VC Dimension

Kontonis Vasilis

27 April, 2017

Corelab, NTUA

Contents

The VC-Dimension

Learnable Infinite-Size Classes

Theorem
Let $\mathcal{H} = \{\mathbb{1}_{[x<a]} : a \in \mathbb{R}\}$. Then $\mathcal{H}$ is PAC learnable, using the ERM rule, with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leqslant \lceil \log(2/\delta)/\epsilon \rceil$.

Proof.

- Let $h^*(x) = \mathbb{1}_{[x<a=*]}$, with $L_D(h^*) = 0$.
- Let $a_0 < a^* < a_1$:

$$D_x(\{x \in (a_0, a^*)\}) = D_x(\{x \in a^*, a_1)\}) = \epsilon$$

- Given training data S, let $b_0 = \max\{x : (x, 1) \in S\}, b_1 = \min\{x : (x, 0) \in S\}$.
- Let $b_S \in (b_0, b_1)$ be an ERM hypothesis.
- $b_0 \geqslant a_0,\ b_1 \leqslant a_1 \implies L_D(h_S) \leqslant \epsilon$. Therefore,

$$D^m\left(L_D(h_S) > \epsilon\right) \leqslant D^m\left(b_0 < a_0 \vee b_1 > a_1\right)$$

- $D^m(b_0 < a_0) \leqslant e^{-\epsilon m}$, $D^m(b_1 > a_1) \leqslant e^{-\epsilon m}$.

Shattering

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

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

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

Definition (Shattering)
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 C to $\{0, 1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.

VC Dimension

Threshold Functions
Let $\mathcal{H}$ be the threshold functions over $\mathbb{R}$. $C = \{c_1\}$ is shattered by $\mathcal{H}$. A set $C = \{c_1, c_2\}, c_1 \leqslant c_2$ is no shattered by $\mathcal{H}$. Consider the labeling $(0, 1)$.

Definition (VC-Dimension)
The VC-dimension of a hypothesis class $\mathcal{H}$, 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, then $\mathcal{H} = \infty$.

NFL Revisited

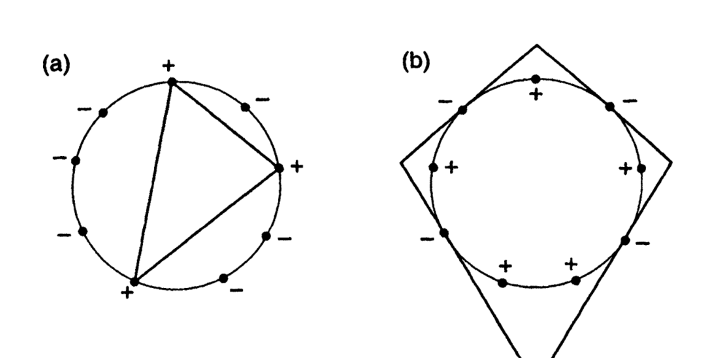
Let m be a training set size. If there exists a set $C \subset \mathcal{X}$ s.t. $|C| = 2m$ and $|\mathcal{H}_C| = 2^{2m}$,that is C is shattered by $\mathcal{H}$, then we cannot learn $\mathcal{H}$ with m samples.

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

Proof.
For any training set of m samples there exists a training set of $2m$ samples shattered by $\mathcal{H}$ $\rightarrow$ No Free Lunch. $\square$

Examples

- Threshold Functions on $\mathbb{R}$: 1.
- Intervals on $\mathbb{R}$: 2.
- Axis Aligned Rectangles on $\mathbb{R}^2$: 4.
- Hyperplanes in $\mathbb{R}^d$: $d+1$.
- Convex d-gons on $\mathbb{R}^2$: $2d+1$.
- Convex polygons on $\mathbb{R}^2$: ∞.
- $\{\sin(\theta x) : \theta \in \mathbb{R}\}$ on $\mathbb{R}$: ∞.



Growth Function

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

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

Intervals
Let $\mathcal{H}$ be the class of intervals on $\mathbb{R}$, then

$$\tau_{\mathcal{H}}(1) = 2$$
$$\tau_{\mathcal{H}}(2) = 4$$
$$\tau_{\mathcal{H}}(3) = 7$$
$$\tau_{\mathcal{H}}(4) = 11$$

In general $\tau_{\mathcal{H}}(n) = (n+1)n/2 + 1 = O(n^2)$

Sauer's Lemma and UC

Lemma (Sauer's Lemma)
Let $\mathcal{H}$ *be a hypothesis class with* $\mathrm{VCdim}(\mathcal{H}) \leqslant d < \infty$. *Then, for all* m, $\tau_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \binom{m}{i}$. *In particular, if* $m > d+1$ *then* $\tau_{\mathcal{H}}(m) \leqslant (em/d)^d$.

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

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

The Fundamental Theorem of Statistical Learning

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:

1. *$\mathcal{H}$ has the UC 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.*

Quantitative Version

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

1. $\mathcal{H}$ has the UC property/is APAC learnable with sample complexity:

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

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

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

Constraint Sampling

Let $\mathcal{Z}$ be a feasible set of linear constraints

$$\gamma_z^T + k_z \geqslant 0, z \in \mathcal{Z}$$

where $K\ell^*|\mathcal{Z}|$.

Claim
The feasible region specified by all constraints can be closely approximated by a sampled subset.

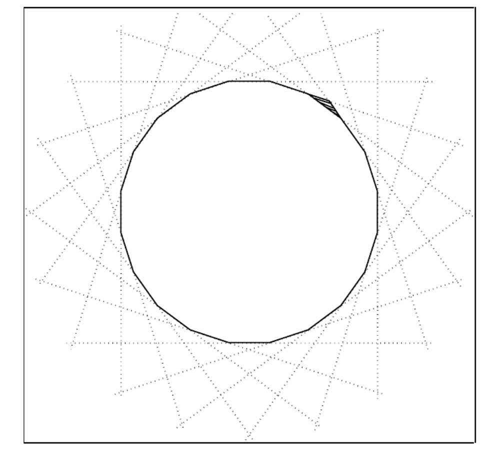
Let ψ be a probability measure over $\mathcal{Z}$. Given $\epsilon \in (0, 1)$ we want a $\mathcal{W} \subseteq \mathcal{Z}$ such that

$$\sup_{\{r \,:\, \gamma_z^T r + k_z \geqslant 0, z \in \mathcal{W}\}} \psi(\{y : \gamma_y^T r + k_y < 0\}) \leqslant \epsilon$$

$$\mathrm{VCdim}(\{\{(\gamma, k) : \gamma^T r + k \geqslant 0\} \;:\; r \in \mathbb{R}^K\}) = K$$

Constraint Sampling

Worst Case Constraint Set:



Questions?

M. J. Kearns and U. Vazirani.
An Introduction to Computational Learning Theory.
The MIT Press, Cambridge, Mass, Aug. 1994.

M. Mohri, A. Rostamizadeh, and A. Talwalkar.
Foundations of Machine Learning.
The MIT Press, Cambridge, MA, Aug. 2012.

S. Shalev-Shwartz and S. Ben-David.
Understanding Machine Learning: From Theory to Algorithms.
Cambridge University Press, New York, NY, USA, 1 edition edition, May 2014.