Introduction to the VC-Dimension

Wei-Ning Chen

December 28, 2018

Outline

Definition (Uniform Convergence)

We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC}$ such that for every $\epsilon, \delta \in (0, 1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then

$$\mathcal{P}(|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon, \forall h \in \mathcal{H}) \geq 1 - \delta$$

Equivalently,

$$\lim_{m \to \infty} \mathcal{P}(\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = 0$$

Remark: Compare to the difinition of PAC:

$$\mathcal{P}(L_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \leq \epsilon) \geq 1 - \delta$$

RECAP

THEOREM (NO-FREE-LUNCH)

Let $\mathcal{A}$ be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

- *There exists a function $f : X \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
- $\mathcal{P}(L_{\mathcal{D}}(\mathcal{A}(S)) \geq \frac{1}{8}) \geq \frac{1}{7}$

In chapter 2, we see that every finite hypothesis class $\mathcal{H}$ is learnable; moreover, the sample complexity is bounded by

$$m_{\mathcal{H}}(\delta, \epsilon) \leq \frac{\log(|\mathcal{H}|)/\delta}{\epsilon}$$

So, what if $|\mathcal{H}| = \infty$?

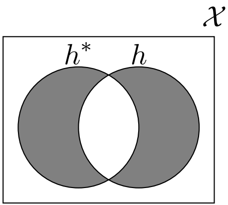
Example: Concentric Circle

Example

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane, that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$. Prove that $\mathcal{H}$ is PAC learnable (assume realizability), and its sample complexity is bounded by

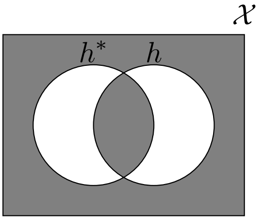
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\log(2\delta)}{\epsilon}.$$

First, we specify $\mathcal{H}_B$.
By definition, if $h \in \mathcal{H}_B$, we have $\mathcal{D}(h(x) \neq h^*(x)) \geq \epsilon$



Example: Concentric Circle

Equivalently, $\mathcal{D}(h(x) = h^*(x)) \leq 1 - \epsilon$

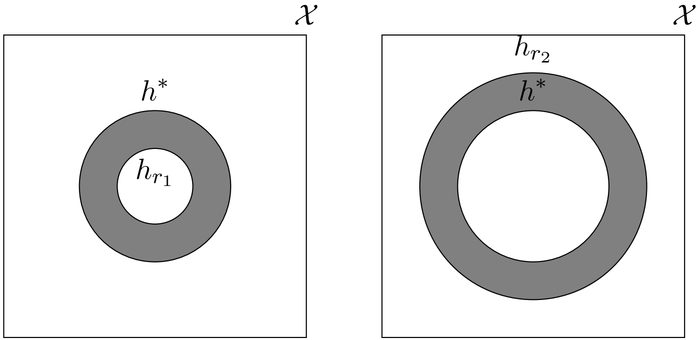


If now $\mathcal{H}$ is finite, we can apply union bound:

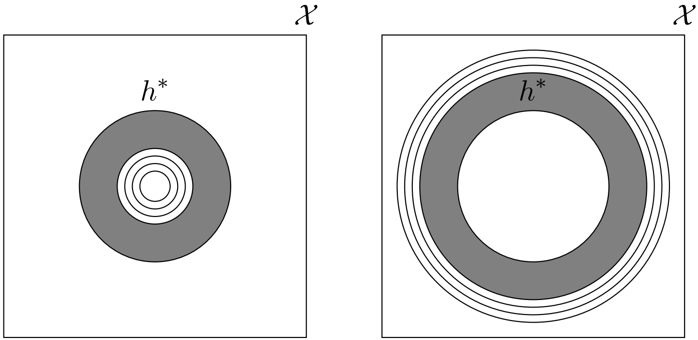
$$\mathcal{D}^m(\bigcup_{h \in \mathcal{H}_B} \forall i = [m] | h(x_i) = h^*(x_i)) \leq |\mathcal{H}|(1-\epsilon)^m \leq \delta$$

We can slightly modify the union bound for the case $|\mathcal{H}| = \infty$.

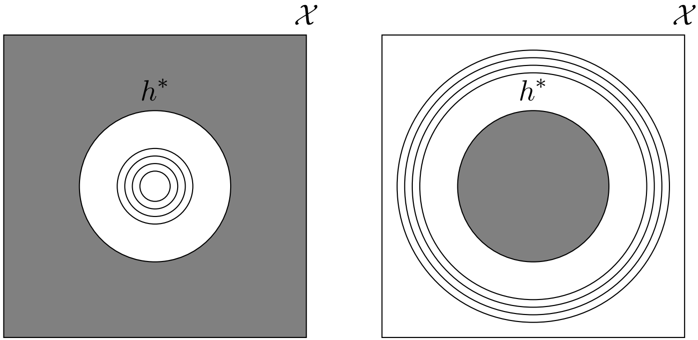
Example: Concentric Circle



Example: Concentric Circle



Example: Concentric Circle



Example: Concentric Circle

Let $S \sim \mathcal{D}^m$, and $r_{\text{min}} = \min_{x \in S} r_x$, $r_{\text{max}} = \max_{x \in S} r_x$.

We have

$$\begin{aligned}
\mathcal{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(h_S) \geq \epsilon) &\leq \mathcal{P}_{S\sim\mathcal{D}^m}(r_{\text{min}} \geq r_1 \cup r_{\text{max}} \leq r_0) \\
&\leq \mathcal{P}_{S\sim\mathcal{D}^m}(r_{\text{min}} \geq r_1) + \mathcal{P}_{S\sim\mathcal{D}^m}(r_{\text{max}} \leq r_0) \\
&\leq 2(1-\epsilon)^m \leq 2e^{-m\epsilon} \leq \delta
\end{aligned}$$

Therefore, for all ϵ and δ, the sample complexity can be bounded by

$$m \leq \frac{\log(2/\delta)}{\epsilon} \qquad \square$$

- In chapter 2, we see that every finite hypothesis class $\mathcal{H}$ is learnable; moreover, the sample complexity is bounded by

$$m_{\mathcal{H}}(\delta, \epsilon) \leq \frac{\log(|\mathcal{H}|)/\delta}{\epsilon}$$

- Also, we see some examples that even the class is infinite-size, it may still be learnable.

- Therefore, we need a measure of $\mathcal{H}$'s complexity

- In this cahpter, we will formally define the complexity of $\mathcal{H}$ (VC dimension) , and show that

$$\mathcal{H} \text{ has uniform convergence property } \iff \text{VCdim}(\mathcal{H}) < \infty$$

The VC-Dimension

Definition (Restriction $\mathcal{H}$ to C)

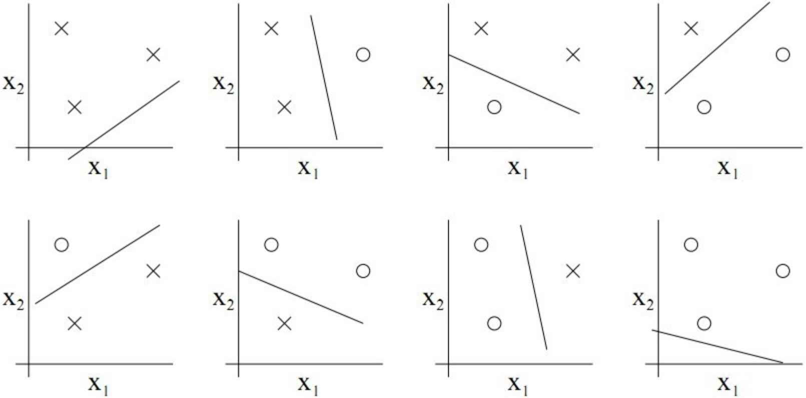
Let $\mathcal{H}$ be a class of function from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, ..., c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to C is the set off all functions from C to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_C = \{h(c_1), ..., h(c_m)) : h \in \mathcal{H}$$

Definition (Shattering)

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

The VC-Dimension



COROLLARY (COROLLARY6.4)

Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$. Let m be a training set size. Assume that there exists a set C of size $2m$ that is shattered by $\mathcal{H}$.
Then, for any learning algorithm $\mathcal{A}$

$$\mathcal{P}(L_{\mathcal{D}}(\mathcal{A}(S)) \geq \frac{1}{8}) \geq \frac{1}{7}$$

Remark: This is a direct result from NFL Theorem

Remark2: If for all m, there exists a set C of size $2m$ that is shattered by $\mathcal{H}$, then $\mathcal{H}$ is not PAC learnable

DEFINITION (VC-DIMENSION)

The VC-dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size, we say that $\mathcal{H}$ has infinite VC-dimension.

Remark: If VCdim($\mathcal{H}$)=d, it means that

$$\exists C \subset \mathcal{X} \text{ that can be shattered by } \mathcal{H},$$

NOT

$$\forall C \subset \mathcal{X} \text{ that can be shattered by } \mathcal{H},$$

Example: Threshold Functions

- Let $\mathcal{H}$ be the all threshold function on $\mathbb{R}$.
- For an arbitrary set $C = \{c_1\}$, $\mathcal{H}$ shatters C, therefore VDdim$(\mathcal{H}) \geq 1$.
- For an arbitrary set $C = \{c_1, c_2\}$, $\mathcal{H}$ does not shatter C. Therefore, VDdim$(\mathcal{H}) < 2$.

EXAMPLE: INTERVALS

- Let $\mathcal{H}$ be the intervals over $\mathbb{R}$; that is, $\mathcal{H} = \{\mathbb{1}_{[a,b]}(x) | a, b \in \mathbb{R}\}$
- It is easy to show that $\text{VCdim}(\mathcal{H}) = 2$

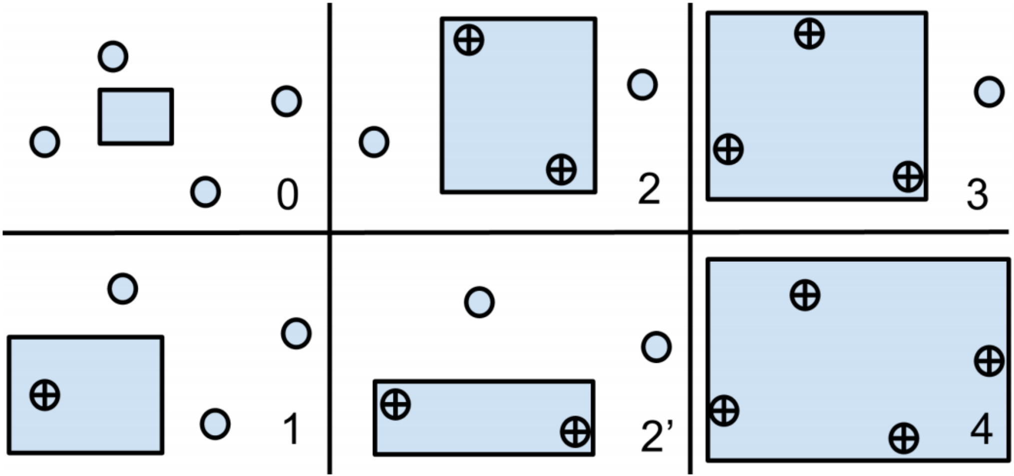
Example: Axis Aligned Rectangles

- Let $\mathcal{H}$ be the the class of axis aligned rectangles:

$$\mathcal{H} = \{\mathbb{1}_{[a,b]\times[c,d]} | a, b, c, d \in \mathbb{R}\}$$

- VDdim$(\mathcal{H}) = 4$

Example: Axis Aligned Rectangles



The Fundamental Theorem of Learning Theory

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 are quivalent:

1. *$\mathcal{H}$ has the uniform covergence 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.*

The growth function measures the maximal ”effective? size of $\mathcal{H}$ on a set of m examples.

Definition (Growth Function)

Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$ is defined as

$$\tau_{\mathcal{H}}(m) = \sup_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

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

Remark: if $\mathrm{VCdim}(\mathcal{H}) = d$, then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$.
However, what interesting is the case $m \geq d$.

THEOREM (SAUER'S LEMMA)

Let $\mathcal{H}$ be a hypothesis with VCdim$(\mathcal{H}) = d$. Then for all m,

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

Remark: if VCdim$(\mathcal{H})$ is finite, then the growth function is polynomial in m.

Theorem (Uniform converges in VC class (Theorem 6.11))

Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}(m)$ be its growth function. Then, for every $\mathcal{D}$ and every δ

$$\mathcal{P}_{S\sim\mathcal{D}^m}(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon) \leq \delta$$

where ϵ can be choose as $\frac{4+\sqrt{(\log(\tau_{\mathcal{H}}(2m)))}}{\delta\sqrt{2m}}$. In other words, this theorem tells us that

$$\text{VCdim}(\mathcal{H}) < \infty \iff \lim_{\ell\to\infty} \frac{\log \tau_{\mathcal{H}}(\ell)}{\ell} = 0 \iff \text{uniform convergence property holds.}$$

THEOREM (THE FUNDAMENTAL THEOREM-QUANTITATIVE VERSION)

Let $\mathcal{H}$ be a hypothesis class from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0-1 loss. Then, there are absolute constants C_1, C_2 such that:

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

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

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

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

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

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

Glivenko-Cantelli Theorem

Definition (Empirical Distribution)

Let $X_1, ..., X_n$ be i.i.d. random variables in $\mathbb{R}$ with common cdf $F(x)$. The empirical distribution function for $X_1, ..., X_n$ is given by

$$F_n(x) = \frac{1}{n} \sum_i \mathbb{1}_{(-\infty,x]}(X_i)$$

Theorem (Glivenko-Cantelli theorem)

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \textit{ almost surely}$$

Remark: $\forall x$, $F_n(x) \to F(x)$ trivially by LLN

GLIVENKO-CANTELLI THEOREM

- More generally, consider a space $\mathcal{X}$ and $\sigma-$field $\mathcal{F}$ generated by borel set with probability measure P and empirical measure P_n

- Then $F(x) = P((-\infty, x])$, and $F_n(x) = P_n((-\infty, x])$

- We can rewrite $\sup\limits_{x\in\mathbb{R}}|F_n(x) - F(x)|$ as $\sup\limits_{c\in\mathcal{C}}|P_n(C) - P(C)|$, where $\mathcal{C} = \{(-\infty, x]|x \in \mathbb{R}\}$

- What happens for the general $\mathcal{C}$?

Glivenko-Cantelli Class

Definition (GC-class)

Let $\mathcal{C} \subset \{C | C$ measurable in $\mathcal{X}\}$. Then if

$$\|P_n - P\|_{\mathcal{C}} = \sup_{c \in \mathcal{C}} |P_n(C) - P(C)|$$

we say class $\mathcal{C}$ is a Glivenko-Cantelli class.

Definition (uniformly GC)

A class is called uniformly Glivenko-Cantelli if the convergence occurs uniformly over all probability measures $\mathcal{P}$ on $(\mathcal{X}, \mathcal{F})$:

$$\sup_{P \in \mathcal{P}(S,A)} \mathbb{E}\|P_n - P\|_{\mathcal{C}} \to 0$$

VC Class

Definition (VC class)

A class with finite VC dimension is called a Vapnik-Chervonenkis class or VC class

Theorem (Vapnik and Chervonenkis, 1968)

A class of sets $\mathcal{C}$ is uniformly GC if and only if it is a Vapnik-Chervonenkis class

VC entropy and Growth function

In previous lecture, we define the growth function as

$$\tau_{\mathcal{H}}(m) = \sup_{C \subset \mathcal{X}:|C|=m} |\mathcal{H}_C|$$

Let's rewrite the growth function as another form:

Definition (Vapnik)

Let $x_1, ...x_m$ be m samples from $\mathcal{X}$. Then define the number

$$N^{\mathcal{H}}(x_1, ..., x_m) = |\{h(x_1), ..., h(x_m)|h \in \mathcal{H}\}| = |\mathcal{H}_{\{x_1,...,x_m\}}|$$

Obviously

$$\tau_{\mathcal{H}}(m) = \sup_{C \subset \mathcal{X}:|C|=m} |\mathcal{H}_C| = \sup_{\{x_1,...,x_m\} \subset \mathcal{H}} N^{\mathcal{H}}(x_1, ..., x_m)$$

The supremum is taken so that the bound works even in the worst distribution.
In general case, we can replace supremum by exapectation w.r.t. a specific distribution, which gives another value:

Definition (VC-entropy, annealed VC-entropy, growth function)

Let $N^{\mathcal{H}}(x_1, ..., x_m)$ be defined as previous. Then we defined VC-entropy as

$$H^{\mathcal{H}}(m) = \mathbb{E} \log N^{\mathcal{H}}(x_1, ..., x_m)$$

The annealed VC-entropy as

$$H^{\mathcal{H}}_{ann}(m) = \log \mathbb{E} N^{\mathcal{H}}(x_1, ..., x_m)$$

And the growth function (with logarithm) as

$$G^{\mathcal{H}}(m) = \log \sup_{x_1, ..., x_m} N^{\mathcal{H}}(x_1, ..., x_m) (= \log \tau_{\mathcal{H}}(m))$$

COROLLARY

$$H^{\mathcal{H}}(m) \leq H^{\mathcal{H}}_{ann}(m) \leq G^{\mathcal{H}}(m)$$

Rmark: The fundamental theorem of learning theory tells us

$$\lim_{m\to\infty} \frac{G^{\mathcal{H}}(m)}{m} = \lim_{m\to\infty} \frac{\log \tau_{\mathcal{H}}(m)}{m} = 0 \Longleftrightarrow \lim_{m\to\infty} \mathcal{P}(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = 0$$

The convergence is uniform for all distribution P in $(\mathcal{X}, \mathcal{F})$

Three milestones of learning theory

Theorem (Vapnik)

1.
$$\lim_{m\to\infty} \mathcal{D}(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = 0$$

is sufficient and necessary that

$$\lim_{m\to\infty} \frac{H^{\mathcal{H}}(m)}{m} = 0$$

2.
$$\lim_{m\to\infty} \mathcal{D}(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \le e^{-c\epsilon^2 m} \text{ (\textit{fast decay})}$$

is sufficient if

$$\lim_{m\to\infty} \frac{H^{\mathcal{H}}_{ann}(m)}{m} = 0$$

Theorem (Vapnik)

3

$$\lim_{m\to\infty} P(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = 0 \text{ ,for all } P \in (\mathcal{X}, \mathcal{F})$$

is sufficient and necessary that

$$\lim_{m\to\infty} \frac{G^{\mathcal{H}}(m)}{m} = 0$$