Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

November 1, 2017

Today:

- Computational Learning Theory
- Vapnik-Chervonenkis (VC) dimension
- Agnostic learning

Recommended reading:

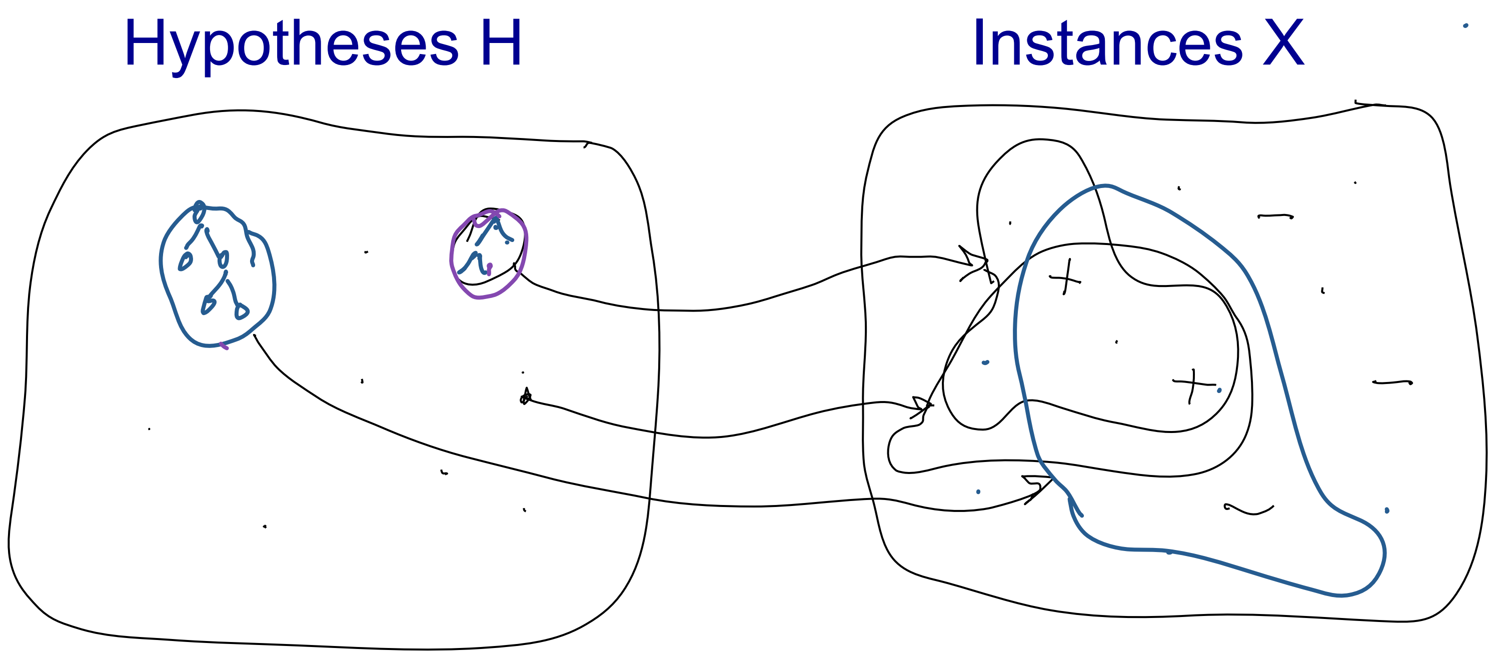
- Prof. Balcan notes: see Piazza syllabus
- Mitchell Ch. 7

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

Function Approximation: The Big Picture



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

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

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 $error_{train}(h)$??

What it means

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

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

Suppose we want this probability to be at most δ

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-δ):

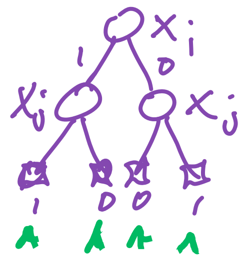
$$error_{true}(h) \leq \frac{1}{m}(\ln|H| + \ln(1/\delta))$$

Example: Depth 2 Decision Trees

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

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



$$|H| = \binom{n}{2} / 2 \cdot 16 = \frac{N(N-1)}{2} 16$$

$$|H| = 8N(N-1)$$

How many training examples *m* suffice to assure that with probability at least 0.99, *any* learner that outputs a consistent depth 2 decision tree will have true error at most 0.05?

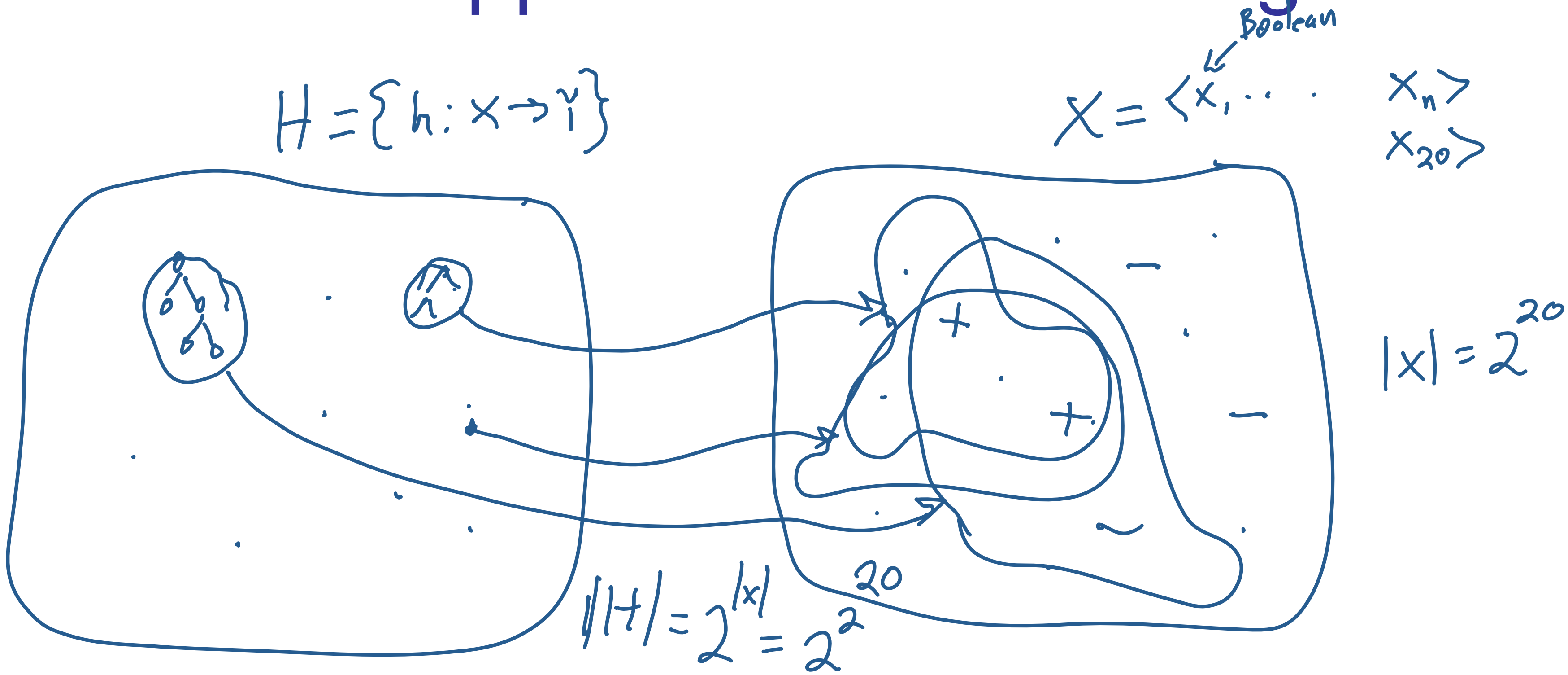
$$m \geq \frac{1}{0.05}\left(\ln(8N^2 - 8N) + \ln\frac{1}{0.01}\right)$$

$N=4 \rightarrow m \geq 184$

$N=10, m \geq 224$

$N=100, m \geq 318$

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?

<u>All</u> 2^{20} instances in X must be labeled!

There is no free lunch!

Inductive inference – generalizing beyond the training data is impossible <u>unless</u> we add more assumptions (e.g. priors over H)

Example: Depth 2 Decision Trees

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

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 up to n

How many training examples *m* suffice to assure that with probability at least 0.99, *any* learner that outputs a consistent depth 2 decision tree will have 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, 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 <u>time that is polynomial</u> 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 <u>time that is polynomial</u> 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 (this holds for all h in H)

Additive Hoeffding Bounds – Agnostic Learning

- Given *m* independent flips of a coin with true Pr(heads) = θ we can bound the error ϵ of the maximum likelihood estimate $\widehat{\theta}$

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

- Relevance to agnostic learning: for any *single* hypothesis *h*

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

- But we must consider all hypotheses in H

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

- So, with probability at least (1-δ) *every* h satisfies

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

General Hoeffding Bounds

- When estimating parameter θ inside [a,b] from *m* examples

$$P(|\widehat{\theta} - E[\widehat{\theta}]| > \epsilon) \leq 2e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$

- When estimating a probability θ is inside [0,1], so

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

- And if we're interested in only one-sided error, then

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

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

Here ε is the difference between the training error and true error of the output hypothesis (this holds for all h in H)

But, the output h with lowest training error might not give us the h^* with lowest true error. How far can true error of h be from h^* ?

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

Here ε is the difference between the training error and true error of the output hypothesis (this holds for all h in H)

But, the output h with lowest training error might not give us the h^* with lowest true error. How far can true error of h be from h^* ?

$$error_{true}(h) \leq error_{true}(h^*) + 2\epsilon$$

best training error hypothesis

best true error hypothesis

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

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 <u>guarantee</u> zero training error (regardless of how it is labeled)

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

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

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

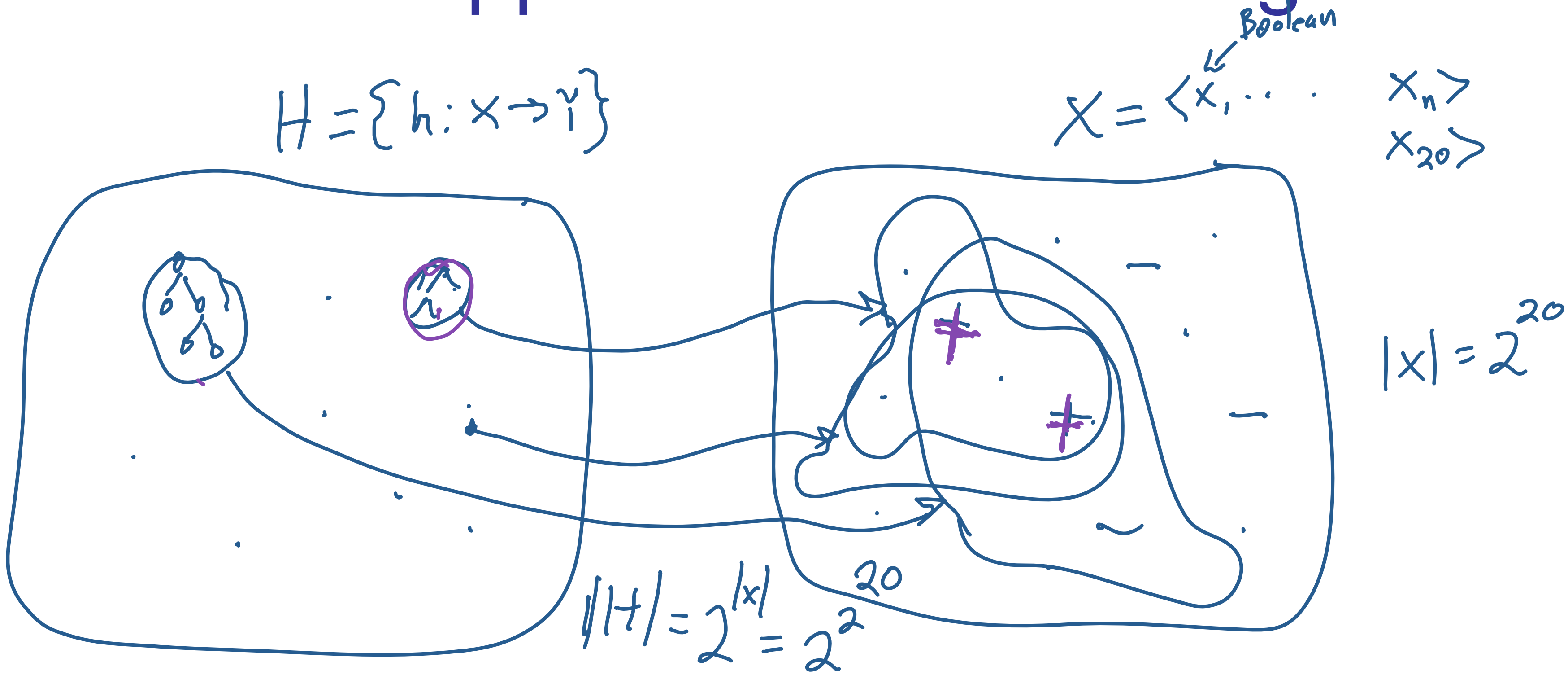
this is the VC dimension of H

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:

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?

<u>All</u> 2^{20} instances in X must be labeled!

There is no free lunch!

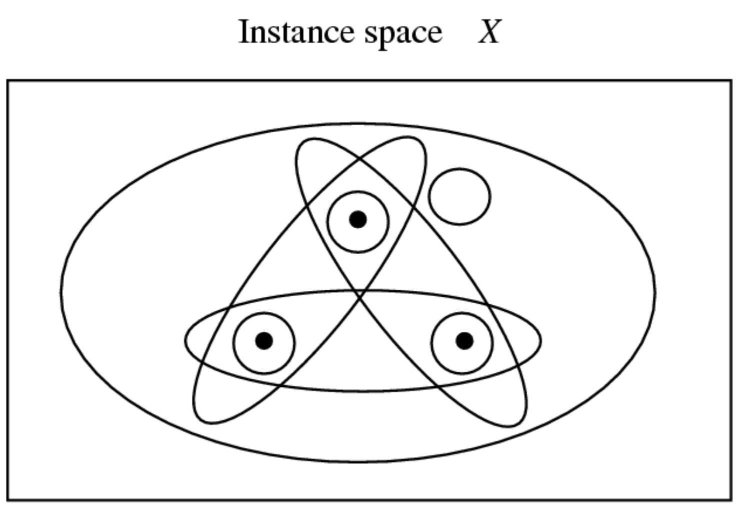
Inductive inference - generalizing beyond the training data is impossible <u>unless</u> we add more assumptions (e.g. priors over H)

Shattering a Set of Instances

a labeling of each member of S as positive or negative

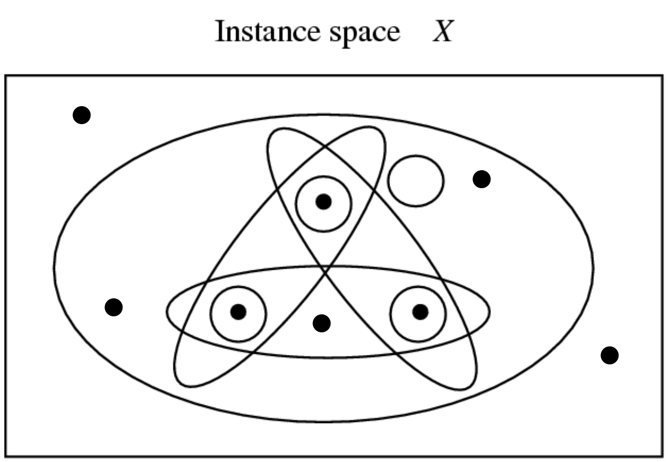
Definition: a **dichotomy** of a set S is a partition of S into two disjoint subsets.

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.



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 $(1-\delta)$?

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

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

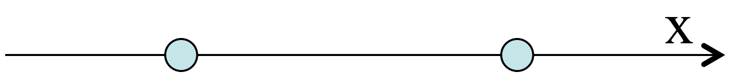
Compare to our earlier results based on $|H|$:

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

VC dimension: examples

Consider 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:

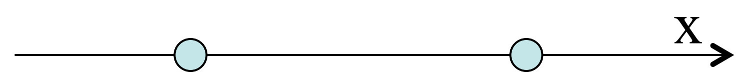
  H3: if $a < x < b$ then $y = 1$ else $y = 0$

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

- Closed intervals:

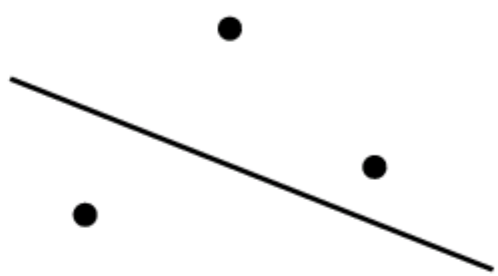
H3: if $a < x < b$ then $y = 1$ else $y = 0$ VC(H3)=2

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

VC dimension: examples

What is VC dimension of lines in a plane?

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

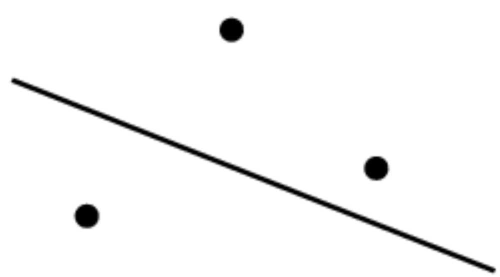


VC dimension: examples

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?

- Decision trees defined over n boolean features

  F: $<X_1, \ldots X_n> \rightarrow Y$

- Decision trees of depth 2 defined over n features

- Naïve Bayes defined over n boolean 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-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 $0 < \varepsilon < 1/8$, 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 δ, L outputs a hypothesis with $error_{\mathcal{D}}(h) > \epsilon$

Shatter coefficient H[m]

for $S \subseteq X$, where $S = \{x_1 \ldots x_m\}$, define $H(S)$ as the set of distinct labelings of S induced by H

$$H(S) \equiv \{\langle h(x_1) \ldots, h(x_m)\rangle \mid h \in H)\}$$

and define $H[m]$ as the maximum number of ways to label m instances of X

$$H[m] \equiv \max_{S \subseteq X, |S|=m} |H(S)|$$

If H can shatter a subset of size m, then $H[m] =$

Note $VCdim(H) \equiv$ largest m for which $H[m] = 2^m$

Shatter coefficient H[m]

Sauer's Lemma: Let $VCdim(H) = d$. Then

1. for all m, $H[m] \leq \Phi_d(m)$, where $\Phi_d(m) \equiv \sum_{i=0}^{d} \binom{m}{i}$
2. for $m > d$,

$$\Phi_d(m) \leq (1+m)^d$$

$$\Phi_d(m) \leq \left(\frac{em}{d}\right)^d$$

Sample Complexity - Summary

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

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

|H|

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

VC(H)

$$m > \frac{2}{\epsilon}(\log_2(1/\delta) + \log_2(3\ H[2m]))$$

H[m]

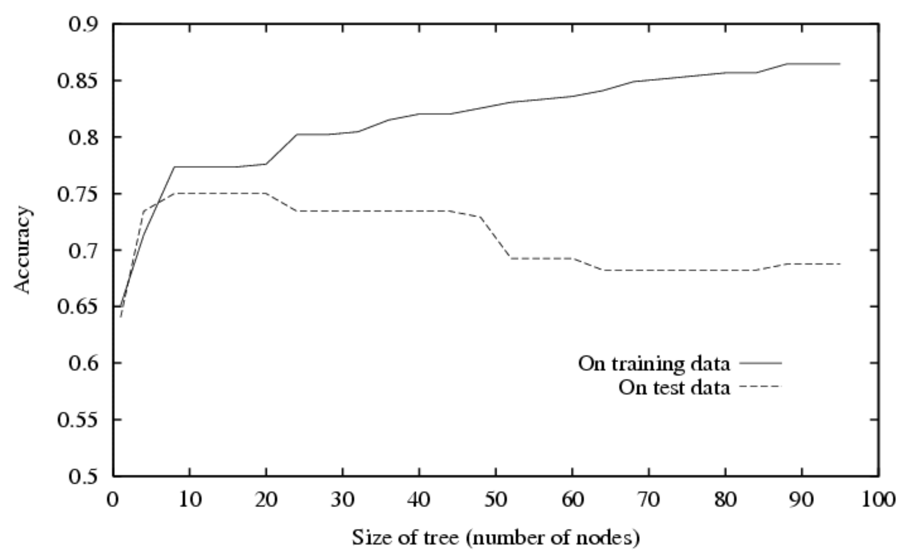
* also Rademacher complexity

Agnostic Learning: VC Bounds

[Schölkopf and Smola, 2002]

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

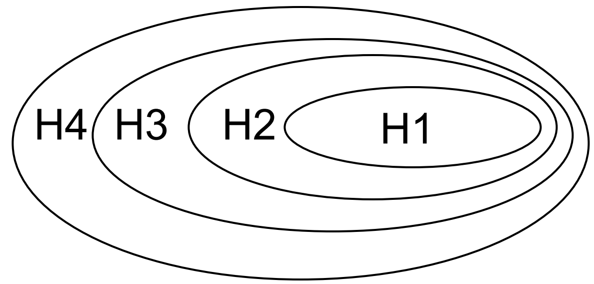
$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}$$



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...

Mistake Bounds

So far: how many examples needed to learn?

What about: how many mistakes before convergence?

Let's consider similar setting to PAC learning:

- Instances drawn at random from X according to distribution $\mathcal{D}$
- Learner must classify each instance before receiving correct classification from teacher
- Can we bound the number of mistakes learner makes before converging?

Mistake Bounds: Find-S

$x = < x_1, \ldots, x_n >, \; x_i \in \{0, 1\}$

Consider Find-S when H = conjunction of boolean literals

FIND-S:

- Initialize h to the most specific hypothesis $x_1 \wedge \neg x_1 \wedge x_2 \wedge \neg x_2 \wedge \ldots \neg x_n \rightarrow y = 1$ else $y = 0$
- For each positive training instance x
  - Remove from h any literal that is not satisfied by x
- Output hypothesis h.

How many mistakes before converging to correct h?

Mistake Bounds: Halving Algorithm

1. Initialize VS ← H
2. For each training example,
   - remove from VS every hypothesis that misclassifies this example

Consider the Halving Algorithm:

- Learn concept using version space CANDIDATE-ELIMINATION algorithm
- Classify new instances by majority vote of version space members

How many mistakes before converging to correct h?

- ... in worst case?
- ... in best case?

Optimal Mistake Bounds

Let $M_A(C)$ be the max number of mistakes made by algorithm A to learn concepts in C. (maximum over all possible $c \in C$, and all possible training sequences)

$$M_A(C) \equiv \max_{c \in C} M_A(c)$$

Definition: Let C be an arbitrary non-empty concept class. The **optimal mistake bound** for C, denoted $Opt(C)$, is the minimum over all possible learning algorithms A of $M_A(C)$.

$$Opt(C) \equiv \min_{A \in learning\ algorithms} M_A(C)$$

$$\boxed{VC(C) \leq Opt(C) \leq M_{Halving}(C) \leq log_2(|C|).}$$

Weighted Majority Algorithm

a_i denotes the i^{th} prediction algorithm in the pool A of algorithms. w_i denotes the weight associated with a_i.

– For all i initialize $w_i \leftarrow 1$

– For each training example $\langle x, c(x) \rangle$

  * Initialize q_0 and q_1 to 0
  * For each prediction algorithm a_i
    · If $a_i(x) = 0$ then $q_0 \leftarrow q_0 + w_i$
      If $a_i(x) = 1$ then $q_1 \leftarrow q_1 + w_i$
  * If $q_1 > q_0$ then predict $c(x) = 1$
    If $q_0 > q_1$ then predict $c(x) = 0$
    If $q_1 = q_0$ then predict 0 or 1 at random for $c(x)$
  * For each prediction algorithm a_i in A do
    If $a_i(x) \neq c(x)$ then $w_i \leftarrow \beta w_i$

when β=0, equivalent to the Halving algorithm…

Weighted Majority

Even algorithms that learn or change over time...

[Relative mistake bound for WEIGHTED-MAJORITY] Let D be any sequence of training examples, let A be any set of n prediction algorithms, and let k be the minimum number of mistakes made by any algorithm in A for the training sequence D. Then the number of mistakes over D made by the WEIGHTED-MAJORITY algorithm using $\beta = \frac{1}{2}$ is at most

$$2.4(k + \log_2 n)$$