VC Dimension

AW

Lecture Overview

Learnability of Infinite classes

We saw that countable union of agnostically learnable (finite) classes are learnable. But what about classes of size $\aleph_1$ classes (classes of same size as real numbers)?

Example: Suppose we need to classify $x \in X \subset \mathbb{R}$. Let $\mathscr{H}$ be the set of threshold functions over the real line: $\mathscr{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0, 1\}$ such that $h_a(x) = \begin{cases} 1 \text{ if } x < a \\ 0 \text{ otherwise} \end{cases}$ Obviously this set is of same size as real numbers.

Proposition

Let $\mathscr{H}$ be the class of threshold functions. Then, $\mathscr{H}$ is PAC learnable, using the ERM rule, with sample complexity of $m_H(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof in SSBD-6.1.

So why is this class learnable but no-free lunch theorem implies that other 'continuous' classes are not learnable?

Shattering

Let $\mathscr{H}$ be a class of functions from (data)set X to $\mathbb{B} = \{0, 1\}$ and let $C = \{c_1, \ldots, c_m\} \subset X$ (i.e. sample). The restriction of $\mathscr{H}$ to C (denoted $\mathscr{H}_C$) is the set of functions from C to $\mathbb{B}$ that can be derived from $\mathscr{H}$. We can identify each function $f \in \mathscr{H}_C$ with m dimensional vector v^f from $\mathbb{B}^m$ in which $v_i^f = f(c_i)$.
If the restriction of $\mathscr{H}$ to C is the set of all functions from C to $\mathbb{B}$ then we say that $\mathscr{H}$ shatters set C

Definition

A hypothesis class $\mathscr{H}$ shatters a finite set $C \subset X$ if the restriction $\mathscr{H}_c$ contains all maps $C \to \mathbb{B}$, i.e. $|\mathscr{H}_C| = 2^{|C|}$.

Why shattering matters? Proof of No-Free-Lunch theorem shows that without restricting the hypothesis class on a set of size $= 2\times$sample size:

- an adversary can construct a distribution for which a learning algorithm *against which adversary is working* will perform poorly
- there is another learning algorithm (against which adversary is not working) that succeeds on the constructed distribution

Understand No-Free-Lunch

Another way to understand no-free lunch theorem:

Proposition

For a hypothesis class $\mathscr{H}$ of functions in $\{f : \mathrm{X} \to \mathbb{B}\}$, let m be a training set size. Suppose there exists a set $\mathrm{C} \subset \mathrm{X}$ of size $2m$ that is shattered by $\mathscr{H}$. Then, for any learning algorithm, A, there exist a distribution D over $\mathrm{X} \times \mathbb{B}$ and a predictor $h \in \mathscr{H}$ such that $\mathrm{L}_{\mathrm{D}}(h) = 0$ but with probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$ we have that $L_D(A(S)) \geq \frac{1}{8}$.

Meaning: If a set C is shattered by $\mathscr{H}$, and we receive a sample S containing half the instances of C, the labels of these instances in S give us no information about the labels of the rest of the instances in C because every possible labeling of the rest of the instances in $C - S$ can be explained by some hypothesis in $\mathscr{H}$.

VC-dimension

Shattering Example: $\mathcal{H}$ =class of threshold functions over $\mathbb{R}$.

- $C = \{c_1 \in \mathbb{R}\}$. Let $a_1 = c_1 + 1$, then $h_{a_1}(c_1) = 1$, and now let $a_2 = c_1 - 1$, then we have $h_{a_2}(c_1) = 0$. Therefore, $\mathcal{H}_C$ is the set of all functions from C to $\mathbb{B}$, and $\mathcal{H}$ shatters C.
- $C = \{c_1, c_2\}$ where $c_i \in \mathbb{R}$ and $c_1 < c_2$. No $h_a \in \mathcal{H}$ can account for the labeling $[(c_1, 1), (c_2, 0)]$, because any threshold a that assigns the label 0 to c_2 must assign the label 0 to c_1 as well. Therefore not all functions from C to $\mathbb{B}$ are included in $\mathcal{H}_C$ so C is not shattered by $\mathcal{H}$.

Definition (VC-dimension)

The VC-dimension of a hypothesis class $\mathcal{H}$ (denoted $VC\dim(\mathcal{H})$), is the maximal size of a set $C \subset X$ 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.

$VC\dim$ and Non-learnability, Examples

Proposition

Let $\mathscr{H}$ be a class of infinite VC-dimension. Then, $\mathscr{H}$ is not PAC learnable.

Proof is obvious: for any size sample we can shatter a set twice the size, so with probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$ we have that $L_D(A(S)) \geq \frac{1}{8}$ for any algorithm A - hence non learnable (can't learn for $(\epsilon, \delta) < (1/8, 1/7)$).

$VC\dim$ example:

- We already worked it out for threshold predictors $\mathscr{H} = \{h_a : a \in \mathbb{R}\}$ for data $x \in \mathbb{R}$. We have shown that they shatter any set of size one but cannot shatter any set of size 2. So $VC\dim(\mathscr{H}) = 1$

More Examples of $VC \dim$ Computation

$VC \dim$ examples:

1. Interval predictors: $\mathscr{H} = \{h_{a,b} : a, b \in R \text{ and } a < b\}$ for data $x \in \mathbb{R}$. Here $h_{a,b} : \mathbb{R} \to \mathbb{B} : x \to \begin{cases} 1 \text{ if } a < x < b \\ 0 \text{ otherwise} \end{cases}$ Let $C = \{c_1, c_2\}$. WLOG $c_1 < c_2$, then shattering maps for C are $h_{a,b}(c_1) = h_{a,b}(c_2) = 0$ if $a, b < c_1$; $h_{a,b}(c_1) = 0, h_{a,b}(c_2) = 1$ if $c_1 < a < c_2 < b$; $h_{a,b}(c_1) = 1, h_{a,b}(c_2) = 0$ if $a < c_1 < b < c_2$; $h_{a,b}(c_1) = h_{a,b}(c_2) = 1$ if $a < c_1 < c_2 < b$. Can $C = \{c_1, c_2, c_3\}$ be shattered by intervals?

More Examples of $VC \dim$ Computation

$VC \dim$ examples:

1. Interval predictors: $\mathscr{H} = \{h_{a,b} : a, b \in R$ and $a < b\}$ for data $x \in \mathbb{R}$. Here $h_{a,b} : \mathbb{R} \to \mathbb{B} : x \to \begin{cases} 1 \text{ if } a < x < b \\ 0 \text{ otherwise} \end{cases}$ Let $C = \{c_1, c_2\}$. WLOG $c_1 < c_2$, then shattering maps for C are $h_{a,b}(c_1) = h_{a,b}(c_2) = 0$ if $a, b < c_1$; $h_{a,b}(c_1) = 0, h_{a,b}(c_2) = 1$ if $c_1 < a < c_2 < b$; $h_{a,b}(c_1) = 1, h_{a,b}(c_2) = 0$ if $a < c_1 < b < c_2$; $h_{a,b}(c_1) = h_{a,b}(c_2) = 1$ if $a < c_1 < c_2 < b$. Can $C = \{c_1, c_2, c_3\}$ be shattered by intervals? If $C = \{c_1, c_2, c_3\}$ for $c_1 < c_2 < c_3$ then map $1, 0, 1$ is not obtainable in $\mathscr{H}$ since any interval that contains c_1 and c_3 also contains c_2. Thus $VC \dim(\mathscr{H}) = 2$.

More Examples of $VC\,\dim$ Computation

$VC\,\dim$ examples:

1. Interval predictors
2. An axis-aligned n-dimensional rectangle classifier $h_{(\bar{l},\overline{u})}$ is given by two vectors $\bar{l}, \overline{u} \in \mathbb{R}^n$ such that $\bar{l} < \overline{u}$ (i.e. $l_i < u_i$ for all $1 \leq i \leq n$). A vector $\overline{x} \in \mathbb{R}^n$ is labeled 1 by this classifier $h_{(\bar{l},\overline{u})}$ if $\bar{l} < \overline{x} < \overline{u}$, i.e. for every i holds $l_i < x_i < u_i$. Otherwise $\overline{x}$ is labeled 0.

More Examples of $VC \dim$ Computation

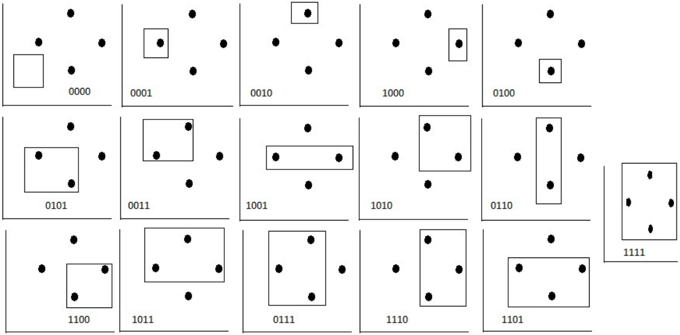
$VC \dim$ examples:

1. Interval predictors
2. Let $\mathscr{H}$ be a class of 2-dimensional axis-aligned rectangle classifiers.
   - Let $C = \{(x_1, y_1), (x_2, y_2), (x_2, y_3), (x_3, y_1)\}$ where $x_1 < x_2 < x_3$ and $y_2 < y_1 < y_3$. Can it be shattered?

More Examples of $VC\,\dim$ Computation

$VC\,\dim$ examples:

1. Interval predictors
2. Let $\mathscr{H}$ be a class of 2-dimensional axis-aligned rectangle classifiers.
   - Let $C = \{(x_1, y_1), (x_2, y_2), (x_2, y_3), (x_3, y_1)\}$ where $x_1 < x_2 < x_3$ and $y_2 < y_1 < y_3$. It is shattered:



More Examples of $VC \dim$ Computation

$VC \dim$ examples:

1. Interval predictors
2. Let $\mathscr{H}$ be a class of 2-dimensional axis-aligned rectangle classifiers.
   - Let $C = \{(x_1, y_1), (x_2, y_2), (x_2, y_3), (x_3, y_1)\}$ where $x_1 < x_2 < x_3$ and $y_2 < y_1 < y_3$. This set is shattered.
   - Let $C = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5)\}$ be set of 5 points.Can this set be shattered?

More Examples of $VC\,\dim$ Computation

$VC\,\dim$ examples:

1. Interval predictors
2. Let $\mathscr{H}$ be a class of 2-dimensional axis-aligned rectangle classifiers.
   - Let $C = \{(x_1, y_1), (x_2, y_2), (x_2, y_3), (x_3, y_1)\}$ where $x_1 < x_2 < x_3$ and $y_2 < y_1 < y_3$. This set is shattered.
   - Let $C = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5)\}$ be set of 5 points. No! Let $x_{\max} = \max_i x_i$, $x_{\min} = \min_i x_i$, $y_{\max} = \max_i y_i$ and $y_{\min} = \min_i y_i$.
     1. Select 4 points $c_1, \ldots, c_4$ out of 5 so that for each number in the list $x_{\max}, x_{\min}, y_{\max}, y_{\min}$ there is at least one point among selected points that it as a coordinate
     2. Label selected points $c_1, \ldots c_4$ by 1 and the point $c_5 = (x_{ns}, y_{ns})$ that was not selected by 0.

     It is impossible to obtain this labeling by an axis-aligned rectangle since $x_{\max} \geq x_{ns} \geq x_{\min}$ and $y_{\max} \geq y_{ns} \geq y_{\min}$, so if all other point are inside the rectangle this one should be there too.

Lecture Overview

1. VC-dimension
2. VC-dimension and Learnability

The Fundamental Theorem of Statistical Learning

Theorem

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathbb{X}$ to $\mathbb{B}$. Then, the following are equivalent:

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

Proof in SSBD chapter 6.5

Relation of VC-dimension to Sample Complexity

Theorem

Let $\mathscr{H}$ be a hypothesis class of functions from a domain X to $\mathbb{B}$ such that $VC\dim(H) = d < \infty$. Then, there are absolute constants C_1 and C_2 such that:

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

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

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

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

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

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

Proof in SSBD chapter 6.5

VC-dimension and uniform convergence

For hypothesis class $\mathcal{H}$ growth function of $\mathcal{H}$ (denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$) is defined as $\tau_{\mathcal{H}}(m) = \max_{\substack{C \subset X \\ |C| = m}} |\mathcal{H}_C|$

Lemma

Let $\mathcal{H}$ has $VC\dim(\mathcal{H}) = d < \infty$. Then, for all m, $\tau_{\mathcal{H}}(m) = \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

Theorem

Let $h \in \mathcal{H}$. Then, for every D and every $\delta \in (0,1)$, with probability of at least $1-\delta$ over the choice of training set $S \sim D^m$ we have

$$|L_D(h) - L_S(h)| \leq \frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

Combining, we get that for $\mathcal{H}$ that has $VC\dim(\mathcal{H}) = d < \infty$ and $m > d+1$ holds

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

For proofs see SSBD chapter 6.5

Reading

SSBD sections 6.2, 6.3, 6.4

Proofs can be omitted without any loss of understanding