ , and  $\delta$  such that  $0 < \delta < 1/2$ , learner  $L$  will with probability at least  $(1 - \delta)$  output a hypothesis  $h \in H$  such that  $\text{error}_{\mathcal{D}}(h) \leq \epsilon$ , in time that is polynomial in  $1/\epsilon$ ,  $1/\delta$ ,  $n$  and  $\text{size}(c)$ .

## PAC Learning

---

Consider a class  $C$  of possible target concepts defined over a set of instances  $X$  of length  $n$ , and a learner  $L$  using hypothesis space  $H$ .

*Definition:*  $C$  is **PAC-learnable** by  $L$  using  $H$  if for all  $c \in C$ , distributions  $\mathcal{D}$  over  $X$ ,  $\epsilon$  such that  $0 < \epsilon < 1/2$ , and  $\delta$  such that  $0 < \delta < 1/2$ , learner  $L$  will with probability at least  $(1 - \delta)$  output a hypothesis  $h \in H$  such that  $\text{error}_{\mathcal{D}}(h) \leq \epsilon$ , in time that is polynomial in  $1/\epsilon$ ,  $1/\delta$ ,  $n$  and  $\text{size}(c)$ .

Sufficient condition:

Holds if learner  $L$  requires only a polynomial number of training examples, and processing per example is polynomial

## Agnostic Learning

---

So far, assumed  $c \in H$

Agnostic learning setting: don't assume  $c \in H$

- What do we want then?
  - The hypothesis  $h$  that makes fewest errors on training data
- What is sample complexity in this case?

$$m \geq \frac{1}{2\epsilon^2}(\ln |H| + \ln(1/\delta))$$

Here  $\epsilon$  is the difference between the training error and true error of the output hypothesis (the one with lowest training error)

## Additive Hoeffding Bounds – Agnostic Learning

- Given  $m$  independent flips of a coin with true  $\Pr(\text{heads}) = \theta$   
we can bound the error  $\epsilon$  in the maximum likelihood estimate  $\hat{\theta}$   

$$\Pr[\theta > \hat{\theta} + \epsilon] \leq e^{-2m\epsilon^2}$$
- Relevance to agnostic learning: for any single hypothesis  $h$   

$$\Pr[\text{error}_{\text{true}}(h) > \text{error}_{\text{train}}(h) + \epsilon] \leq e^{-2m\epsilon^2}$$
- But we must consider all hypotheses in  $H$   

$$\Pr[(\exists h \in H) \text{error}_{\text{true}}(h) > \text{error}_{\text{train}}(h) + \epsilon] \leq |H|e^{-2m\epsilon^2}$$
- So, with probability at least  $(1-\delta)$  every  $h$  satisfies  

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

## General Hoeffding Bounds

- When estimating parameter  $\theta$  inside  $[a,b]$  from  $m$  examples  

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$
- When estimating a probability  $\theta$  is inside  $[0,1]$ , so  

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{-2m\epsilon^2}$$
- And if we're interested in only one-sided error, then  

$$P((E[\hat{\theta}] - \hat{\theta}) > \epsilon) \leq e^{-2m\epsilon^2}$$

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

Question: If  $H = \{h \mid h: X \rightarrow Y\}$  is infinite, what measure of complexity should we use in place of  $|H|$  ?

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

Question: If  $H = \{h \mid h: X \rightarrow Y\}$  is infinite, what measure of complexity should we use in place of  $|H|$  ?

Answer: The largest subset of  $X$  for which  $H$  can guarantee zero training error (regardless of the target function  $c$ )

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

Question: If  $H = \{h \mid h: X \rightarrow Y\}$  is infinite, what measure of complexity should we use in place of  $|H|$  ?

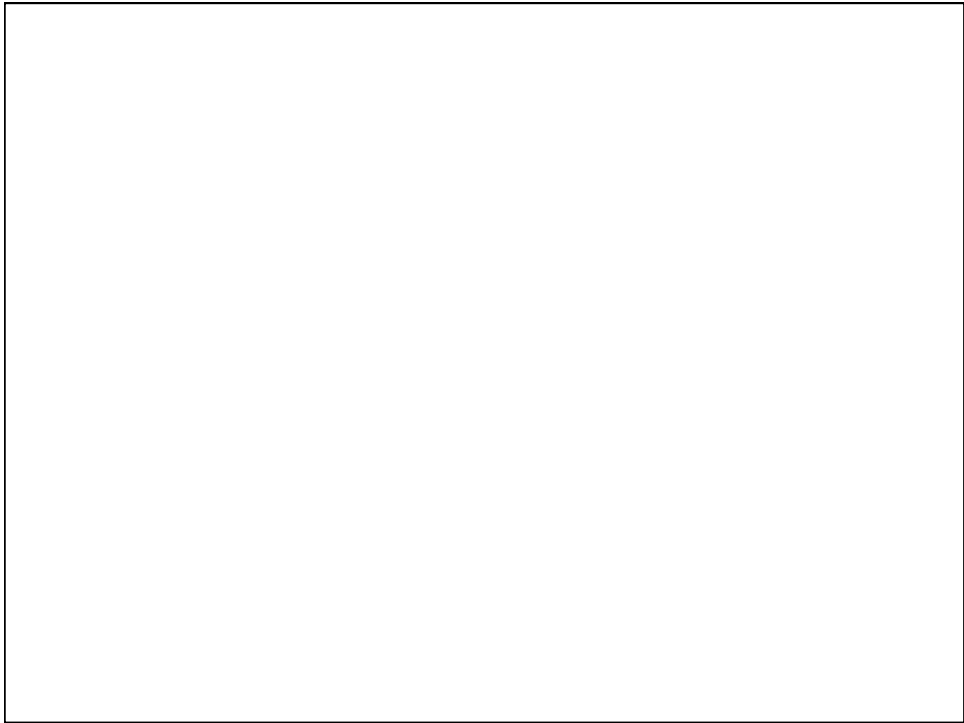
Answer: The largest subset of  $X$  for which  $H$  can guarantee zero training error (regardless of the target function  $c$ )

**VC dimension of  $H$  is the size of this subset**

Question: If  $H = \{h \mid h: X \rightarrow Y\}$  is infinite, what measure of complexity should we use in place of  $|H|$  ?

Answer: The largest subset of  $X$  for which  $H$  can guarantee zero training error (regardless of the target function  $c$ )

Informal intuition:



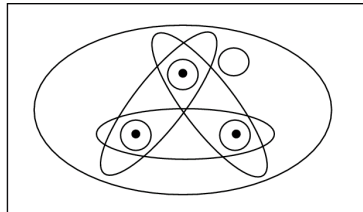
## Shattering a Set of Instances

*Definition:* a **dichotomy** of a set  $S$  is a partition of  $S$  into two disjoint subsets.

a labeling of each member of  $S$  as positive or negative

*Definition:* a set of instances  $S$  is **shattered** by hypothesis space  $H$  if and only if for every dichotomy of  $S$  there exists some hypothesis in  $H$  consistent with this dichotomy.

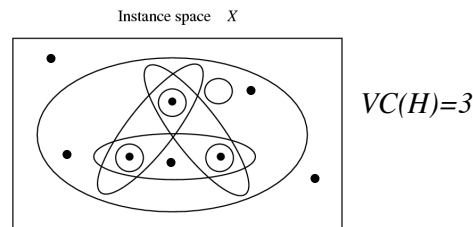
Instance space  $X$



## The Vapnik-Chervonenkis Dimension

---

*Definition:* The **Vapnik-Chervonenkis dimension**,  $VC(H)$ , of hypothesis space  $H$  defined over instance space  $X$  is the size of the largest finite subset of  $X$  shattered by  $H$ . If arbitrarily large finite sets of  $X$  can be shattered by  $H$ , then  $VC(H) \equiv \infty$ .



## Sample Complexity based on VC dimension

How many randomly drawn examples suffice to  $\epsilon$ -exhaust  $VS_{H,D}$  with probability at least  $(1-\delta)$ ?

ie., to guarantee that any hypothesis that perfectly fits the training data is probably  $(1-\delta)$  approximately ( $\epsilon$ ) correct

$$m \geq \frac{1}{\epsilon} (4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon))$$

Compare to our earlier results based on  $|H|$ :

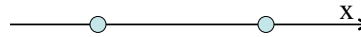
$$m \geq \frac{1}{\epsilon} (\ln(1/\delta) + \ln |H|)$$



## VC dimension: examples

Consider  $X = \mathbb{R}$ , want to learn  $c: X \rightarrow \{0,1\}$

What is VC dimension of



- Open intervals:

H1: if  $x > a$  then  $y = 1$  else  $y = 0$

H2: if  $x > a$  then  $y = 1$  else  $y = 0$

or, if  $x > a$  then  $y = 0$  else  $y = 1$

- Closed intervals:

H3: if  $a < x < b$  then  $y = 1$  else  $y = 0$

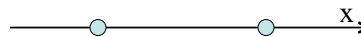
H4: if  $a < x < b$  then  $y = 1$  else  $y = 0$

or, if  $a < x < b$  then  $y = 0$  else  $y = 1$

## VC dimension: examples

Consider  $X = \mathbb{R}$ , want to learn  $c: X \rightarrow \{0,1\}$

What is VC dimension of



- Open intervals:

H1: if  $x > a$  then  $y = 1$  else  $y = 0$        $VC(H1)=1$

H2: if  $x > a$  then  $y = 1$  else  $y = 0$        $VC(H2)=2$   
or, if  $x > a$  then  $y = 0$  else  $y = 1$

- Closed intervals:

H3: if  $a < x < b$  then  $y = 1$  else  $y = 0$        $VC(H3)=2$

H4: if  $a < x < b$  then  $y = 1$  else  $y = 0$        $VC(H4)=3$   
or, if  $a < x < b$  then  $y = 0$  else  $y = 1$

## VC dimension: examples

What is VC dimension of lines in a plane?

- $H_2 = \{ ((w_0 + w_1x_1 + w_2x_2) > 0 \rightarrow y=1) \}$



## VC dimension: examples

What is VC dimension of

- $H_2 = \{ ((w_0 + w_1x_1 + w_2x_2) > 0 \rightarrow y=1) \}$ 
  - $VC(H_2)=3$
- For  $H_n$  = linear separating hyperplanes in  $n$  dimensions,  $VC(H_n)=n+1$



For any finite hypothesis space  $H$ , can you give an upper bound on  $VC(H)$  in terms of  $|H|$  ?  
(hint: yes)

### More VC Dimension Examples to Think About

- Logistic regression over  $n$  continuous features
  - Over  $n$  boolean features?
- Linear SVM over  $n$  continuous features
- Decision trees defined over  $n$  boolean features  
 $F: \langle X_1, \dots, X_n \rangle \rightarrow Y$
- Decision trees of depth 2 defined over  $n$  features
- How about 1-nearest neighbor?

## Tightness of Bounds on Sample Complexity

How many examples  $m$  suffice to assure that any hypothesis that fits the training data perfectly is probably  $(1-\delta)$  approximately  $(\epsilon)$  correct?

$$m \geq \frac{1}{\epsilon} (4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon))$$

How tight is this bound?

## Tightness of Bounds on Sample Complexity

How many examples  $m$  suffice to assure that any hypothesis that fits the training data perfectly is probably  $(1-\delta)$  approximately  $(\epsilon)$  correct?

$$m \geq \frac{1}{\epsilon} (4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon))$$

How tight is this bound?

**Lower bound on sample complexity** (Ehrenfeucht et al., 1989):

Consider any class  $C$  of concepts such that  $VC(C) > 1$ , any learner  $L$ , any  $0 < \epsilon < 1/8$ , and any  $0 < \delta < 0.01$ . Then there exists a distribution  $\mathcal{D}$  and a target concept in  $C$ , such that if  $L$  observes fewer examples than

$$\max \left[ \frac{1}{\epsilon} \log(1/\delta), \frac{VC(C) - 1}{32\epsilon} \right]$$

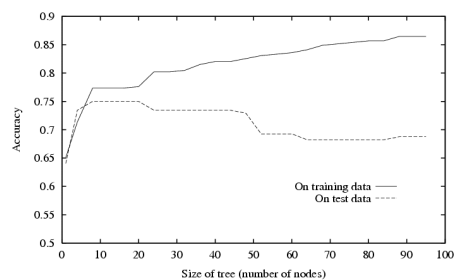
Then with probability at least  $\delta$ ,  $L$  outputs a hypothesis with  $error_{\mathcal{D}}(h) > \epsilon$

## Agnostic Learning: VC Bounds

[Schölkopf and Smola, 2002]

With probability at least  $(1-\delta)$  every  $h \in H$  satisfies

$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}$$

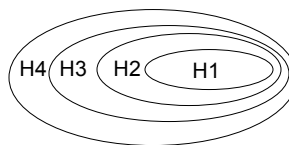


## Structural Risk Minimization

[Vapnik]

Which hypothesis space should we choose?

- Bias / variance tradeoff



SRM: choose  $H$  to minimize bound on expected true error!

$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}$$

\* unfortunately a somewhat loose bound...