Learning Theory

Machine Learning 10-601B

Seyoung Kim

Many of these slides are derived from Tom Mitchell, Ziv-Bar Joseph. Thanks!

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 some random process (e.g., nature) proposes instances, and teacher labels them?

   - instances drawn according to $P(X)$

2. If learner proposes instances as queries to teacher?

   - learner proposes x, teacher provides f(x)

   Active learning

3. If teacher (who knows f(x)) proposes training examples?

   - teacher proposes sequence $\{<x^1, f(x^1)>, \ldots <x^n, f(x^n)>$

Learning Theory

- In general, we are interested in
  - Sample complexity: How many training examples are needed for a learner to converge to a successful hypothesis?
  - Computational complexity: How much computational effort is needed for a learner to converge to a successful hypothesis?
  - The two are related. Why?

Problem Setting for Learning from Data

Given:

- Set of instances $X = \{x_1, \ldots, x_n\}$ for *n* input features
- Sequence of input instances drawn at random from $P(X)$
- Set of hypotheses $H = \{h : X \rightarrow \{0, 1\}\}$
- Set of possible target functions $C = \{c : X \rightarrow \{0, 1\}\}$
- teacher provides noise-free label $c(x)$

Learner observes a sequence *D* of training examples of the form $<x, c(x)>$ for some target concept $c \in C$

- Instances x are drawn from *P(X)*
- Teacher provides target value $c(x)$

Problem Setting for Learning from Data

Goal: Then, learner must output a hypothesis $h \in H$ estimating *c* such that

$$h = \arg\min_{h \in H} \; error_{train}(h)$$

h is evaluated on subsequent instances drawn from P(X)

Randomly drawn instances, noise-free classification

Function Approximation with Training Data

- Given X = {x: x is boolean and x={x_1, …, x_n} } with n input features

- How many possible input values? |X| = 2^n

- How many possible label assignments? 2^{2^n}

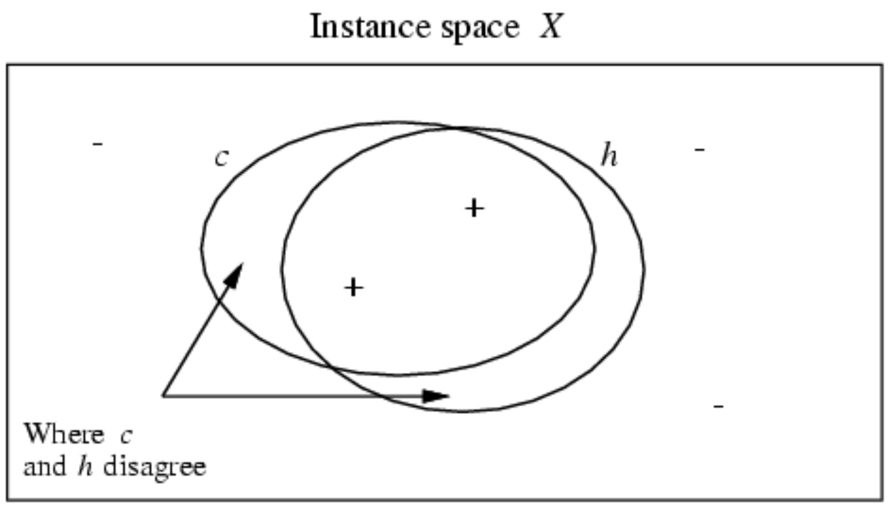
- The size of hypothesis space that can represent all possible label assignments? |H|= 2^{2^n}

In order to find h that is identical to c, we need observations for all data |X| = 2^n

In practice, we are limited by training data!

Need to introduce inductive bias

True Error of a Hypothesis



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

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

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

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

training examples D

True error of hypothesis h with respect to c

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

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

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

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

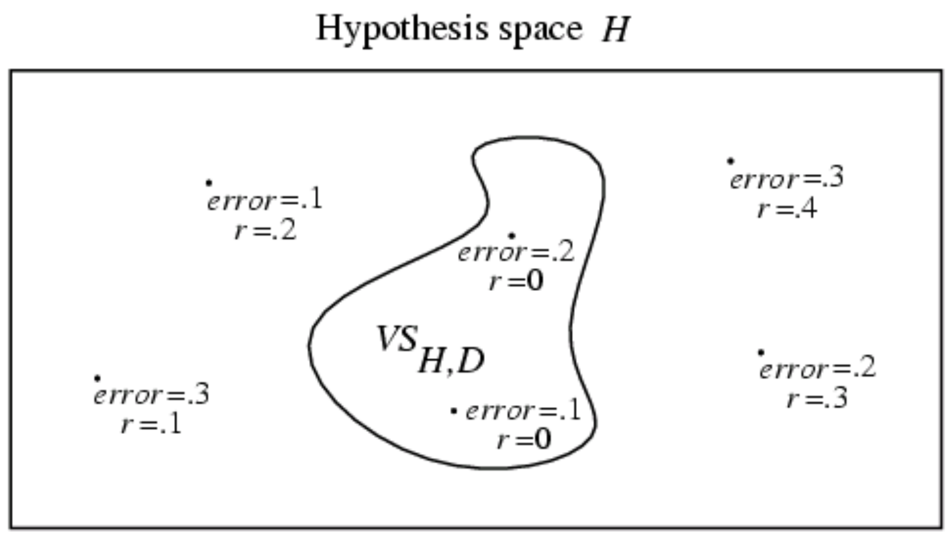
A hypothesis h is **consistent** with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example $\langle x, c(x) \rangle$ in D.

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

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.

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

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 ϵ-**exhausted** if every hypothesis h in $VS_{H,D}$ has true error less than ϵ.

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

How many examples will ϵ-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 ϵ-exhausted (with respect to c) is less than

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

How many examples will ϵ-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 ϵ-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 $VS_{H,D}$)

Proof:

- Given, hypothesis space H, input space X, m labeled examples, target function $C = \{c : X \rightarrow \{0, 1\}\}$, error tolerance ε
- Let h_1, h_2, …, h_K be hypotheses with true error > ε. Then,

Probability that h_1 will be consistent with first training example ≤ (1- ε)

Probability that h_1 will be consistent with m independently drawn training examples ≤ $(1- \varepsilon)^m$

Probability that at least one of h_1, h_2, …, h_K (K bad hypotheses) will be consistent with m examples ≤ $K(1- \varepsilon)^m$

$\leq |H|(1- \varepsilon)^m$ since $K \leq |H|$

$\leq |H| e^{-\varepsilon m}$ since for $0 \leq \varepsilon \leq 1$, $(1- \varepsilon) \leq e^{-\varepsilon m}$

What it means

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

↑

Suppose we want this probability to be at most δ

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

1. How many training examples suffice?

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

Learning Conjunctions of Boolean Literals

How many examples are sufficient to assure with probability at least $(1-\delta)$ that

every h in $VS_{H,D}$ satisfies $error_{\mathcal{D}}(h) \leq \epsilon$

Use our theorem:

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

Suppose H contains conjunctions of constraints on up to n boolean attributes (i.e., n boolean literals).

E.g.,

X=< X1, X2, … Xn >

Each h ∈ H constrains each Xi to be 1, 0, or "don't care"

In other words, each h is a rule such as:

If X2=0 and X5=1

Then Y=1, else Y=0

Learning Conjunctions of Boolean Literals

How many examples are sufficient to assure with probability at least $(1-\delta)$ that

every h in $VS_{H,D}$ satisfies $error_{\mathcal{D}}(h) \leq \epsilon$

Use our theorem:

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

Suppose H contains conjunctions of constraints on up to n boolean attributes (i.e., n boolean literals). Then $|H| = 3^n$, and

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

or

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

Example: Simple decision trees

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

Consider Boolean classification problem

- instances: $X = \langle X_1 \ldots X_N \rangle$ where each X_i is boolean
- Each hypothesis in H is a decision tree of depth 1

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: Simple decision trees

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

Consider Boolean classification problem

- instances: $X = <X_1 \ldots X_n>$ where each X_i is boolean
- Each hypothesis in H is a decision tree of depth 1

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?

|H| = 4 n, epsilon = 0.05, delta = 0.01

$$m \geq \frac{1}{0.05}(\ln(4N) + \ln\frac{1}{0.01})$$

N=4	m≥148
N=10	m≥166
N=100	m≥212

Example: H is Decision Tree with depth=2

Consider classification problem f:X→Y:

- instances: $X = <X_1 \ldots X_N>$ 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?

Example: H is Decision Tree with depth=2

Consider classification problem f:X→Y:

- instances: $X = \langle X_1 \ldots X_N \rangle$ 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?

|H| = N(N-1)/2 x 16

$$m \geq \frac{1}{0.05}(\ln(8N^2 - 8N) + \ln\frac{1}{0.01})$$

N=4	m≥184
N=10	m≥224
N=100	m≥318

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

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, ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$,
learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $size(c)$.

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, ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$,
learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $size(c)$.

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 ε 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(heads) = θ
  we can bound the error ϵ in the maximum likelihood estimate $\widehat{\theta}$

$$\Pr[\theta > \hat{\theta} + \epsilon] \leq e^{-2m\epsilon^2}$$

- Relevance to agnostic learning: for any <u>*single*</u> hypothesis *h*

$$\Pr[error_{true}(h) > error_{train}(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

- But we must consider all hypotheses in H

$$\Pr[(\exists h \in H) error_{true}(h) > error_{train}(h) + \epsilon] \leq |H| e^{-2m\epsilon^2}$$

- Now we assume this probability is bounded by δ. Then, we have

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

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

Question: If H = {h | h: X → 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 | h: X → 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 | h: X → 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

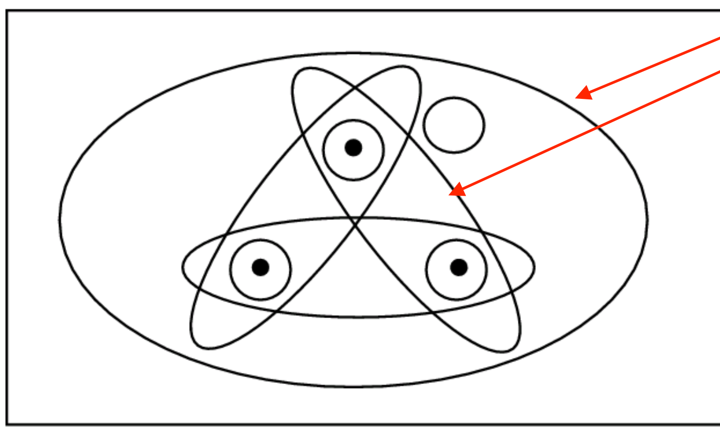
Shattering a Set of Instances

a labeling of each member of S as positive or negativea

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

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



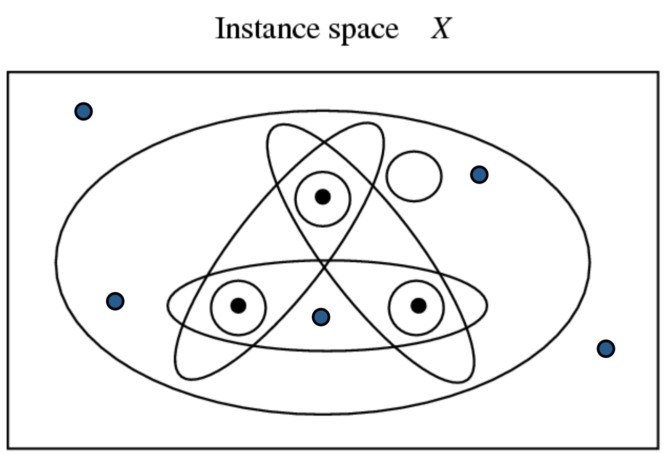
Each ellipse corresponds to a possible dichotomy

Positive: Inside the ellipse

Negative: Outside the ellipse

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 $VC(H) \equiv \infty$.



VC(H)=3

Sample Complexity based on VC dimension

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

ie., to guarantee that any hypothesis that perfectly fits the training data is probably (1-δ) approximately (ε) correct

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

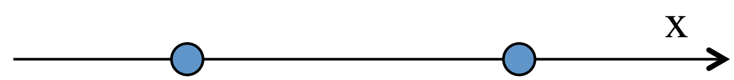
Compare to our earlier results based on $|H|$:

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

VC dimension: examples

Consider 1-dim real valued input X, want to learn c:X→{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:

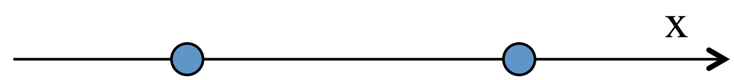
  H3: if $a < x < b$ then $y = 1$ else $y = 0$

  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 1-dim real valued input X, want to learn c:X→{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:

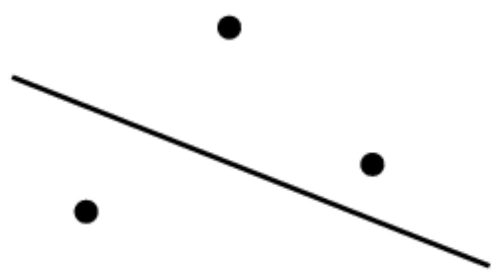
H3: if $a < x < b$ then $y = 1$ else $y = 0$ VC(H3)=2

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?

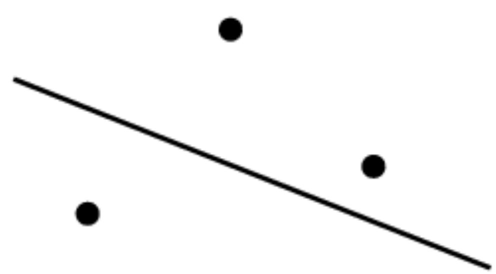
- $H_2 = \{ ((w_0 + w_1x_1 + w_2x_2)>0 \rightarrow y=1) \}$



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)

Assume VC(H) = K, which means H can shatter K examples.

For K examples, there are 2^K possible labelings. Thus, $|H| \geq 2^K$

Thus, $K \leq \log_2 |H|$

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: $\langle X_1, \dots X_n \rangle \rightarrow Y$

- 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-δ) approximately (ε) 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-δ) approximately (ε) 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 < ε < 1/8, and any 0 < δ < 0.01. Then there exists a distribution and a target concept in C, such that if L observes fewer examples than $\mathcal{D}$

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

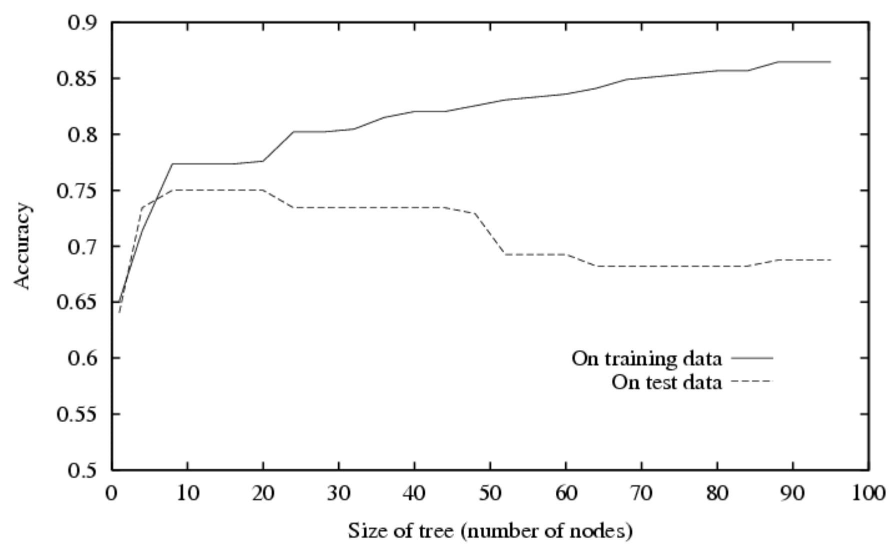
Then with probability at least δ, L outputs a hypothesis with $error_{\mathcal{D}}(h) > \epsilon$

Agnostic Learning: VC Bounds for Decision Tree

[Schölkopf and Smola, 2002]

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

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



What You Should Know

- Sample complexity varies with the learning setting
  - Learner actively queries trainer
  - Examples arrive at random

- Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  - For ANY consistent learner (case where $c \in H$)
  - For ANY "best fit" hypothesis (agnostic learning, where perhaps c not in H)

- VC dimension as a measure of complexity of H

- Conference on Learning Theory: http://www.learningtheory.org
- Avrim Blum's course on Machine Learning Theory:
  - https://www.cs.cmu.edu/~avrim/ML14/