Advanced Machine Learning

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

University of Bucharest, 2nd semester, 2019-2020

Assignment 1

Deadline: Friday, 17th of April

1. **(0.5 points)** Consider $\mathcal{H} = \{h_{\theta_1} : \mathbb{R} \to \{0,1\}, h_{\theta_1}(x) = \mathbf{1}_{[x \geq \theta_1]}(x) = \mathbf{1}_{[\theta_1, +\infty)}(x), \theta_1 \in \mathbb{R}\} \cup \{h_{\theta_2}(x) = \mathbf{1}_{[x < \theta_2]}(x) = \mathbf{1}_{(-\infty, \theta_2)}(x), \theta_2 \in \mathbb{R}\}$. Compute VCdim($\mathcal{H}$).

2. **(0.75 points)** Consider $\mathcal{H}$ to be the class of all centered in origin sphere classifiers in the 3D space. A centered in origin sphere classifier in the 3D space is a classifier h_r that assigns the value 1 to a point if and only if it is inside the sphere with radius $r > 0$ and center given by the origin **O**(0,0,0).

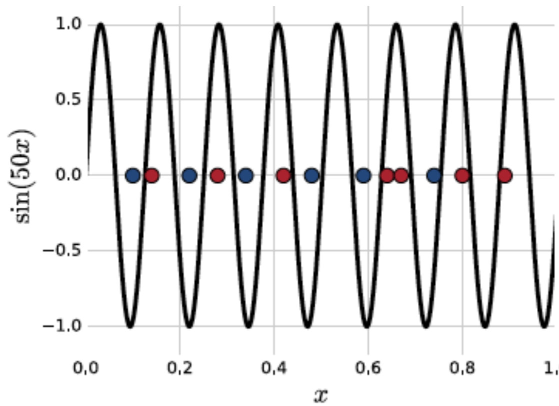
Material for assignment 1 (updated on 2nd of April).

Recap - VCdim($\mathcal{H}_{\sin}$)

VCdim($\mathcal{H}_{\text{thresholds}}$) = 1, VCdim($\mathcal{H}_{\text{intervals}}$) = 2, VCdim($\mathcal{H}_{\text{lines}}$) = 3, VCdim($\mathcal{H}_{\text{rec}^2}$) = 4

Consider $\mathcal{H} = \mathcal{H}_{\sin}$ be the set of sin functions:

$\mathcal{H}_{\sin} = \{h_\theta: \mathbf{R} \rightarrow \{0,1\} \mid h_\theta(x) = \lceil \sin(\theta x) \rceil, \theta \in \mathrm{R}\}$, $\lceil -1 \rceil = 0$



Show that VCdim($\mathcal{H}_{\sin}$) = ∞ based on the following lemma:

Let x $\in$ (0, 1) and let $0.x_1x_2x_3\ldots$ be the binary representation of x. Then, for any natural number m, provided that there exist $k \geq m$ such that $x_k = 1$, we have:

$$\lceil \sin(2^m \pi x) \rceil = 1 - x_m$$

Recap - VCdim($\mathcal{H}S_0^{\,n}$)

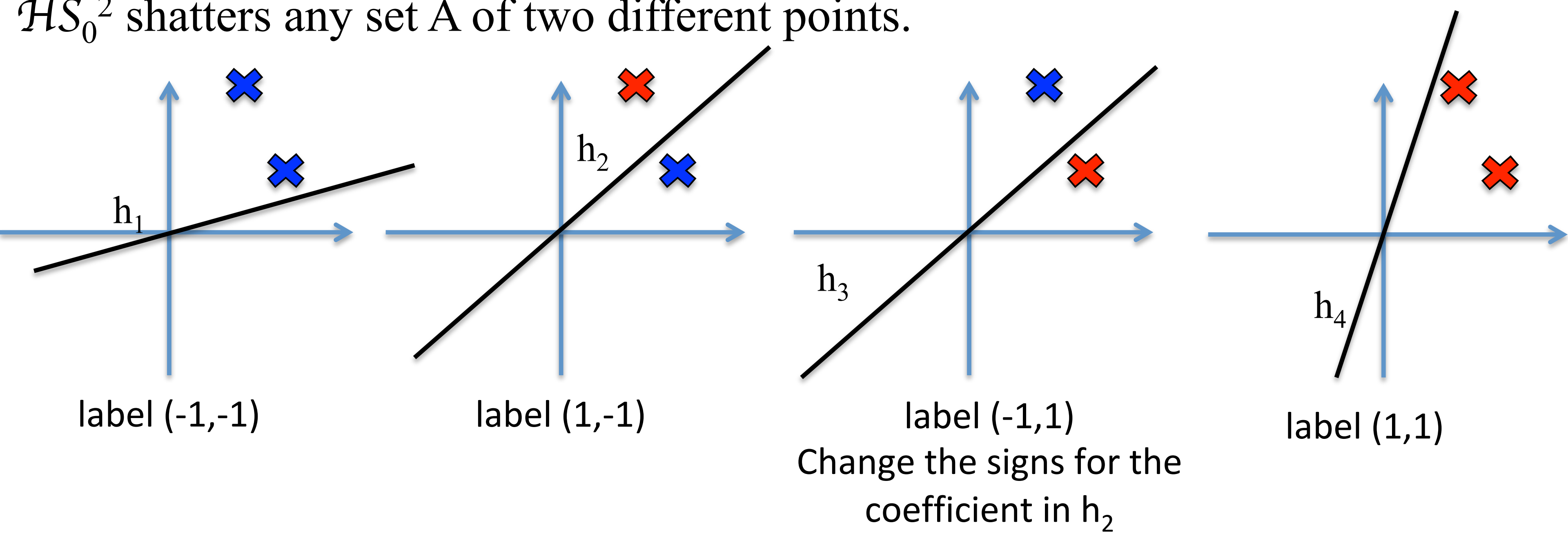
$\mathcal{H}S_0^{\,n} = \{h_{w,0}: \mathbf{R}^n \rightarrow \{-1, 1\},\ h_{w,0}(x) = sign\left(\sum_{i=1}^{n} w_i x_i\right) \mid w \in \mathbf{R}^n\}$

For n = 2 we have:

$\mathcal{H}S_0^{\,2} = \{h_{w1,w2}: \mathbf{R}^2 \rightarrow \{-1, 1\},\ h_{w1,w2}(x) = sign(w_1x_1 + w_2x_2) \mid (w_1,w_2) \in \mathbf{R}^2\}$

What is the VCdim($HS_0^{\,2}$) ?

$\mathcal{H}S_0^{\,2}$ shatters any set A of two different points.



Does $HS_0^{\,2}$ shatter a set A of three points?

Difficult to reason geometrically… choose the algebraic proof.

Recap - VCdim($\mathcal{HS}_0^n$)

We will show that VCdim($\mathcal{HS}_0^n$) = n.

Proof: 1st part

We first show that VCdim($\mathcal{HS}_0^n$) $\geq n$.

We find a set A consisting of n points in $\mathbf{R}^n$ that is shattered by $\mathcal{HS}_0^n$.

Take A = $\{e_1, e_2, \ldots, e_n\}$ to be the orthonormal basis of $\mathbf{R}^n$.

$e_1 = (1, 0, 0, \ldots, 0)$; $e_2 = (0, 1, 0, \ldots, 0)$; …. ; $e_n = (0, 0, 0, \ldots, 1)$

We want to proof that $\mathcal{HS}_0^n$ shatters A, so that VCdim($\mathcal{HS}_0^n$) $\geq n$. This is equivalent to proof that for every B $\subseteq$ A, there is a function $h_B \in \mathcal{HS}_0^n$ such that h_B gives label +1 to all elements in B and label -1 to all elements of A \ B.

Pick B subset of A, B $\subseteq \{e_1, e_2, \ldots, e_n\}$. Choose $w = (w_1, w_2, \ldots, w_n)$ such that:

$$w_i = \begin{cases} 1, & \text{if } e_i \in B \\ -1, & \text{if } e_i \notin B \end{cases}$$

Then, $h_B(e_i) = \text{sign}(\langle w, e_i \rangle) = w_i$ will generate the labels +1 for elements in B, -1 for elements not in B

Recap - VCdim($\mathcal{H}S_0^n$)

Proof: 2nd part

We now show that VCdim($\mathcal{H}S_0^n$) $< n + 1$.

We will prove that given any set A = $\{x_1, x_2, \ldots, x_{n+1}\}$ of $n + 1$ points in $\mathbf{R}^n$, A cannot be shattered by $\mathcal{H}S_0^n$.

The points $\{x_1, x_2, \ldots, x_{n+1}\}$ "live" in $\mathbf{R}^n$, a vector space with dimension n. So, $\{x_1, x_2, \ldots, x_{n+1}\}$ are linearly dependent and there exist coefficients $a_1, a_2, \ldots$ a_{n+1} not all of them 0 such that:

$$\sum_{i=1}^{n+1} a_i x_i = 0$$

Take P ⊆ { 1, 2, …, n +1} the set of strictly positive coefficients a_i and N ⊆ { 1, 2, …, n +1} the set of negative coefficients of a_i. So we have:

$$\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$$

Recap - VCdim($\mathcal{H}S_0^{\,n}$)

Take P $\subseteq$ { 1, 2, …, n +1} the set of positive coefficients a_i and N $\subseteq$ { 1, 2, …, n +1} the set of negative coefficients of a_i. Both P and N cannot be at the same time empty. So we have:

$$\sum_{i\in P} a_i x_i = \sum_{j\in N} |a_j| x_j$$

Assume that A is shattered by $\mathcal{H}S_0^{\,n}$ and take B = $\{x_i | \ i \in P\}$. In particular, there exist h_B such that it realizes the label consisting of +1 for all $x_i \in$ B and -1 for all $x_i \notin$ B.

So, we have that $h_B(x_i) = 1$, if $x_i \in$ B, meaning that $\langle w_B, x_i\rangle \geq 0$ if $x_i \in$ B
and $h_B(x_i) = -1$, if $x_i \notin$ B, meaning that $\langle w_B, x_i\rangle < 0$ if $x_i \notin$ B

So, we have that $h_B\left(\sum_{i\in P} a_i x_i\right) = sign(\left\langle w_B, \sum_{i\in P} a_i x_i \right\rangle) = sign(\sum_{i\in P} a_i \left\langle w_B, x_i \right\rangle)$

But $a_i > 0$ (because i $\in$ P) and also $\langle w_B, x_i\rangle \geq 0$ as $x_i \in$ B, so we obtain that:

$$h_B\left(\sum_{i\in P} a_i x_i\right) = sign(\left\langle w_B, \sum_{i\in P} a_i x_i \right\rangle) = sign(\sum_{i\in P} a_i \left\langle w_B, x_i \right\rangle) = 1$$

Recap - VCdim($\mathcal{H}S_0^n$)

Take $P \subseteq \{1, 2, \ldots, n+1\}$ the set of positive coefficients a_i and $N \subseteq \{1, 2, \ldots, n+1\}$ the set of negative coefficients of a_i. Both P and N cannot be at the same time empty. So we have:

$$\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$$

Assume that A is shattered by $\mathcal{H}S_0^n$ and take $B = \{x_i | \, i \in P\}$. In particular, there exist h_B such that it realizes the label consisting of +1 for all $x_i \in B$ and -1 for all $x_i \notin B$.

So, we have that $h_B(x_i) = 1$, if $x_i \in B$, meaning that $\langle w_B, x_i \rangle \geq 0$ if $x_i \in B$ and $h_B(x_i) = -1$, if $x_i \notin B$, meaning that $\langle w_B, x_i \rangle < 0$ if $x_i \notin B$

On the other hand, we have that $h_B\left(\sum_{j \in N} |a_j| x_j\right) = sign\left(\left\langle w_B, \sum_{j \in N} |a_j| x_j \right\rangle\right) = sign\left(\sum_{j \in N} |a_j| \langle w_B, x_j \rangle\right)$

But $|a_j| > 0$ and also $\langle w_B, x_j \rangle < 0$ as $x_j \notin B$, so we obtain that:

$$h_B\left(\sum_{j \in N} |a_j| x_j\right) = sign\left(\left\langle w_B, \sum_{j \in N} |a_j| x_j \right\rangle\right) = sign\left(\sum_{j \in N} |a_j| \langle w_B, x_j \rangle\right) = -1$$

Recap - VCdim($\mathcal{HS}_0^{\,n}$)

Take P ⊆ { 1, 2, …, n +1} the set of positive coefficients a_i and N ⊆ { 1, 2, …, n +1}the set of negative coefficients of a_i. Both P and N cannot be at the same time empty. So we have:

$$\sum_{i\in P} a_i x_i = \sum_{j\in N} |a_j| x_j$$

Assume that A is shattered by $\mathcal{HS}_0^{\,n}$ and take B = $\{x_i | \ i \in P\}$. In particular, there exist h_B such that it realizes the label consisting of +1 for all $x_i \in$ B and -1 for all $x_i \notin$ B. So $h_B(x_i) = 1$, if $x_i \in$ B and $h_B(x_i) = 1$, if $x_i \notin$ B

$$\sum_{i\in P} a_i x_i = \sum_{j\in N} |a_j| x_j$$

$$h_B\left(\sum_{i\in P} a_i x_i\right) = sign(\left\langle w_B, \sum_{i\in P} a_i x_i \right\rangle) = sign(\sum_{i\in P} a_i \langle w_B, x_i \rangle) = 1$$

$$h_B\left(\sum_{j\in N} |a_j| x_j\right) = sign\left(\left\langle w_B, \sum_{j\in N} |a_j| x_j \right\rangle\right) = sign\left(\sum_{j\in N} |a_j| \langle w_B, x_j \rangle\right) = -1$$

So, this is a contradiction.

Recap - VCdim($\mathcal{HS}_0^n$)

Proof:

1st part – show that VCdim($\mathcal{HS}_0^n$) $\geq$ n

A = $\{e_1, e_2, \ldots, e_n\}$, the orthonormal basis of $\mathbf{R}^n$ is shattered by $\mathcal{HS}_0^n$.

2nd part – show that VCdim($\mathcal{HS}_0^n$) < n + 1

Any set A = $\{x_1, x_2, \ldots, x_{n+1}\}$ of n + 1 points in $\mathbf{R}^n$ cannot be shattered by $\mathcal{HS}_0^n$. Provide an algebraic proof, based on the fact that $\{x_1, x_2, \ldots, x_{n+1}\}$ are linearly dependent in $\mathbf{R}^n$.

So, VCdim($\mathcal{HS}_0^n$) = n

Similarly, it can be shown that VCdim($\mathcal{HS}^n$) = n + 1

The fundamental theorem of statistical learning

The fundamental theorem of statistical learning

Theorem (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 statements 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.

A finite VC- dimension guarantees learnability*. *Hence, the VC-dimension characterizes PAC learnability.

Proof

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.

Proof:

1 → 2 follows from lecture 4: uniform convergence property → every sample S is ε-representative → ERM is a successful agnostic PAC learner

2 → 3, 3 → 4 (lecture 5), 2 → 5 follow immediately

4 → 6 (lecture 5), 5 → 6 – follow from the No-Free Lunch theorem

Need to prove 6 → 1 (the hardest part)

Remember – lecture 3: uniform convergence property

Definition (*uniform convergence*)

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* wrt a domain $\mathcal{Z}$, loss function ℓ if:

- there exists a function $m_H^{UC} : (0,1)^2 \to N$
- such that for all $(\varepsilon, \delta) \in (0,1)^2$
- and for any probability distribution $\mathcal{D}$ over $\mathcal{Z}$

if S is a sample of m $\geq m_H^{UC}(\varepsilon,\delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ε-representative.

Definition (ε – representative sample)

A sample S is called ε – representative wrt domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ and distribution $\mathcal{D}$ if:

$$\forall h \in \mathcal{H}, \ \ |L_S(h) - L_\mathcal{D}(h)| \leq \epsilon.$$

Lemma

Let S be a sample that is ε/2 – representative wrt domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ and distribution $\mathcal{D}$. Then any output of $\mathrm{ERM}_{\mathcal{H}}(S)$ i.e any $h_S \in \mathrm{argmin}_h L_S(h)$ satisfies:

$$L_\mathcal{D}(A(S)) \leq \min_{h\in\mathcal{H}} L_\mathcal{D}(h) + \epsilon$$

Proof for $6 \to 1$

We want to prove that finite VC-dimension $\to$ *uniform convergence property*

Two steps:

1. (Sauer's lemma) If $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$, then even though $\mathcal{H}$ might be infinite, when restricting it to a finite set $C \subseteq \mathcal{X}$, its "effective" size, $|\mathcal{H}_C|$, is only $O(|C|^d)$. That is, the size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$.

2. we have shown in lecture 4 that finite hypothesis classes enjoy the uniform convergence property. We generalize this result and show that uniform convergence holds whenever the hypothesis class has a "small effective size." By "small effective size" we mean classes for which $|\mathcal{H}_C|$ grows polynomially with $|C|$.

The Growth function

Definition

Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$, denoted by τ_H, where $\tau_{\mathcal{H}}$: $\mathbf{N} \rightarrow \mathbf{N}$, is defined as:

$$\tau_H(m) = \max_{C \subseteq X : |C| = m} |H_C|$$

In other words, $\tau_H(m)$ is the maximum number of different functions from a set C of size m to {0,1} that can be obtained by restricting $\mathcal{H}$ to C.

Observation: if VCdim($\mathcal{H}$) = d then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$. In such cases, $\mathcal{H}$ induces all possible functions from C to {0,1}.

What happens when m becomes larger than the VC-dimension?

Answer given by the Sauer's lemma: the growth function $\tau_{\mathcal{H}}$ increases polynomially rather than exponentially with m.

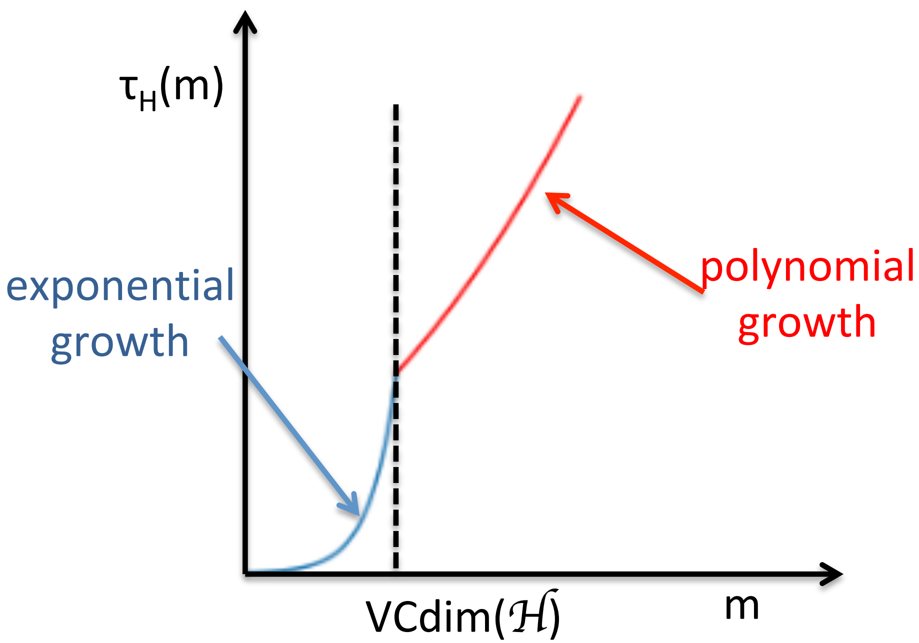
The Sauer's lemma

Lemma (Sauer – Shelah – Perles)

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

$$\tau_H(m) \leq \sum_{i=0}^{d} C_m^i$$

In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d = O(m^d)$



The Sauer's lemma - proof

Lemma (Sauer – Shelah – Perles)

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

$$\tau_H(m) \leq \sum_{i=0}^{d} C_m^i$$

In particular, if m > d +1 then $\tau_{\mathcal{H}}(m) \leq (em/d)^d = O(m^d)$

Proof

To prove the lemma it suffices to prove the following stronger claim:

For any $C = \{c_1, c_2, \dots, c_m\}$ we have:

$$|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|, \text{ for all } \mathcal{H} \text{ a hypothesis class}$$

The reason why this claim is sufficient to prove the lemma is that if VCdim($\mathcal{H}$) $\leq$ d then no set B whose size is larger than d is shattered by $\mathcal{H}$ and therefore:

$$\tau_H(m) = \max_{C \subseteq X : |C| = m} |H_C| \leq \max_{C \subseteq X : |C| = m} |\{B \subseteq C : |B| \leq d\}| \leq \sum_{i=0}^{d} C_m^i$$

The Sauer's lemma - proof

We will employ induction over the size of C

First step: Fix $\mathcal{H}$ and consider $|C| = 1$.

If $|\mathcal{H}_C| = 1 \leq |\{B \subseteq C: \mathcal{H}$ shatters $B\}| = 1$ ($\mathcal{H}$ shatters the empty set).

If $|\mathcal{H}_C| = 2 \leq |\{B \subseteq C: \mathcal{H}$ shatters $B\}| = 2$ ($\mathcal{H}$ shatters the empty set and C)

Induction step:

Assume the claim holds for $|C| \leq m$ and prove it for $|C| = m+1$.

Fix $\mathcal{H}$ and consider $C = \{c_1, c_2, \ldots, c_m, c_{m+1}\}$ and $C' = \{c_1, c_2, \ldots, c_m\}$.

Take $Y_0 = \{g: C' \rightarrow \{0, 1\} \mid$ exists $h \in \mathcal{H}$ such that $h(c) = g(c)$ for all $c \in C'$ and $h(c_{m+1}) = 0$ OR $h(c_{m+1}) = 1\}$

So, $Y_0 = \mathcal{H}_{C'}$

Y_0

c_1	c_2	…	c_m	c_{m+1}
1	1	0	1	0
1	1	0	1	1
0	1	1	1	1
1	0	0	1	0
1	0	0	0	1
…	…	…	…	…

The Sauer's lemma - proof

We will employ induction over the size of C

First step: Fix $\mathcal{H}$ and consider $|C| = 1$.

If $|\mathcal{H}_C| = 1 \leq |\{B \subseteq C: \mathcal{H} \text{ shatters } B\}| = 1$ ($\mathcal{H}$ shatters the empty set).

If $|\mathcal{H}_C| = 2 \leq |\{B \subseteq C: \mathcal{H} \text{ shatters } B\}| = 2$ ($\mathcal{H}$ shatters the empty set and C)

Induction step:

Assume the claim holds for $|C| \leq m$ and prove it for $|C| = m+1$.

Fix $\mathcal{H}$ and consider $C = \{c_1, c_2, \ldots, c_m, c_{m+1}\}$ and $C' = \{c_1, c_2, \ldots, c_m\}$.

Take $Y_0 = \{g: C' \rightarrow \{0, 1\} |$ exists $h \in \mathcal{H}$ such that $h(c) = g(c)$ for all $c \in C'$ and $h(c_{m+1}) = 0$ OR $h(c_{m+1}) = 1\} = \mathcal{H}_{C'}$

If there exists two different function h_1 and h_2 in $\mathcal{H}$ that agree with g on C' then they will disagree on c_{m+1}: $h_1(c_{m+1}) \neq h_2(c_{m+1})$. They are two different functions in $\mathcal{H}$ but they will be counted only once in Y_0.

The Sauer's lemma - proof

Take $Y_0 = \{g: C' \rightarrow \{0, 1\} \mid \text{exists } h \in \mathcal{H} \text{ such that } h(c) = g(c) \text{ for all } c \in C' \text{ and } h(c_{m+1}) = 0 \text{ OR } h(c_{m+1}) = 1\} = \mathcal{H}_{C'}$

Take $Y_1 = \{g: C' \rightarrow \{0, 1\} \mid \text{exists } h_1, h_2 \in \mathcal{H} \text{ such that } h_1(c) = g(c) \text{ for all } c \in C' \text{ and } h_1(c_{m+1}) = 0 \text{ AND } h_2(c) = g(c) \text{ for all } c \in C' \text{ and } h_2(c_{m+1}) = 1\}$

	c_1	c_2	…	c_m	c_{m+1}	
	1	1	0	1	0	h_1
Y_1	1	1	0	1	1	h_2
Y_0	0	1	1	1	1	
	1	0	0	1	0	
	1	0	0	0	1	
	…	…	…	…	…	

The Sauer's lemma - proof

Take $Y_0 = \{g: C' \rightarrow \{0, 1\} |$ exists $h \in \mathcal{H}$ such that $h(c) = g(c)$ for all $c \in C'$ and $h(c_{m+1}) = 0$ OR $h(c_{m+1}) = 1\} = \mathcal{H}_{C'}$

Take $Y_1 = \{g: C' \rightarrow \{0, 1\} |$ exists $h_1, h_2 \in \mathcal{H}$ such that $h_1(c) = g(c)$ for all $c \in C'$ and $h_1(c_{m+1}) = 0$ AND $h_2(c) = g(c)$ for all $c \in C'$ and $h_2(c_{m+1}) = 1\}$

We have that $Y_1 \subseteq Y_0$

Y_1 contains only those restriction $h_{C'}$ that come from two different functions h_1 and h_2 from $\mathcal{H}$

Y_0 might contain restrictions $h_{C'}$ that come from a single h from H.

For simplicity let's assume that $C = \mathcal{X}$, $\mathcal{X}$ is the domain of $\mathcal{H}$.

We have that $|H| = |Y_0| + |Y_1|$

The Sauer's lemma - proof

Take $Y_0 = \{g: C' \rightarrow \{0, 1\} \mid \text{exists } h \in \mathcal{H} \text{ such that } h(c) = g(c) \text{ for all } c \in C'$ and $h(c_{m+1}) = 0 \text{ OR } h(c_{m+1}) = 1\} = \mathcal{H}_{C'}$

Take $Y_1 = \{g: C' \rightarrow \{0, 1\} \mid \text{exists } h_1, h_2 \in \mathcal{H} \text{ such that } h_1(c) = g(c) \text{ for all } c \in$ C' and $h_1(c_{m+1}) = 0$ AND $h_2(c) = g(c)$ for all $c \in C'$ and $h_2(c_{m+1}) = 1\}$

	c_1	c_2	…	c_m	c_{m+1}	
Y_1	1	1	0	1	0	h_1
	1	1	0	1	1	h_2
Y_0	0	1	1	1	1	
	1	0	0	1	0	
	1	0	0	0	1	
	…	…	…	…	…	

The Sauer's lemma - proof

Now, we will apply our induction hypothesis on Y_0

$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subseteq C : \mathcal{H} \text{ shatters } B \text{ and } c_{m+1} \notin B\}|$

Take $\mathcal{H}' = \{h_1 \in \mathcal{H}$ such that there exists $h_2 \in \mathcal{H}$ s. t. for all $c \in C'$ we have $h_1(c) = h_2(c)$ but $h_1(c_{m+1}) \neq h_2(c_{m+1})\}$

Then $Y_1 = \mathcal{H}'_{C'}$ = set of function on C' with two extensions on c_{m+1}

Use the induction hypothesis here, on Y_1 :

$|Y_1| = |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B\}| = |\{B \subseteq C : \mathcal{H} \text{ shatters } B \text{ and } c_{m+1} \in B\}|$

So, we have that $|\mathcal{H}| = |\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|$

$\tau_{\mathcal{H}}$ grows polynomially

Corollary

Let H be a hypothesis class with VCdim(H) = d. Then for all m ≥ d:

$$\tau_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d)$$

Proof:

From the Sauer lemma we have:

$$\tau_H(m) \leq \sum_{i=0}^{d} C_m^i \underset{\uparrow\ m \geq d}{\leq} \sum_{i=0}^{d}\left(C_m^i \times \left(\frac{m}{d}\right)^{d-i}\right) \underset{\uparrow\ m \geq d}{\leq} \sum_{i=0}^{m}\left(C_m^i \times \left(\frac{m}{d}\right)^{d-i}\right) = \left(\frac{m}{d}\right)^d \sum_{i=0}^{m}\left(C_m^i \times \left(\frac{d}{m}\right)^{i}\right)$$

$$\tau_H(m) \leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{m}\left(C_m^i \times \left(\frac{d}{m}\right)^{i}\right) \underset{\uparrow\ \text{Newton's binomial formula}}{=} \left(\frac{m}{d}\right)^d \left(1+\frac{d}{m}\right)^m \underset{\uparrow\ 1-x \leq e^{-x}}{\leq} \left(\frac{m}{d}\right)^d \left(e^{\frac{d}{m}}\right)^m = \left(\frac{em}{d}\right)^d$$

Proof for 6 → 1

We want to prove that finite VC-dimension → *uniform convergence property*

Two steps:

1. (Sauer's lemma) If VCdim($\mathcal{H}$) = d < ∞, then even though $\mathcal{H}$ might be infinite, when restricting it to a finite set $C \subseteq \mathcal{X}$, its "effective" size, $|\mathcal{H}_C|$, is only $O(|C|^d)$. That is, the size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$.

2. we have shown in lecture 4 that finite hypothesis classes enjoy the uniform convergence property. We generalize this result and show that uniform convergence holds whenever the hypothesis class has a "small effective size." By "small effective size" we mean classes for which $|\mathcal{H}_C|$ grows polynomially with $|C|$.

Uniform converge holds for $\mathcal{H}$ with small effective size

Theorem

Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of at least $1 - \delta$ over the choice of S ~ $\mathcal{D}^m$ we have:

$$\left|L_D(h) - L_S(h)\right| \leq \frac{4 + \sqrt{\log(\tau_H(2m))}}{\delta\sqrt{2m}}$$

Proof:

- in the book, is beyond the scope of this lecture

Proof for 6 → 1

We want to prove that finite VC-dimension → *uniform convergence property.*

Combine the last result with Sauer lemma: $\tau_{\mathcal{H}}(m) \le (em/d)^d = O(m^d)$ to obtain: for every $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ we have:

$$|L_D(h) - L_S(h)| \le \frac{4 + \sqrt{\log(\tau_H(2m))}}{\delta\sqrt{2m}} \le \frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}} \le \frac{2\sqrt{d\log(2em/d)}}{\delta\sqrt{2m}}$$

Sauer lemma
$\tau_{\mathcal{H}}(2m) \le (2em/d)^d$

consider m such that
$4^2 \le d\log(2em/d)$

$$|L_D(h) - L_S(h)| \le \frac{1}{\delta}\frac{\sqrt{2d\log(2em/d)}}{\sqrt{m}} < \varepsilon$$

This leads (see the calculation in the book) to:

$$m \ge 4\frac{2d}{(\delta\varepsilon^2)}\log(\frac{2d}{\delta\varepsilon^2}) + \frac{4d\log(2\varepsilon/d)}{(\delta\varepsilon^2)}$$

Proof for 6 → 1

We want to prove that finite VC-dimension → *uniform convergence property.*

for every $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$ we have that if:

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

then the sample S is ε-representative

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

So, we have that: $m_H^{UC}(\varepsilon,\delta) \leq 4\frac{2d}{(\delta\varepsilon^2)}\log(\frac{2d}{\delta\varepsilon^2}) + \frac{4d\log(2\varepsilon/d)}{(\delta\varepsilon^2)}$

The derived bound is not the tightest possible, there exist another bound much tighter (see next).

The fundamental theorem of statistical learning – quantitative version

Theorem

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 VCdim($\mathcal{H}$) = d <∞. Then, there are absolute constants C_1, C_2 such that:

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

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

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

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

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

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

The VC dimension determines (along with ε, δ) the samples complexities of learning a class. It gives us a lower and an upper bound.

Intuition for deriving the lower bounds

The PAC case (realizable case)

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

Pick a set $A = \{x_1, x_2, \ldots, x_d\}$ of size d (=VCdim($\mathcal{H}$)) that is shattered by $\mathcal{H}$. Choose the following (adversarial) probability distribution $\mathcal{D}$ over $\mathcal{X}$:

$\mathcal{D}(x_1) = 1\text{-}4\varepsilon$, $\mathcal{D}(x_i) = 4\varepsilon/(d\text{-}1)$, $i = 2,3,\ldots,d$, $\mathcal{D}(x) = 0$, for all x in $\mathcal{X} \setminus A$

By the No Free Lunch theorem as long as a sample S hits $B = \{x_2,\ldots x_d\}$ at most (d-1)/2 times, the probability of making an error over B is $\geq 1/4$. This happens because we see less then half of the domain B points. So, our expected error with respect to $\mathcal{D}$ is $4\varepsilon/4 = \varepsilon$.

If the sample S has size *m*, then roughly $4m\varepsilon$ points will hit $B = \{x_2, \ldots, x_d\}$. So, to make less than ε errors we need to have $4m\varepsilon > (d\text{-}1)/2$, $m > (d\text{-}1)/8\varepsilon$