Advanced Machine Learning

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

University of Bucharest, 2^{nd} semester, 2020-2021

Assignment 1

Deadline: Sunday, 25th of April 2021

Upload your solutions at: https://tinyurl.com/AML-2021-ASSIGNMENT1

1. **(0.5 points)** Give an example of a finite hypothesis class $\mathcal{H}$ with VCdim($\mathcal{H}$) = 2021. Justify your choice.

2. **(0.5 points)** Consider $\mathcal{H}_{\text{balls}}$ to be the set of all balls in $\mathbf{R}^2$:
   $\mathcal{H}_{\text{balls}} = \{B(x,r), x \in \mathbb{R}^2, r \geq 0\}$, where $B(x,r) = \{y \in \mathbb{R}^2 | \, \| y - x \|_2 \leq r\}$
   As mentioned in the lecture, we can also view $\mathcal{H}_{\text{balls}}$ as the set of indicator functions of the balls B(x,r) in the plane: $\mathcal{H}_{\text{balls}} = \{ h_{x,r}: \mathbb{R}^2 \to \{0,1\}, h_{x,r} = \mathbf{1}_{B(x,r)}, x \in \mathbb{R}^2, r > 0\}$.
   Can you give an example of a set A in $\mathbf{R}^2$ of size 4 that is shattered by $\mathcal{H}_{\text{balls}}$? Give such an example or justify why you cannot find a set A of size 4 shattered by $\mathcal{H}_{\text{balls}}$.

3. **(1 point)** Let $\mathcal{X} = \mathbf{R}^2$ and consider $\mathcal{H}_\alpha$ the set of concepts defined by the area inside a right triangle ABC with the two catheti AB and AC parallel to the axes (Ox and Oy) and with AB/AC = α (fixed constant > 0). Consider the realizability assumption. Show that the class $\mathcal{H}_\alpha$ can be (ϵ, δ) − PAC learned by giving an algorithm A and determining an upper bound on the sample complexity $m_H(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied.

Assignment 1

Deadline: Sunday, 25th of April 2021

Upload your solutions at: https://tinyurl.com/AML-2021-ASSIGNMENT1

4. **(1 point)** 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). Consider the realizability assumption.
   a. show that the class $\mathcal{H}$ can be (ϵ, δ) – PAC learned by giving an algorithm A and determining an upper bound on the sample complexity $m_H(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied. **(0.5 points)**
   b. compute VCdim($\mathcal{H}$). **(0.5 points)**

5. **(1 point)** Let $\mathcal{H} = \{h_\theta : \mathbb{R} \to \{0,1\}, h_\theta(x) = \mathbf{1}_{[\theta,\theta+1]\cup[\theta+2,+\infty)}(x), \theta \in \mathbb{R}\}$. Compute VCdim($\mathcal{H}$).

6. **(1 point)** Let $\mathcal{X}$ be an instance space and consider $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ a hypothesis space with finite VC dimension. For each $x \in \mathcal{X}$, we consider the function $z_{\mathbf{x}}$: $\mathcal{H} \to \{0,1\}$ such that $z_{\mathbf{x}}(h) = h(\mathrm{x})$ for each $h \in \mathcal{H}$. Let $Z = \{z_{\mathbf{x}}: \mathcal{H} \to \{0,1\}, x \in \mathcal{X}\}$. Prove that $\mathrm{VCdim}(Z) < 2^{\mathrm{VCdim(H)}+1}$.

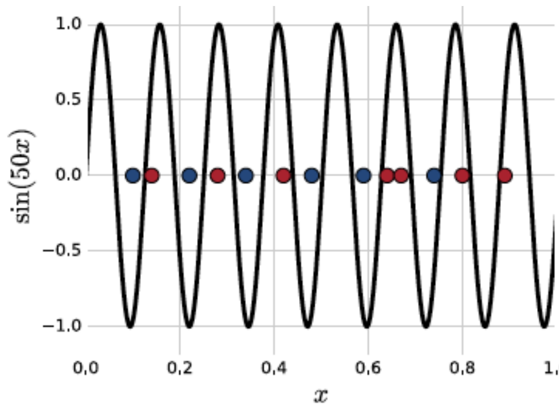
Ex-officio: 0.5 points

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:

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



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

Let x ∈ (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 ≥ m such that x_k = 1, we have:

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

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

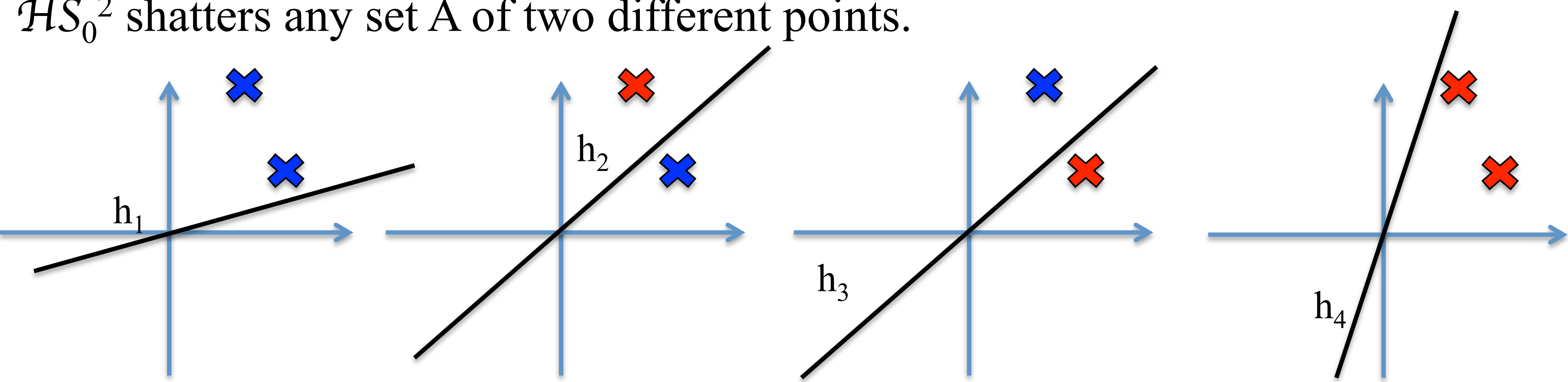
$\mathcal{H}S_0^{\,n} = \{h_{w,0}: \mathbf{R}^n \rightarrow \{-1, 1\},\ h_{w,0}(x) = sign\left(\sum_{i=1}^{n} w_i x_i\right) \mid w \in \mathbf{R}^n\}$

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.



label (-1,-1)

label (1,-1)

label (-1,1)
Change the signs for the coefficient in h_2

label (1,1)

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

Difficult to reason geometrically… choose the algebraic proof.

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

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 from the definition

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

Need to prove 6 → 1 (the hardest part)

Remember – lecture 4: uniform convergence property

Definition (*uniform convergence*)

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* wrt a domain $\mathcal{Z}=\mathcal{X} \times \mathcal{Y}$, loss fct. ℓ if:

- there exists a function $m_H^{UC} : (0,1)^2 \to \mathrm{N}$
- 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 $\rightarrow$ 1

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

Two steps:

1. (Sauer's lemma) If 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 $\tau_{\mathcal{H}}$: **N**→**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 ≤ 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 Growth function

Definition

Let $\mathcal{H}$ be a hypothesis class. *$\tau_H(m)$ 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.*

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

Example: consider $\mathcal{H} = \mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line $\mathcal{H}_{\text{thresholds}} = \{h_a : \mathrm{R} \to \{0, 1\}, h_a(x) = \mathbf{1}_{[x<a]}, a \in \mathbf{R}\}$, $|\mathcal{H}_{\text{thresholds}}| = \infty$.

We know that VCdim($\mathcal{H}$) = 1.

Consider $C = \{c_1, c_2, \ldots, c_m\}$ a set of m points, with $c_i < c_j$. What is *$\tau_H(m)$?*

The Growth function

Definition

Let $\mathcal{H}$ be a hypothesis class. *$\tau_H(m)$ 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.*

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

Example:

Consider C = $\{c_1, c_2, \ldots, c_m\}$ a set of m points, with $c_i < c_j$. What is $\tau_H(m)$?

$\mathcal{H}_C$ can have at most m+1 different functions: take $a_1 < c_1 < a_2 < c_2 < \ldots < a_m < c_m < a_{m+1}$, than we will have $|\mathcal{H}_C| = \{h_{a\_1}, h_{a\_2}, \ldots, h_{a\_m+1}\} = m+1$ as

h_{a_1} labels points $c_1, c_2, \ldots, c_m$ with labels (0, 0, 0, …., 0 , 0)

h_{a_2} labels points $c_1, c_2, \ldots, c_m$ with labels (1, 0, 0, …., 0 , 0)

h_{a_3} label points $c_1, c_2, \ldots, c_m$ with labels (1, 1, 0, …., 0 , 0)

……..

h_{a_m} label points $c_1, c_2, \ldots, c_m$ with labels (1, 1, 1, …., 1 , 0)

h_{a_m+1} label points $c_1, c_2, \ldots, c_m$ with labels (1, 1, 1, …., 1 , 1)

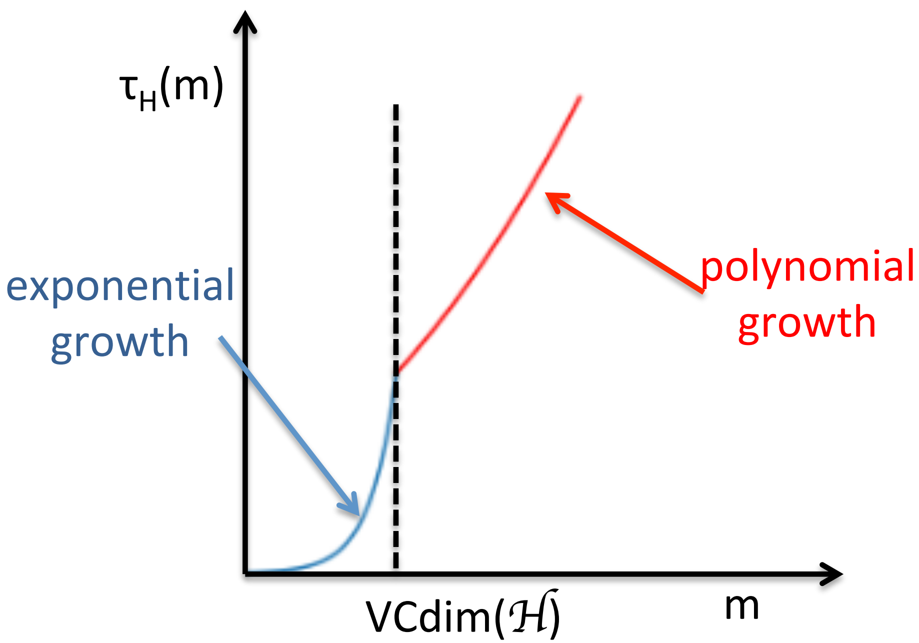
The Sauer's lemma

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 $\tau_{\mathcal{H}}(m) \leq (em/d)^d = O(m^d)$



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 $C = \{c_1, c_2, \ldots, c_m\}$ 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 $|\mathcal{H}_C| = 1 \leq |\{B \subseteq C\text{: } \mathcal{H} \text{ shatters } B\}| = 1$ ($\mathcal{H}$ shatters the empty set).

If $|\mathcal{H}_C| = 2 \leq |\{B \subseteq C\text{: } \mathcal{H} \text{ shatters } B\}| = 2$ ($\mathcal{H}$ 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 $C' = \{c_1, c_2, \ldots, c_m\}$, so $C = C' \cup \{c_{m+1}\}$

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

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

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

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

Take $Y_1 = \{g\colon C' \to \{0, 1\} \mid$ 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 $h_2(c_{m+1}) = 1\}$

	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:

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

Proof:

From the Sauer lemma we have:

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

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

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 ⊆ $\mathcal{X}$, 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 $S \sim \mathcal{D}^m$ 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: $\tau_{\mathcal{H}}(m) \leq (em/d)^d = O(m^d)$ 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:

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

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

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

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

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

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

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:

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

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 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 $4m\varepsilon > (d\text{-}1)/2$, $m > (d\text{-}1)/8\varepsilon$