Learning Theory Overview

Dimitris Diochnos

University of Virginia
Department of Computer Science

Spring 2017
CS 6501 - Learning Theory
Outline

1. Preliminaries
2. PAC Learning and VC-Dimension
Learning Theory in One Line

Find a Good Approximation of a Function with High Probability
Learning Theory

Goal (Good Approximation with High Probability)
There is a function \( c \) over a space \( X \). One wants to come up (in a reasonable amount of time) with a function \( h \) such that \( h \) is a good approximation of \( c \) with high probability.

Description (Parameters and Terminology)
- \( X \): Instance Space
- \( c \in C \): Target Concept
- \( h \in H \): Hypothesis
- Good Approximation: Small Error \( \varepsilon \)
- High Probability: Confidence \( 1 - \delta \)
- Reasonable Amount of Time: Polynomial in \( n, 1/\varepsilon, 1/\delta \)

Example
\[
X = \{0, 1\}^n \quad \quad \quad c = x_1 \land x_2 \land x_3 \quad \quad \quad h = x_1 \land x_4
\]
Probably Approximately Correct (PAC) Learning

- There is an arbitrary, unknown distribution $\mathcal{D}$ over $X$.
- Learn from examples $(x, c(x))$, where $x \sim \mathcal{D}$.
- $\text{error}(h, c) = \Pr(h(x) \neq c(x))$.

Goal (Valiant, 1984)

$$\Pr(\text{error}(h, c) \leq \varepsilon) \geq 1 - \delta.$$
Efficiently PAC Learning Conjunctions

Let $X = \{x_1, x_2, x_3, x_4, x_5\}$ and $c = x_1 \land \overline{x}_3 \land x_4$.

- Request $m$ examples and look on the positive ones.

<table>
<thead>
<tr>
<th>example</th>
<th>hypothesis $h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>((11010), +)</td>
<td>$x_1 \land \overline{x}_1 \land x_2 \land \overline{x}_2 \land x_3 \land \overline{x}_3 \land x_4 \land \overline{x}_4 \land x_5 \land \overline{x}_5$</td>
</tr>
<tr>
<td>((10010), +)</td>
<td>$x_1 \land x_2 \land \overline{x}_3 \land x_4 \land \overline{x}_5$</td>
</tr>
<tr>
<td>((10011), +)</td>
<td>$x_1 \land \overline{x}_3 \land x_4$</td>
</tr>
<tr>
<td></td>
<td>$x_1 \land \overline{x}_3 \land x_4$</td>
</tr>
</tbody>
</table>

Theorem (PAC Learning of Finite Concept Classes)

For every distribution $D$, drawing $m \geq \frac{1}{\varepsilon} \cdot \left( \ln |C| + \ln \frac{1}{\delta} \right)$ examples guarantees that any consistent hypothesis $h$ satisfies $\Pr(\text{error}(h, c) \leq \varepsilon) \geq 1 - \delta$.

- For conjunctions $|C| = 3^n + 1$.
- Efficiently PAC learning because the algorithm runs in poly-time.
- What about infinite concept classes (e.g. halfspaces)?
Different Classifications and the Growth Function

- $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is a set of $m$ examples.

**Number of Classifications $\Pi_H(x)$ of $x$ by $H$:** Distinct vectors $(h(x_1), h(x_2), \ldots, h(x_m))$ as $h$ runs through $H$.

- $\Pi_H(x) \leq 2^m$. 

Different Classifications and the Growth Function

- \( x = (x_1, x_2, \ldots, x_m) \) is a set of \( m \) examples.

**Number of Classifications** \( \Pi_\mathcal{H}(x) \) **of** \( x \) **by** \( \mathcal{H} \): Distinct vectors

\( (h(x_1), h(x_2), \ldots, h(x_m)) \) as \( h \) runs through \( \mathcal{H} \).

- \( \Pi_\mathcal{H}(x) \leq 2^m \).

**Growth Function**: \( \Pi_\mathcal{H}(m) = \max\{\Pi_\mathcal{H}(x) : x \in X^m\} \).

**Example**

Rays on a line:

\( h_\vartheta(x) = \begin{cases} + & \text{if } x \geq \vartheta \\ - & \text{otherwise} \end{cases} \)

\( \Pi_\mathcal{H}(m) = m + 1 \).
The Vapnik-Chervonenkis Dimension

**Definition**

A sample \( x \) of size \( m \) is *shattered* by \( \mathcal{H} \), or \( \mathcal{H} \) *shatters* \( x \), if \( \mathcal{H} \) can give all \( 2^m \) possible classifications of \( x \).

**Definition (VC dimension)**

\[
VC\text{-dim}(\mathcal{C}) = \max \{ m : \prod_{\mathcal{C}}(m) = 2^m \}
\]

- Our ray example has \( VC\text{-dim}(\text{Rays}) = 1 \).
  - One point is shattered.
  - Two points are not shattered (+, −)

- Lower Bound \( \implies \) Explicit construction that achieves \( 2^m \).
- Upper Bound \( \implies \) For any sample \( x \) of length \( m \) we can not achieve \( 2^m \).
Configurations of 3 Points in 2D
Halfspaces Shatter 3 Points in 2D

Question
Can we shatter 4 points?
Can Halfspaces Shatter 4 Points in 2D?
Halfspaces can *not* Shatter 4 Points in 2D

Theorem (Radon)

Any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two (disjoint) sets whose convex hulls intersect.

Corollary

- $VC\text{-}dim(\text{HALFSPACES}) = 3$ in 2 dimensions.
- $VC\text{-}dim(\text{HALFSPACES}) = d + 1$ in $d \geq 1$ dimensions.
Sauer’s Lemma

Lemma (Sauer’s Lemma)

Let $d \geq 0$ and $m \geq 1$ be given integers and let $\mathcal{H}$ be a hypothesis space with $VC\text{-dim}(\mathcal{H}) = d$. Then

$$\Pi_{\mathcal{H}}(m) \leq 1 + \binom{m}{1} + \binom{m}{2} + \cdots + \binom{m}{d} = \Phi(d, m).$$

Proposition

For all $m \geq d \geq 1$, $\Phi(d, m) < \left(\frac{em}{d}\right)^d$. 
VC-Dimension

Theorem

Let $\mathcal{C}$ have finite $\text{VC-dim}(\mathcal{C}) = d \geq 1$ and moreover let $0 < \delta, \varepsilon < 1$. Then,

$$m \geq \left\lceil \frac{4}{\varepsilon} \cdot \left( d \cdot \lg \left( \frac{12}{\varepsilon} \right) + \lg \left( \frac{2}{\delta} \right) \right) \right\rceil$$

samples guarantee that any consistent hypothesis has small error with high probability (in the PAC-learning sense).

- We still need an efficient algorithm to efficiently PAC-learn the class.