Machine Learning

CSE427

Mahbub Majumdar
Typeset by:
Syeda Ramisa Fariha

BRAC University
66 Mohakhali
Dhaka, Bangladesh

July 30, 2018

Superhumungous Thanks

These slides were typeset by Syeda Ramisa Fariha.

Without her tremendous dedication, these slides would not exist.

Introduction

- **Goal:** Understand which $\mathcal{H}$ are PAC Learnable
- So far, we saw that,
  - *Finite $\mathcal{H}$ are Learnable.*
  - *Class of all functions over an infinite size domains are not learnable.*
- Examples where, infinite size classes are learnable:
  - *aligned axis predictors*
  - *circle predictors,...*
- Thus, *finiteness* doesn't seem to be a necessary condition for learnability
- We want a better measure of learnability

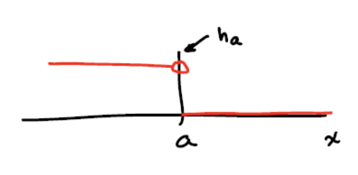
Introduction

- in 1971, *Vapnik and Chervonenkis* in the context of statistics invented the *"VC"* dimension idea
- This was applied to PAC Learning theory by Blomer, etc., in 1989.

Infinite-Size Classes Can Be Learnable

- Let's first show that infinite size classes can be learnable
- Simplest example: *Threshold function*
- Example: $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$ where, $h_a : \mathbb{R} \to \{0, 1\}$ such that

$$h_a(\chi) = \underbrace{\mathbb{I}(\chi < a)}_{\text{Indicator Function}}$$



- Since ∞ possible choices for a, $\mathcal{H}$ is infinite

Infinite-Size Classes Can Be Learnable

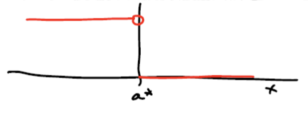
- Let's show that $\mathcal{H}$ is *PAC learnable* with sample complexity

$$m_{\mathcal{H}} \leq \frac{\log \frac{2}{\epsilon}}{\epsilon}$$

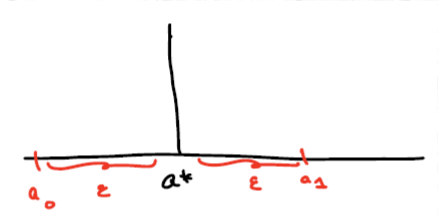
- Let $a^* =$ threshold such that

$$h^*(\chi) = \mathbb{I}(\chi < a^*)$$

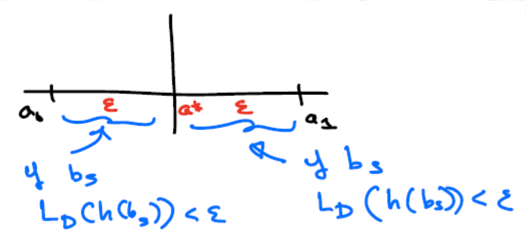
This is the true labeling function and $L_D(h^*) = 0$



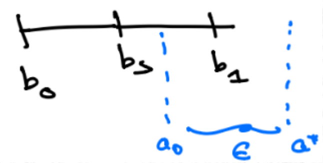
- Now, how can you get an error $< \epsilon$?



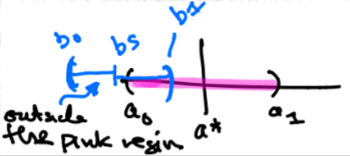
- If the predicted value b_s falls in the region (a_o, a_*) or (a_*, a_1) then the error $< \epsilon$.



- Let $b_0 < b_s < b_1$ be a similar ϵ sized region around b_s just as $a_0 < a^* < a_1$ is an ϵ sized region



- The idea is a^* is associated to the sandwich region (a_0, a_1)
- b_s is associated to the sandwich region (b_0, b_1)
- If $b_s < a_0$ then error $> \epsilon$



Infinite-Size Classes Can Be Learnable

- Similarly if $b_s > a_1$, then error $> \epsilon$

$$\underset{S \sim D^m}{\mathbb{P}}\Big[L_D(h_S) > \epsilon\Big] \leq \underset{S \sim D^m}{\mathbb{P}}\Big[b_s < a_0 \text{ or } b_s > a_1\Big]$$

- Now, use the *Union Bound*,

$$\begin{aligned}\underset{S \sim D^m}{\mathbb{P}}\Big[b_s < a_0 \text{ or } b_s > a_1\Big] &\leq \underset{S \sim D^m}{\mathbb{P}}\Big[b_s < a_0\Big] + \underset{S \sim D^m}{\mathbb{P}}\Big[b_s > a_1\Big] \\ &= (1-\epsilon)^m + (1-\epsilon)^m \\ &= 2(1-\epsilon)^m \\ &\leq 2e^{-\epsilon m}\end{aligned}$$

Infinite-Size Classes Can Be Learnable

- Setting this equal to δ,

$$2e^{-\epsilon m} = \delta$$
$$\Rightarrow m = \frac{1}{\epsilon} \ln \frac{2}{\delta}$$

Thus, if $m \geq \frac{1}{\epsilon} \ln \frac{2}{\delta}$ then $\mathcal{H}$ is *PAC Learnable*.

VC dimension of $\mathcal{H}$ correctly measures it's learnability

Motivation: No Free Lunch Theorem

$$\mathcal{H} = \{\text{all possible functions from } C \subseteq X \text{ to } \{0,1\}\}$$

- If choose no more than $\frac{|C|}{2}$ sample points $\rightarrow$ *average error* $\geq \frac{1}{4}$
- There exists an h with zero error, since $\mathcal{H}$ includes all possible functions from $C \rightarrow \{0,1\}$ in $\mathcal{H}$.
- In the special case $|C| = \infty$, we need ∞ samples for *PAC Learnability*
- Infinite sized $\mathcal{H}$ like this, where all possible hypotheses are included, are not PAC Learnable

VC Dimension

- If we consider distributions like this, that are concentrated on $C \subseteq X$, we should study how X behaves on C.

Definition: Restriction of $\mathcal{H}$ to C

Let $\mathcal{H}$ be the class of functions from X to $\{0, 1\}$. Let

$$C = \{C_1, C_2, \ldots, C_n\} \subset X$$

The restriction of $\mathcal{H}$ to C is the set of functions from C to $\{0, 1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_c = \{h(C_1), h(C_2), \ldots, h(C_m) \mid h \in \mathcal{H}\}$$

where we can represent each function from $C \to \{0, 1\}$ as a vector in $\{0, 1\}^{|C|}$

Shattering

- **Example: Two Points**

  Suppose, $C = \{0, 1\}$ is just two points,

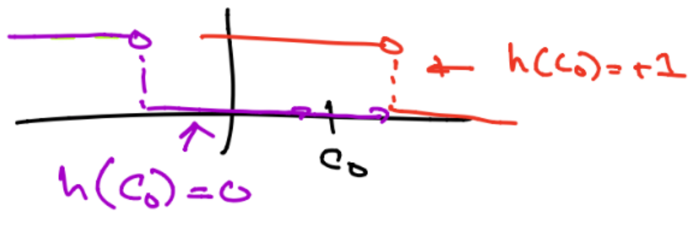
  $$\mathcal{H}_C = \{(0,0), (0,1), (1,0), (1,1)\}$$

  where each set of 0's and 1's is a *hypothesis*.

- If the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$, then we say, $\mathcal{H}$ shatters C

- **Example: One Point**

Suppose $C = \{C_0\}$. Take $\mathcal{H}$ = *Class of all hypothesis functions.*



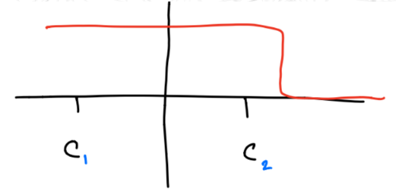
Thus, C_0 gets assigned all possible values. We say that the class of threshold functions shatters $C = \{C_0\}$

Definition: Set Shattering

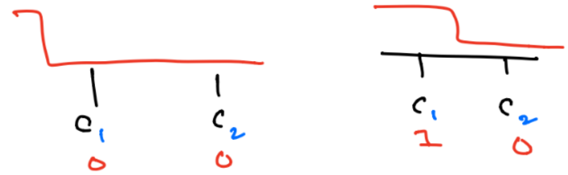
A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset X$, if the restriction of $\mathcal{H}$ to C is the set of all functions from $C \to \{0,1\}$, i.e. $|\mathcal{H}_C| = 2^{|C|}$.

Example: Two points

Let $C = \{C_1, C_2 \mid C_1 \leq C_2\}$. Let's look at the possible labelings with threshold functions in $\mathcal{H}$.



Shattering Examples



- Question: Can we get the labeling $\{0, 1\}$?
- Is $C = \{C_1, C_2\}$ shattered by $\mathcal{H}$?
- Answer: Any labeling assigning 1 to C_2 will assign 1 to C_1 also

$\Rightarrow$ C is not shattered by $\mathcal{H}$

Shattering

- Going back to the *NFL Theorem*:

$\rightarrow$ Whenever some c is shattered by $\mathcal{H}$

$\rightarrow$ the adversary is not restricted by $\mathcal{H}$

$\rightarrow$ the adversary can choose a perfect labeling function and

$\rightarrow$ the adversary can achieve zero loss

Shattered Sets and NFL Theorem

Corollary:

Let $\mathcal{H}$ be a hypothesis class of functions from X to $\{0,1\}$. Let m be the training set size. Assume there exists a set $C \subset X$ of size $2m$ that is shattered by $\mathcal{H}$. Then for any learning algorithm A, there exists a distribution D over $X \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $L_D(h) = 0$, but with probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$, we have

$$L_D(A(S)) \geq \frac{1}{8}$$

- Thus if $\mathcal{H}$ shatters C, with $C = |2m|$,

$\rightarrow$ then we can't learn $\mathcal{H}$ using m samples.

Shattered Sets and NFL Theorem

- Intuitively, if C shatters $\mathcal{H}$,

$\rightarrow$ a sample of less than half the instances of C, gives no information about the other half of C.

- Any labeling of the rest of the instances in C can be explained by an $h \in \mathcal{H}$

VC Dimension

Definition: VC Dimension

The VC Dimension of $\mathcal{H}$ is the maximal size of a set $C \subset X$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ shatters sets of arbitrarily large size, then VC-dim$(\mathcal{H}) \equiv \infty$.

Theorem: Infinite VC Dimension

If VC-dim$(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC Learnable

Proof: Every set of C of size $2m$ is shattered by $\mathcal{H}$, where m is the size of the training set. Now apply NFL to C.

Comments

- It turns out that if VC dim($\mathcal{H} < \infty$), then $\mathcal{H}$ is learnable.

- In order to find the VC-Dim of $\mathcal{H}$, we need to show that
  a) There exists a C of size d, that is shattered by $\mathcal{H}$
  b) Every set C of size $d + 1$ is not shattered by $\mathcal{H}$

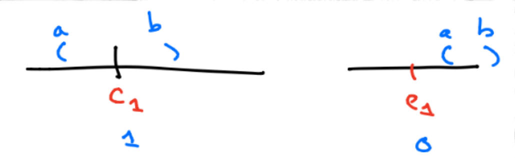
Example: Threshold Functions

- Let's calculate the VC-dim for Threshold Functions.
- $C = \{C_1\}$ is shattered VC-dim$(\mathcal{H}) \geq 1$
- $C = \{C_1, C_2\}$ is *not* shattered. Therefore, VC-dim$(\mathcal{H}) < 2$
- Thus VC-dim$(\mathcal{H}_{threshold}) = 1$
- Thus, it makes sense that the infinite class $\mathcal{H}_{threshold}$ is PAC Learnable

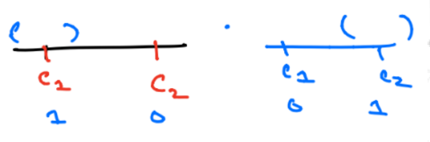
- Let's calculate the VC-dim of Intervals.

$$h_{ab}(x) = \begin{cases} 1 & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

- $C = \{C_1\}$ is shattered by $\mathcal{H}_{ab}$



- $C = \{C_1, C_2\}$ is shattered

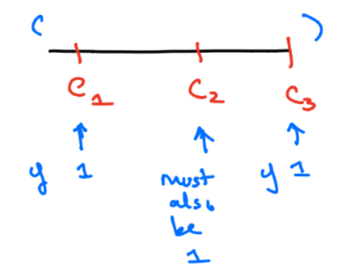


Example: Intervals

- Thus, VC-dim$(\mathcal{H}_{ab}) \geq 2$

Example: Intervals

- $C = \{C_1, C_2, C_3\}$



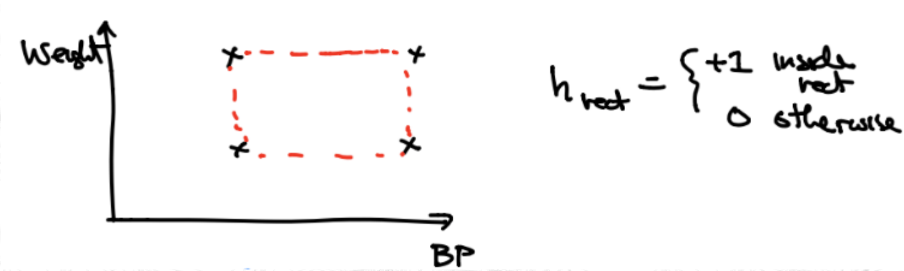
- If $h(C_1) = 1, h(C_3) = 1 \Rightarrow h(C_2) = 1$

- If $h(C_1) = 1, h(C_3) = 1 \Rightarrow h(C_2) = 1$, thus $(1, 0, 1)$ is not possible

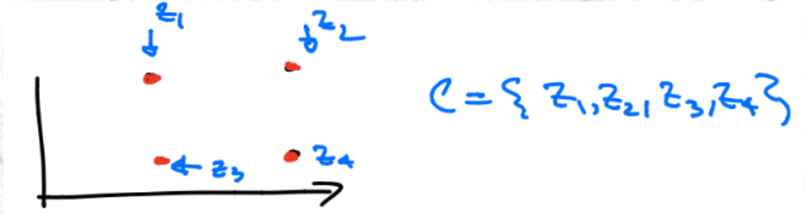
$\Rightarrow$ $VC\ dim(\mathcal{H}_{ab}) < 3$

$\Rightarrow$ $VC\ dim(\mathcal{H}_{ab}) = 2$

Recall, that aligned-axis rectangles look like:

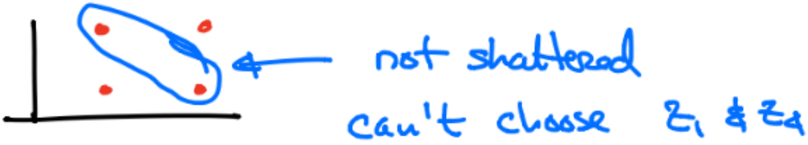


Example: Axis Aligned Rectangles



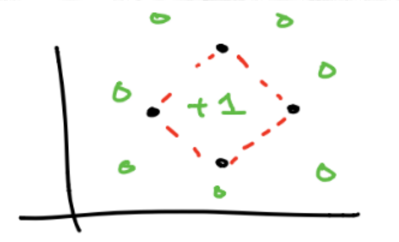
- Label the corner points as $\{z_1\}, \{z_2\}, \{z_3\}, \{z_4\}$
- $\{z_1, z_2\}, \{z_2, z_3\}, \{z_3, z_4\}, \{z_2, z_4\}$ are all shattered

- But, $\{z_1, z_4\}$ and $\{z_2, z_3\}$ are not shattered



Example: Axis Aligned Rectangles

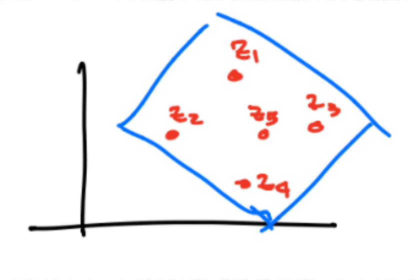
- Solution: rotate rectangle.



- Now, see VC-dim($\mathcal{H}_{rect}$) ≥ 4

Example: Axis Aligned Rectangles

- But if $C = \{z_1, z_2, z_3, z_4, z_5\}$



- The point z_5 is always captured along with the rest of the points.

$\Rightarrow$ VC-dim$(\mathcal{H}_{rect}) < 5$

$\Rightarrow$ VC-dim$(\mathcal{H}_{rect}) = 4$

Remarks: Finite Classes

- Suppose $\mathcal{H}$ is finite.
- Suppose $|C| = k$. This means C has k points.
- Suppose, the labels are binary. Then $|\mathcal{H}| = 2^k$.
- Suppose VC-dim$(\mathcal{H}) = d$.
- The VC-dim of $\mathcal{H}$ restricted to C, can be at most k.
- Thus, $d \leq k$.
- This means
$$|\mathcal{H}| \geq 2^d$$
- Or equivalently, taking logarithms
$$\log_2 |\mathcal{H}| \geq \text{VC-dim}(\mathcal{H}) = d$$

Remarks: Finite Classes

- C can't be shattered if $d < 2^{|C|}$
- Will prove that classes with finite VC dimension are learnable
- Example: Threshold functions: $\mathcal{H}_a = \{h_a \mid a \in \mathbb{R}\}$ is an infinite class. but it has finite VC-dim$(\mathcal{H}_a) = 1$

$\Rightarrow$ Threshold function class is *learnable*

VC Dimension and the Number of Parameters

- It looks like VC-dim $=$ number of parameters, example: threshold functions parameter by $a \in \mathbb{R}$
- But not always true that number of parameters $=$ VC dimension

The Fundamental Theorem of Statistical Learning

- We saw that infinite VC dimension class is not learnable.
- Opposite statement is also true. This gives rise to:

Theorem 6.7 (The Fundamental Theorem of Statistical Learning). *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the* $0-1$ *loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

- VC Dimension determines *Learnability and Sample Complexity*

The Fundamental Theorem of Statistical Learning - Quantitative Version

Theorem 6.8 (The Fundamental Theorem of Statistical Learning – Quantitative Version). *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the* 0−1 *loss. Assume that* $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. *Then, there are absolute constants C_1, C_2 such that*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}^{UC}(\epsilon,\delta) \le C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}(\epsilon,\delta) \le C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

3. *$\mathcal{H}$ is PAC learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \le m_{\mathcal{H}}(\epsilon,\delta) \le C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

The proof of this theorem is given in Chapter 28.

Note: *Lower bound and upper bound are determined by the VC Dimension=d*

Proof of Theorem 6.7 **Proof:**

a) $1 \rightarrow 2$, We proved this in *Chapter 4*

b) $2 \rightarrow 3$, $3 \rightarrow 4$, $2 \rightarrow 5$ are clearly true

c) $4 \rightarrow 6$, $5 \rightarrow 6$ follow from *NFL Theorem*

- Difficult part is $1 \rightarrow 6$

Comments:

- Suppose $VC\ dim(\mathcal{H}) = d$, $C \subseteq \chi$
- Then effective size of $\mathcal{H}_C$ is $|\mathcal{H}_C|$

  $|\mathcal{H}_C| = \mathcal{O}(|C|^d)$
- For binary predictors, $|\mathcal{H}_C| = |C|^d$
- As $d = VC\ dim(\mathcal{H})$ grows, m increases

Proof of Theorem 6.7

Lower Bound Analysis For Agnostic PAC Learning

- As $d = VC\ dim(\mathcal{H})$ grows, m increases
- As $\epsilon \to 0$, $m \to 0$
- As $\delta \to 0$, $m \to 0$, but *logarithmically* only
- Look at *(1)* in Theorem 6.8, since
  ERM + Uniform Convergence = PAC Corollary 4.4, page 32, (1)
  says all you need is an ERM Learning Paradigm
- However, as we will see to apply ERM $\to$ need to minimize error over all $h \in \mathcal{H}$
- But then we have to worry about *computational complexity*

$\Rightarrow$ How long will it take to calculate the error for every $h \in \mathcal{H}$

$\Rightarrow$ ∞ time if $|\mathcal{H}| = \infty$

Lower Bound Analysis For Agnostic PAC Learning

- For *agnostic PAC Learning*,

$$\Rightarrow C_1 \frac{d + log\frac{1}{\delta}}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + log\frac{1}{\delta}}{\epsilon^2}$$

While for *PAC*,

$$\Rightarrow C_1 \frac{d + log\frac{1}{\delta}}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{dlog\frac{1}{\epsilon} + log\frac{1}{\delta}}{\epsilon}$$

  - This is because for agnostic PAC, there is no guarantee that you will succeed
  - Therefore, we need more samples to get an accurate, reliable predictor

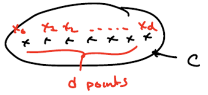
Proof of Theorem 6.7

Lower Bound Analysis For Agnostic PAC Learning

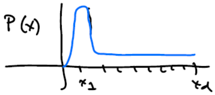
- first consider lower bound
- NFL tells us that
$$\frac{d}{2} = \frac{|c|}{2} < m_{\mathcal{H}}(\epsilon, \delta)$$
- How does the ϵ - dependence of the lower bound come in ?
- $\mathcal{H}$ should work for any distribution on C. Choose the most difficult distribution to predict from.

Lower Bound Analysis For Agnostic PAC Learning

- choose the largest set C that is shattered by $\mathcal{H}$. It has $dim = d$

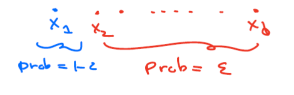


- This set will be very hard to learn from if you keep sampling the same point



Lower Bound Analysis For Agnostic PAC Learning

- Let the probability of sampling x_1 be ϵ, $P(x_1) = 1 - \epsilon, P(x_i = \frac{\epsilon}{d-1})$ $P(x\ outside\ C) = 0$
-



- Then to learn from x_2.........x_d need to sample at least half the points, which is $\frac{d-1}{2}$
- Let m be number of points sample from $x_2, \ldots, x_d$
- the expected number of points in the $\epsilon-$ region is :

$$m\epsilon$$

- Thus need

$$m\epsilon > \frac{d-1}{2}$$

Proof of Theorem 6.7

Lower Bound Analysis For Agnostic PAC Learning

$$\Rightarrow m > \frac{d-1}{2\epsilon}$$

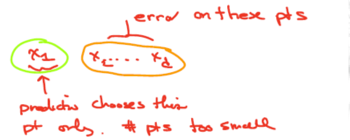
$$m > \mathcal{O}(\frac{d}{\epsilon})$$

- Thus we expect for the PAC case, the lower bound to be $\mathcal{O}(\frac{d}{\epsilon})$
- Can check this with a more complete derivation (Chapter 28)

Note: error for less than $\frac{d-1}{2}$ point is $\frac{1}{4} \times \epsilon = \frac{\epsilon}{4}$
Note: $\underset{S\sim}{\mathbb{E}}(L_D(A(S))) \geq \frac{\epsilon}{4}$ of $m \leq \frac{d-1}{2\epsilon}$ for any ϵ

- think about the points x_2.........x_d as the points where the classifier makes a mistake

Lower Bound Analysis For Agnostic PAC Learning

- 

  Now give argument for Agnostic PAC that lower bound $\sim \theta(\frac{1}{\epsilon^2})$
- **Why are things harder in Agnostic case ?**
- Recall, PAC case one h has zero error. Now no such assumption
- Consider a Binomial process
  $\rightarrow$ like trying to figure out the probability of candidate A getting a vote instead of candidate B

Proof of Theorem 6.7

Lower Bound Analysis For Agnostic PAC Learning

- 2 candidates $A$and B
- want to predict probability of winning p
- want accuracy ϵ
  $\Rightarrow$ thus, $\hat{p} - \epsilon \leq p \leq \hat{p} + \epsilon$
  $\Rightarrow$ here $\hat{p}$ is the estimate of p
- For binomial process

$$\mathrm{Var}(x) = p(1-p)$$

$$\text{Average Var } = \frac{p(1-p)}{n}$$

$$\text{std-dev } = \frac{\sqrt{p(1-p)}}{n}$$

Lower Bound Analysis For Agnostic PAC Learning

- graph of $p(1-p) =$

- max of $p(1-p)$ is at $p = \frac{1}{2}$

$$\sqrt{p(1-p)} = \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}$$

- consider the 2 sigma range

Proof of Theorem 6.7

Lower Bound Analysis For Agnostic PAC Learning 2σ range is

$\hat{p} \pm \frac{2\sqrt{p(1-p)}}{\sqrt{n}}$

$\Rightarrow \hat{p} \pm \frac{1}{\sqrt{n}}$

- accuracy $= \frac{1}{\sqrt{n}}$

$$\epsilon = \frac{1}{\sqrt{n}} \Rightarrow n = \frac{1}{\epsilon^2}$$

- Thus for more than $\frac{1}{\epsilon^2}$ points accuracy $\leq \epsilon$
- That's why for Agnostic PAC lower bound $\sim \frac{1}{\epsilon^2}$

Proof of Theorem 6.7

Upper Bound Analysis For Agnostic PAC Learning:
Sauer-Shelah-Perles Lemma

- Lower bound analysis is less interesting because it doesn't tell you how many samples you need for machine learning
- It just says the number of samples is at least ..... this many
- Lower bound analysis uses probability theory. Doesn't use the VC dimension. Only way we used VC dimension $\Rightarrow$ number of points in a set that has all possible behaviours.
- Upper bounds depend on a very special property of VC dimension $\Rightarrow$ Sauer Lemma

Upper Bound Analysis For Agnostic PAC Learning:
Sauer-Shelah-Perles Lemma

- Define the Growth function

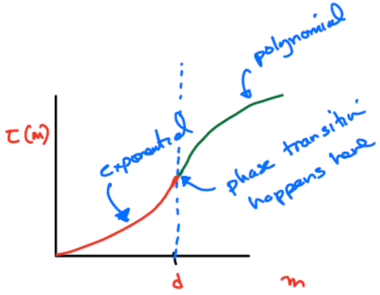
$$\tau_{\mathcal{H}}(m) = \max_{C \subset X, |C| = m} \left(|\mathcal{H}_C| \right)$$

- $\tau_{\mathcal{H}}$ is the maximum number of different functions from C which has size m to $\{0, 1\}$
- $\mathcal{H}_C$ = class restricted to C
- If $\text{VCdim}(\mathcal{H}) = d$ then for any $m \leq d$

$$\tau_{\mathcal{H}}(m) = 2^m$$

Upper Bound Analysis For Agnostic PAC Learning:
Sauer-Shelah-Perles Lemma

- Then $\mathcal{H}_C$ includes all possible function from $C \to \{0, 1\}$
- Sauer lemma tells us what happens when $m > d$



Sauer-Shelah-Perles Lemma

Sauer-Shelah-Perles Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) \leq d < \infty$. Then for all m

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

In particular if $m > d$ then

$$\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d} \sim m^{d}$$

Proof of Sauer-Shelah-Perles Lemma

Proof:

- Let $C = \{c_1, ...c_m\}$
- It's enough to prove for all $\mathcal{H}$

$$|\mathcal{H}_C| \leq \left| \left\{ \begin{array}{l} \textit{all the subsets } B \textit{ of } C, \\ \textit{that are shattered by } \mathcal{H} \end{array} \right\} \right|$$

  more concisely,

$$|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$$

- This is sufficient to prove Sauer because if
  a $VCdim(\mathcal{H}) \leq d$
  then no set with size $> d$ is shattered by $\mathcal{H}$

Proof of Sauer-Shelah-Perles Lemma

Therefore

$$|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$$

$\leq$ number of subsets of size up to d of C, where

$$|C| = m$$
$$= \sum_{i=0}^{d} \binom{m}{i}$$

- when $m > d$ then

$$\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$$

(thus this is a technical algebraic fact $\rightarrow$ Appendix: A-5)

Proof of Sauer-Shelah-Perles Lemma

- Now to prove $|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$, use induction
- $m = 0 \rightarrow$ one point

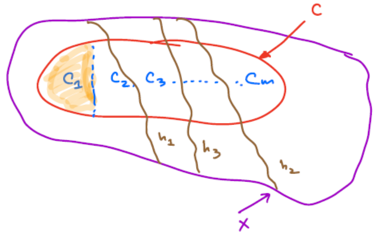
$$|\mathcal{H}_C| = 1$$

$$|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}| = |\phi| = 1$$

- If start induction at $m = 1$ then $|\mathcal{H}_C = 2|$,
  $\{B \subseteq C \mid B \; \textit{shattered}\} = 2$
- **Induction Step :**
  Assume $|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$ holds for sets of size $k < m$
- **Need to prove :**
  $|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$ holds for sets of size m

Proof of Sauer-Shelah-Perles Lemma

- fix $\mathcal{H}$
- $C = \{c_1 .....c_m\}$
  $C' = \{c_2 .....c_m\}$



Let Y_0 be the set of the functions $\{c_2, ...c_m\}$

Proof of Sauer-Shelah-Perles Lemma

$$Y_0 = \{\text{the functions in } \mathcal{H} \text{ restricted to } \mathcal{H}_{C'}\}$$
$$= \text{ these are the functions which map}$$

$$\{f : (c_2....c_m) \to \{0,1\}\}$$

- these functions are $m-1$ dimensional vectors

$$\begin{pmatrix} h(c_2) \\ h_(c_3) \\ . \\ . \\ . \\ h(c_m) \end{pmatrix}$$

Proof of Sauer-Shelah-Perles Lemma

- Now define

$$Y_1 = \left\{ \begin{array}{l} \text{the set of all functions on } C' \\ \text{that can be extended to } c_1 \end{array} \right\}$$

$$= \left\{ (y_2, y_3 \ldots y_m) \mid \begin{array}{l} (0, y_2 \ldots y_m) \epsilon \mathcal{H}_C \text{ and} \\ (1, y_2 \ldots y_m) \epsilon \mathcal{H}_C \end{array} \right\}$$

- Then

$$|\mathcal{H}_C| = |Y_0| + |Y_1|$$

Here,
$Y_0 =$ the number of functions living only in C'
$Y_1 =$ the number of functions in C'that can be extended to c_1

Proof of Sauer-Shelah-Perles Lemma

- Now using the induction step

$$|Y_0| = |\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$$

$$= |\{B \subseteq C' \mid c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}|$$

- Now define $\mathcal{H}' \subseteq \mathcal{H}$ to be the pairs of functions that agree on C' but differ on c_1

$$\begin{aligned} \mathcal{H}' = \{h, h' \in \mathcal{H} &\text{ such that} \\ &(1 - h'(c_1), h'(c_2) ... h'(c_m)) \\ = &\ (h(c_1), h(c_2) .... h(c_m)) \\ &(h(c_1), h(c_2), ... h(c_m))\} \end{aligned}$$

Proof of Sauer-Shelah-Perles Lemma

- Therefore,
  $\Rightarrow$ if $\mathcal{H}'$ shatters $B \subseteq C$, then it also shatters $\{c_1\} \cup B$ and *vice versa*
- Now

$$Y_1 = \mathcal{H}'_{C'}$$

- Therefore :

$$|Y_1| = |\mathcal{H}'_{C'}| \leq \left| \left\{ B \subseteq C' \mid \mathcal{H} \text{ shatters } B \right\} \right|$$

$$= \left| \left\{ B \subseteq C' \mid \mathcal{H} \text{ shatters } B \cup \{c_1\} \right\} \right|$$

$$= \left| \left\{ B \subseteq C \mid c_1 \in B \text{ and } \mathcal{H}' \text{ shatters } B \right\} \right|$$

$$\leq \left| \{ B \subseteq C \mid c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B \} \right|$$

Proof of Theorem 6.7

Proof of Sauer-Shelah-Perles Lemma

- Putting everything together

$$
\begin{aligned}
|\mathcal{H}_C| &= |Y_0| + |Y_1| \\
&\le |\{B \subseteq C \mid c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| \\
&\le |\{B \subseteq C \mid c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| \\
&= |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|
\end{aligned}
$$

This is what we wanted to prove !

Proof of Sauer-Shelah-Perles Lemma

- **What are the consequences of this?**
  $\tau_{\mathcal{H}}(m) = \max\limits_{|C|=m} |\mathcal{H}_C|$ The max is taken over all sets $C \subseteq \chi$ such that $|C| = m$
- Note different C with $|C| = m$ may have different $|\mathcal{H}_C|$
  Example: $\mathcal{H} =$ *aligned axis rectangle*

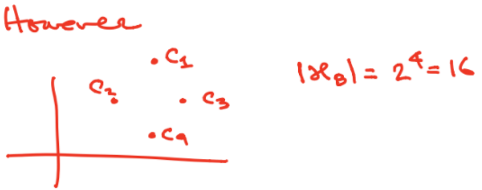
$C = \{c_1, c_2, c_3, c_4\}$

$|\mathcal{H}_B| = 2^4 - 2 = 14$

because and $\emptyset$ not possible

Proof of Theorem 6.7

Proof of Sauer-Shelah-Perles Lemma



- If $\tau_{\mathcal{H}}(m) \sim m^d$ for $m > d$, then
- $$D^m(S|L_{D,f}(h) > \epsilon) leq |\mathcal{H}_B| e^{-\epsilon m} \tag{1}$$

  where $|\mathcal{H}_B| e^{-\epsilon m} = \textit{number of bad hypothesis}$

Proof of Sauer-Shelah-Perles Lemma

- Recall that,

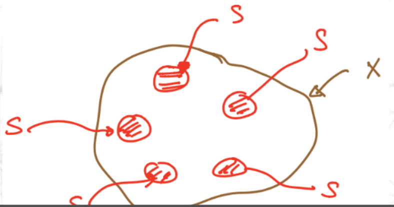
  $|\mathcal{H}_B| =$

$$\left\{ \begin{array}{c} \text{number of bad hypothesis} \\ \text{generated by sampling from} \\ \text{misleading training sets} \end{array} \right\}$$

- How do you choose a training set?

  $\Rightarrow$ Select a $|C| = m$ subset of χ

  $\Rightarrow$ Therefore, the number of ways to choose a bad hypothesis is $|\mathcal{H}_B| \leq |\mathcal{H}|_C$, where $C = S$



Proof of Sauer-Shelah-Perles Lemma

- $$\max_{\substack{S\subseteq\chi \\ |S|=m}} |\mathcal{H}_B| \leq \max_{\substack{S\subseteq\chi \\ |S|=m}} |\mathcal{H}|_S = \tau(m)$$

- Therefore equation (1) becomes,

$$D^m(S|L_{D,f}(h) > \epsilon) leq \tau(m) e^{-m\epsilon}$$

- Thus we see that, if

$$\tau(m) \sim polynomial$$

Then

$$\tau(m)e^{-m\epsilon} \Rightarrow 0$$

as $m \Rightarrow \infty$

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size Now lets go back to proving $1 \to 6$ in **Theorem 6.7**

- to prove this we need to show that finite VCdim leads to uniform convergence
- Recall that if a class $\mathcal{H}$ is uniformly convergent if for every distribution D , there is a $m_{\mathcal{H}}(\epsilon, \delta)$ such that S is ϵ-representative
- And recall that S is ϵ-representative if for every $h \epsilon \mathcal{H}$

$$|L_S(h) - L_D(h)| \leq \epsilon$$

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size

- Since $|L_S(h) - L_D(h)| \leq \epsilon$ should hold for every $h \in \mathcal{H}$, choose the h fro which

$$|L_S(h) - L_D(h)|$$

  is largest
- Formally, this is

$$\underset{h \in \mathcal{H}}{\text{SUP}} |L_S(h) - L_D(h)|$$

  SUP $\equiv$ least upper hand

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size The expected max of $|L_S(h) - L_D(h)|$ is

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{h\in\mathcal{H}}|L_D(h) - L_S(h)|\right]$$

- One thing we have been doing up to this point is
- We constructed our prediction after seeing the data
- This is generally an illegal move in statistics

Uniform Convergence for Classes of Small Effective Size

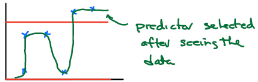
- For example
  a. Suppose I construct predictors
     $f_0 =$ all zero
     $f_1 =$ all one
     Then I evaluate the error by evaluating it over the data

Then we evaluate the true error

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size

- b Suppose, instead we look at the data first and the construct select our predictors. Then we might select



- The predictor that is selected will then look overfitted and the true error will not really be zero, even if it appears so.
- One way to select the predictor after seeing the data is to evaluate the predictor's true error $L_D(h)$ on a data set disjoint/independent from the data set S
  This is call the **double-sampling trick**

Uniform Convergence for Classes of Small Effective Size

- Evaluate h on a set S'

$$L_D(h) = \mathop{\mathbb{E}}_{S' \sim D^m} (L_{S'}(h))$$

then

$$\mathop{\mathbb{E}}_{S \sim D^m} \left[\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \right]$$

$$\mathop{\mathbb{E}}_{S \sim D^m} \left[\sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{S' \sim D^m} (L_{S'}(h)) - L_S(h) \right| \right]$$

$$\leq \mathop{\mathbb{E}}_{S \sim D^m} \sup_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{S' \sim D^m} |(L_{S'}(h)) - L_S(h)|$$

(using triangle inequality $|x + y + z| \leq |x| + |y| + |z|$)

Uniform Convergence for Classes of Small Effective Size

$$\leq \underset{S\sim D^m}{\mathbb{E}} \underset{S'\sim D^m}{\mathbb{E}} \sup_{h\in\mathcal{H}} |(L_{S'}(h)) - L_S(h)|$$

$$\text{(using } \sup \mathbb{E} \leq \mathbb{E} \ \sup\text{)}$$

- we have reduced this to a double sample over S and S'
- Since $L_{S'}(h)$ and $L_S(h)$ are empirical errors

$$\underset{S\sim D^m}{\mathbb{E}} \left(\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)| \right)$$

$$\leq \underset{S',S\sim D^m}{\mathbb{E}} \left(\sup_{h\in\mathcal{H}} |L_{S'}(h) - L_S(h)| \right)$$

$$= \underset{S',S\sim D^m}{\mathbb{E}} \left(\sup \frac{1}{m} \,|\sum_{i=1}^{m} \ell(h, z_i') - \ell(h, z_i) \right) \ldots\ldots (1)$$

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size

- the sample S

$$\{z'_1 = (x'_1, y'_1),\ z'_2 = (x'_2, y'_2)......, z'_m\}$$

and

$$\{z_1 = (x_1, y_1), ... \ z_m = (x_m, y_m)\}$$

are $2m$ vectors

- Our answer is symmetric in z_i and z_i' if we replace

$$z_i \leftrightarrow z_i'$$

then,

$$l(h, z_i') - l(h, z_i) \leftrightarrow -(l(h, z_i') - l(h, z_i))$$

We can write equation (1) as

$$\mathbb{E}_{S \in D^m} \sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)|$$

$$\leq \mathbb{E}_{S, S' \sim D^m} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \delta_i (l(h, z_i') - l(h, z_i))$$

Uniform Convergence for Classes of Small Effective Size

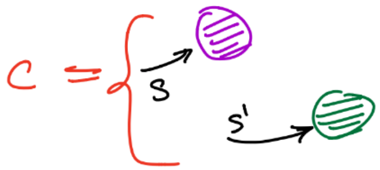
Here, $\delta_i = \pm 1$
$l(h, z_i) - l(h, z_i) = positive$
Here, The sign is isolated in $\delta_i = \{+1, -1\}$
Since we are looking for the *expected* value, we can overage over δ_i
Let the distribution for δ_i be $u_\pm$. Then we have equivalently,

$$\mathop{\mathbb{E}}_{\delta \sim u_\pm^m} \mathop{\mathbb{E}}_{S,S' \in D^m} SUP \frac{1}{m} |\sum_{i=1}^{m} \delta_i (l(h, z_i' - l(h, z_i))|$$

$$\mathop{\mathbb{E}}_{S,S' \in D^m} \mathop{\mathbb{E}}_{\delta \sim u_\pm^m} SUP \frac{1}{m} |\sum_{i=1}^{m} \delta_i (l(h, z_i' - l(h, z_i))|$$

Uniform Convergence for Classes of Small Effective Size

Now fix S and S. Let C be the set of S and S



- $|C| = 2m$
- Take the supremum over $h \epsilon \mathcal{H}_C$

Example of Supremum

-

$$G = \{\frac{1}{n} | n \in N\}$$

$$\text{SUP G} = \text{max value of the set G}$$
$$= \text{max G}$$

Uniform Convergence for Classes of Small Effective Size

- Sometimes the supremum of G isn't in the set G
  For example for,

$$G = \{X \mid X \in \mathbb{R} \text{ and } X < 1\}$$

  then $\text{SUP } G = 1,$ but $1 \notin G$
- when the SUP of G is in G then $\text{SUP } G = \max G$
- therefore

$$\mathop{\mathbb{E}}_{\delta \sim u_{\pm}^{m}} \left(\mathop{\text{SUP}}_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i} \right| \right)$$

$$\mathop{\mathbb{E}}_{\delta \sim u_{\pm}^{m}} \left(\mathop{max}_{h \in \mathcal{H}_C} \frac{1}{m} \left| \sum_{i} \right| \right)$$

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size

- now pick h in $\mathcal{H}_C$ and define

$$\theta_n = \frac{1}{m} \sum_{i=1}^{m} \sigma_i(\ell(h, z_i') - \ell(h, z_i))$$

- Since

$$\mathbb{E}(\theta_n) = 0$$

- θ_n is an average of variables
- range of $\theta_n \in [-1, +1]$
- we can then use Hoeffding's inequality for $e > 0$

$$\mathbb{P}[|\theta_n| > e] \leq 2e^{-2m\rho^2}$$

Uniform Convergence for Classes of Small Effective Size

- Apply the union bound

$$\mathbb{P}\left(\max_{h \in \mathcal{H}_C} |\theta_n| \geq \rho \right) \leq 2|\mathcal{H}_C| e^{-2m\rho^2}$$

- use Appendix A.4 if

$$\mathbb{P}(|X - X_0| > t) \leq 2be^{\frac{-t^2}{a^2}}$$

$$(\text{Here, } X_0 \in \mathbb{R}, t \geq 0, a > 0\ )$$

then

$$\mathbb{E}(|X - X_0|) \leq a\left(2 + \sqrt{\log b}\right)$$

$$(\text{Here, } b \geq e)$$

Uniform Convergence for Classes of Small Effective Size

- $$\underset{S\sim D^m}{\mathbb{E}}\left(\underset{h\in\mathcal{H}}{\text{SUP}}|L_D(h)-L_S(h)|\leq\frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}\right)$$
- Implications for m
  $\Rightarrow$ for $m > d$
  $$\tau_{\mathcal{H}}(2m)\leq\left(\frac{2em}{d}\right)^d$$
- $$\mathbb{E}(\text{SUP}|L_S(h)-L_D(h)|)$$
  $\rightarrow$ $=$ (probability of large$|L_S - L_D|$) $\times$ $|L_S - L_D|$

Uniform Convergence for Classes of Small Effective Size

- $$\delta|L_S(h) - L_D(h)| \leq \frac{4 + \sqrt{d \log\left(\frac{2em}{d}\right)}}{\sqrt{2m}}$$
- $$|L_S(h) - L_D(h)| \leq \frac{4 + \sqrt{d \log\left(\frac{2em}{d}\right)}}{\delta\sqrt{2m}}$$
- $$\text{Assume } \sqrt{d \log\left(\frac{2em}{d}\right)} \geq 4$$

$$\Rightarrow \ |L_S(h) - L_D(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d \log\left(\frac{2em}{d}\right)}{m}}$$

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size

- If

$$|L_S(h) - L_D(h)| \leq \epsilon$$

then

$$\frac{1}{\delta}\sqrt{\frac{2d\log\left(\frac{2em}{d}\right)}{m}} \leq \epsilon$$

$$\Rightarrow \quad \frac{1}{\delta^2}\frac{2d\log\frac{2em}{d}}{\epsilon^2} \leq m$$

Proof of Theorem 6.7

Uniform Convergence for Classes of Small Effective Size

- do some Algebra (lemma A.2)

$$m \geq 4\frac{2d}{(\epsilon\delta)^2}\log\frac{2d}{(\epsilon\delta)^2} + \frac{4d\log\frac{2e}{d}}{(\epsilon\delta)^2}$$

- Thus if $\mathrm{VCdim}(\mathcal{H}) = d < \infty$ then there is an m to give PAC learning