Understanding Machine Learning – A theory Perspective

Shai Ben-David

University of Waterloo

MLSS at MPI Tubingen, 2017

Some infinite classes are learnable

Examples:

- Initial segments of the real line.
- The class of singletons over any domain set

Other classes of the same "size" are not learnable

- The class of all finite subsets of an infinite domain.

Proof: *Note that for possible value for* $m=m_H(1/8,1/8)$ *there is a domain subset* A_m *of double the size for which every possible* $F{:}A_m \rightarrow \{0,1\}$ *agrees with some h in H.*

A combinatorial characterization of PAC learnable classes

Shattering:

A class H shatters a domain subset A if
For every B subset of A
There is some h_B in H so that for all x in A
$h_B(x)=1$ if and only if x is in B.

- **Examples:**

The Vapnik Chervonenkis dimension

Given a class of binary valued functions, H,
The Vapnik-Chervonenkis dimension of H is

$$VCdim(H) = \sup \{|A| : H \text{ shatters } A\}$$

First connection to PAC learning

Note that our proof of the No Free Lunch Theorem shows, in fact, that:

For any class H, $m_H(1/8, 1/8) > \mathrm{VCdim}(H)/2$

Corollary: If VCdim(H) is infinite then H is not PAC learnable.

The fundamental theorem (qualitative)

Theorem: Given a class H of binary valued functions the following statements are equivalent:

a) *H has the Uniform Convergence Property*
b) *ERM is an agnostic PAC learner for H*
c) *H is agnostic PAC learnable*
d) *H is PAC learnable*
e) *VCdim(H) is finite*

Main tool for (e) implies (a)

The Shatter function

For a class H define a function $\Pi_H: N \rightarrow N$ as $\Pi_H(m) = \max_{\{A: |A|=m\}} |\{h|_A : h \text{ in } H\}|$

Some basic properties of the shatter function:

1. For every $m \leq VCdim(H)$, $\Pi_H(m) = 2^m$
2. For every $m > VCdim(H)$, $\Pi_H(m) < 2^m$

The Sauer (Shelah, Perles) lemma

For every class H of finite VC-dimension, d,

For every m,

$$\Pi_H(m) \leq \Sigma_{i=0}^{d} \text{ (m choose i)} \leq m^d$$

A typical corollary

The number of linear partitions of a set of points in the plain.

Quantitative version of the Fundamental Theorem

For some constants C_1, C_2, for every d and every class H of binary valued functions such that VCdim(H)=d,

1. H has Uniform Convergence property with

   $C_1(d+\log(1/\delta))/\varepsilon^2 < m^{uc}{}_H(\varepsilon, \delta) < C_1(d+\log(1/\delta))/\varepsilon^2$

2. H is agnostic PAC learnable with

   $C_1(d+\log(1/\delta))/\varepsilon^2 < m_H(\varepsilon, \delta) < C_1(d+\log(1/\delta))/\varepsilon^2$

3. H is PAC learnable with

   $C_1(d+\log(1/\delta))/\varepsilon < m_H(\varepsilon, \delta) < C_1(d+\log(1/\delta))/\varepsilon$