VC-dimension

Chapter 6 of Understanding Machine Learning: From Theory to Algorithms

Argyris Mouzakis
amouzakis17@hotmail.com

November 23, 2018

Overview

1. Introduction to the VC-dimension
2. The Fundamental Theorem of Statistical Learning
3. The end

The problem of characterization

- How do we figure out if a class is PAC-learnable;

Conjecture

The cardinality of the class determines whether it's PAC-learnable.

Finite classes

Theorem

Suppose $\mathcal{H}$ is a finite hypothesis class. $\mathcal{H}$ is PAC-learnable with sample complexity $\mathcal{O}\left(\frac{\ln\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon}\right)$.

- This settles the problem for finite classes.
- What about infinite classes?

Infinite hypothesis classes

Theorem

Suppose $\mathcal{H}$ is a class consisting of all classifiers $h : \mathcal{X} \to \{0, 1\}$. If $|\mathcal{X}| = \infty$, then $\mathcal{H}$ is not PAC-learnable.

- Can this be generalized for all infinite classes?

PAC-learnable infinite classes

- There are infinite classes that are PAC-learnable.
- Remember the concentric circles.

Theorem

Consider $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$ with $h_r(x) = \mathbb{1}\{||x|| \leqslant r\}, \forall x \in \mathbb{R}^2$. $\mathcal{H}$ is PAC-learnable with sample complexity $\mathcal{O}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon}\right)$.

- It's not the only one.

Learning intervals

Theorem

Consider $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ with $h_a(x) = \mathbb{1}\{x \leqslant a\}, \forall x \in \mathbb{R}$. $\mathcal{H}$ is PAC-learnable with sample complexity $\mathcal{O}\left(\frac{\ln\left(\frac{2}{\delta}\right)}{\epsilon}\right)$.

- We will provide merely a sketch of the proof.

Algorithm for interval learning

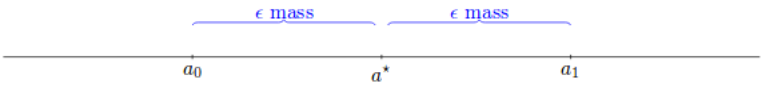
Algorithm 1: Real intervals

Data: training set S
Result: classifier h_S with 0 empirical risk

$b_0 = -\infty, b_1 = +\infty$;
for $(x, y) \in S$ **do**
    if $y == 1$ *and* $x > b_0$ **then**
        $b_0 = x$;
    else if $y == 0$ *and* $x < b_1$ **then**
        $b_1 = x$;
choose randomly $a \in (b_0, b_1)$;
$h_S = h_a$;

- Time complexity: $\mathcal{O}(m)$.

Analysis of algorithm



Sketch of Proof

- $a^\star$ is the a we are looking for.
- a_0, a_1 are such that $\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)]$.
- $\mathbb{P}_{S\sim\mathcal{D}^m}[L_\mathcal{D}(h_S) > \epsilon] \leqslant \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0 \vee b_1 > a_1]$
- Apply union bound.
- Find an upper bound (the same) for both probabilities.
- Demand the sum to be less than δ and this completes the proof. ■

Remarks

- This algorithm achieves the desired results with the aforementioned sample complexity.
- By using the idea from the concentric circles we can have one-sided error and achieve time and sample complexity $\mathcal{O}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon}\right)$.

Conclusion

The cardinality can't characterize the PAC-learnability of a class.

- Then what can?

Number of parameters

- Note that the infinite hypothesis classes we encountered before had smaller sample complexity than the one we proved for finite hypothesis classes.
- The elements of each of those classes can be described accurately using a single parameter (r in the circles example and a in the intervals example).
- On the other hand, we didn't make any assumptions about the members of $\mathcal{H}$ when we studied finite hypothesis classes.
- We need to know all its $|\mathcal{H}|$ members to describe it.

Parameters and PAC-learnability

Conjecture

The number of parameters (degrees of freedom) required to describe the elements of a hypothesis class determines the PAC-learnability as well as the sample complexity.

- This yields:

Corollary

A class is PAC-learnable if it has finite number of degrees of freedom.

- Better guess than the previous one.
- Still wrong.

Sines

Counterexample

The class $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$ where $h_\theta(x) = \lceil \sin(\theta x) \rceil, \forall x \in \mathbb{R}$ (consider $\lceil -1 \rceil = 0$) is not PAC-learnable.

- Apparently, correlation does not imply causation.
- We need to introduce a new measure.
- That's the VC-dimension!

Restriction of hypothesis class

Definition

Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, c_2, \ldots, c_m\} \subset \mathcal{X}$. 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:

$$\mathcal{H}_C = \{(h(c_1), h(c_2), \ldots, h(c_m)) : h \in \mathcal{H}\}$$

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

- Some of the functions in $\mathcal{H}$ may be the same when restricted to C.
- We refer to $|\mathcal{H}_C|$ as the effective size of $\mathcal{H}$ with respect to C.
- Clearly, $|\mathcal{H}_C| \leqslant |\mathcal{H}|$.

Shattering

- There are $2^{|C|}$ partitions of C.
- Each element of $\mathcal{H}_C$ corresponds to one of them.

Definition

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

The VC-dimension

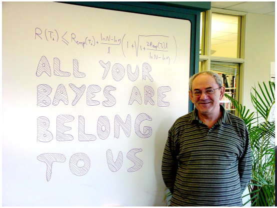
Definition

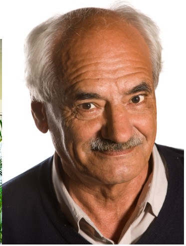
The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $VCdim\left(\mathcal{H}\right)$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.

- To show that $VCdim\left(\mathcal{H}\right) \geqslant d$, it suffices to find one $C \subset \mathcal{X}$ with $|C| = d$ that is shattered by $\mathcal{H}$.
- If $VCdim\left(\mathcal{H}\right) < d$, there is no $C \subset \mathcal{X}$ with $|C| = d$ that is shattered by $\mathcal{H}$.
- Based on the above, to show that $VCdim\left(\mathcal{H}\right) = d$, we have to show that $VCdim\left(\mathcal{H}\right) \geqslant d \wedge VCdim\left(\mathcal{H}\right) < d + 1$.

Notes

- Introduced by Vladimir Vapnik and Alexej Chervonenkis.





- Intuively, the VC-dimension of a hypothesis class is a combinatorial measure that quantifies its expressive power.

NFL reminder

- The NFL theorem stated:

Theorem

Let 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 $\frac{|\mathcal{X}|}{2}$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

- There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.
- With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geqslant \frac{1}{8}$.

VC and NFL

- An alternative formulation involving the VC-dimension is:

Corollary

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 $C \subset X$ of size $2m$ that is shattered by $\mathcal{H}$. Then, for any learning algorithm, A, there exist a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in H$ such that $L_{\mathcal{D}}(h) = 0$ but with probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geqslant \frac{1}{8}$.

- $VCdim(\mathcal{H}) \geqslant 2m$.
- The structure of $\mathcal{H}$ is such that, despite there being a hypothesis corresponding to distribution $\mathcal{D}$, we can't find it.

Infinite VC-dimension and PAC-learnability

Theorem

If a hypothesis class has infinite VC-dimension, it is not PAC-learnable.

- The above result is an immediate consequence of the previous corollary.
- Having a finite VC-dimension is a necessary condition for PAC-learnability. Is it also sufficient?
- Yes!
- The theorem will be presented after some examples of VC calculations.

1. Introduction to the VC-dimension
   - The road to the VC-dimension
   - Definition and Properties of the VC-dimension
   - Examples

2. The Fundamental Theorem of Statistical Learning

3. The end

Intervals

Example

Consider the class $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ with $h_a(x) = \mathbb{1}\{x \leqslant a\}, \forall x \in \mathbb{R}$. We have $VCdim(\mathcal{H}) = 1$.

Sketch of Proof

- It is easy to show that $\mathcal{H}$ shatters all sets with only one member.
- But when it comes to $C = \{x_1, x_2\}$ with $x_1 \leqslant x_2$, it's impossible to have the configuration $(1, 0)$.
- So $VCdim(\mathcal{H}) = 1$. ■

More intervals

Example

Consider the class $\mathcal{H} = \left\{h_{(a,b)} : a, b \in \mathbb{R}\right\}$ with $h_{(a,b)}(x) = \mathbb{1}\{x \in (a,b)\}, \forall x \in \mathbb{R}$. We have $VCdim(\mathcal{H}) = 2$.

Sketch of Proof

- It is easy to show that $\mathcal{H}$ shatters all sets with two members.
- But when it comes to $C = \{x_1, x_2, x_3\}$ with $x_1 \leqslant x_2 \leqslant x_3$, it's impossible to have the configuration $(1, 0, 1)$.
- So $VCdim(\mathcal{H}) = 2$. ■

Circles

Example

Consider the class $\mathcal{H}_r = \{h_r : r \in \mathbb{R}_+\}$ with $h_r(x) = \mathbb{1}\{||x|| \leqslant r\}, \forall x \in \mathbb{R}^2$.
We have $VCdim(\mathcal{H}) = 1$.

Sketch of Proof

- It is easy to show that $\mathcal{H}$ shatters all sets with only one member.
- But when it comes to $C = \{x_1, x_2\}$ with $||x_1|| \leqslant ||x_2||$, it's impossible to have the configuration $(1, 0)$.
- So $VCdim(\mathcal{H}) = 1$. ■

The Fundamental Theorem

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. Then, the following are equivalent:

- $\mathcal{H}$ has the uniform convergence property.
- Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.
- $\mathcal{H}$ is agnostic PAC learnable.
- $\mathcal{H}$ is PAC learnable.
- Any ERM rule is a successful PAC learner for $\mathcal{H}$.
- $\mathcal{H}$ has a finite VC-dimension.

- We are mainly interested in $(6) \Rightarrow (1)$.

Growth Function

Definition

Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as:

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

- For $m \leqslant VCdim(\mathcal{H})$ there is at least one set of cardinality m that is shattered by $\mathcal{H}$, so $\tau_{\mathcal{H}}(m) = 2^m$ (exponential to m).
- What about $m > VCdim(\mathcal{H})$?

Sauer's lemma

Lemma

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

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

In particular, if $m \geqslant d + 1$, then:

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

Remarks

- This answers our question about the growth function's values for $m > VCdim(\mathcal{H})$.
- Despite increasing exponentially at first, asymptotically, the growth function increases polynomially.
- Its value transcends machine learning theory.
- Stated and proven independently by Sauer, Shelah and Perles.
- Many alternative proofs and generalizations.
- Speaking of which, let's prove this!

Proof of Sauer's lemma

Proof

- Let $C = \{x_1, x_2, \ldots, x_m\} \subset \mathcal{X}$.
- Consider the set $\{B \subseteq C : \mathcal{H}\ shatters\ B\}$.
- The subsets of C shattered by $\mathcal{H}$ cannot be more than those with up to d elements.
- Therefore:

$$|\{B \subseteq C : \mathcal{H}\ shatters\ B\}| \leqslant \sum_{i=0}^{d} \binom{m}{i}$$

Proof of Sauer's lemma (cont'd)

Proof (cont'd)

- It suffices to prove that $\tau_{\mathcal{H}}(m) \leqslant |\{B \subseteq C : \mathcal{H} \; shatters \; B\}|$.
- An even more powerful statement is:

$$\forall C = \{x_1, x_2, \ldots, x_m\} \subset \mathcal{X}, \forall \mathcal{H}, |\mathcal{H}_C| \leqslant |\{B \subseteq C : \mathcal{H} \; shatters \; B\}|$$

- We will prove this by induction.
- The part involving the upper bound for $m \geqslant d + 1$ will not be presented here.

Proof of Sauer's lemma (cont'd)

Proof (cont'd)

- For $m = 1$ there are two cases according to d.
  - For $d = 0$:
  $$|\mathcal{H}_C| = 1 = \binom{1}{0}$$
  so the inequality holds.
  - For $d > 0$:
  $$|\mathcal{H}_C| = 2 = \binom{1}{0} + \binom{1}{1}$$
  so the inequality holds.
- The base case holds.

Proof of Sauer’s lemma (cont’d)

Proof (cont’d)

- Assume the lemma holds for all $k < m$.
- For some $\mathcal{H}$ and $C = \{c_1, c_2, \ldots, c_m\}$, denote $C' = \{c_2, \ldots, c_m\}$ and define the sets:

$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

- As a result, we have $\mathcal{H}_{C'} = Y_0$ and $|\mathcal{H}_C| = |Y_0| + |Y_1|$.

Proof of Sauer's lemma (cont'd)

Proof (cont'd)

- By the inductive hypothesis, we have:

$$|Y_0| = |\mathcal{H}_{C'}| \leqslant \left|\left\{B \subseteq C' : \mathcal{H}\ shatters\ B\right\}\right| =$$

$$= |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H}\ shatters\ B\}|\ (1)$$

- We define the set $\mathcal{H}'$, containing pairs of classifiers that differ on c_1:

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H}\ s.t.\ \left(1 - h'(c_1), h'(c_2), \ldots, h'(c_m)\right) =$$

$$= (h(c_1), h(c_2), \ldots, h(c_m))\} \subseteq \mathcal{H}$$

Proof of Sauer's lemma (cont'd)

Proof (cont'd)

- $\mathcal{H}'$ shatters $B \subseteq C' \Leftrightarrow \mathcal{H}'$ shatters $B \cup \{c_1\}$.
- $Y_1 = \mathcal{H}'_{C'}$ so by inductive hypothesis:

$$|Y_1| = \left|\mathcal{H}'_{C'}\right| \leqslant \left|\left\{B \subseteq C' : \mathcal{H}' \; shatters \; B\right\}\right| =$$

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

$$= \left|\left\{B \subseteq C : c_1 \in B \wedge \mathcal{H}' \; shatters \; B\right\}\right| =$$

$$= |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \; shatters \; B\}| \quad (2)$$

- Relations (1), (2) and $|\mathcal{H}_C| = |Y_0| + |Y_1|$ complete the proof. ■

Uniform Convergence for Classes of 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:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leqslant \frac{4 + \sqrt{\log\left(\tau\left(2m\right)\right)}}{\delta\sqrt{2m}}$$

- No proof.
- We will use this, along with Sauer's lemma to complete the proof of the fundamental theorem.

Proof of the fundamental theorem

Proof

- Combining Sauer's lemma with the previous theorem, we get that, with probability at least $1 - \delta$:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leqslant \frac{4 + \sqrt{d \log\left(\frac{2em}{d}\right)}}{\delta\sqrt{2m}}$$

- Assume that m is big enough so that:

$$\sqrt{d \log\left(\frac{2em}{d}\right)} \geqslant 4$$

Proof of the fundamental theorem (cont'd)

Proof (cont'd)

- The previous give:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leqslant \frac{1}{\delta}\sqrt{\frac{2d\log\left(\frac{2em}{d}\right)}{m}}$$

- We demand the previous to be less than ϵ.
- A sufficient condition for that is:

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

■

Any questions?

Thank you!