Computational Learning Theory Overview

Dimitris Diochnos

University of Oklahoma
School of Computer Science

CS 5970 – Computational Learning Theory
Fall 2020

Outline

Learning Theory in One Line

Find a Good Approximation of a Function
with High Probability

Computational 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)

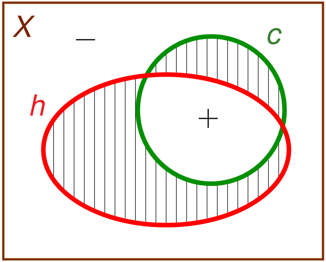
- X: Instance Space
- $c \in \mathcal{C}$: Target Concept — $h \in \mathcal{H}$: Hypothesis
- Good Approximation: Small Error ε
- High Probability: Confidence $1 - \delta$
- Reasonable Amount of Time: Polynomial in n, $1/\varepsilon$, $1/\delta$, size(c)

Example

$X = \{0, 1\}^n$ $\quad$ $c = x_1 \wedge x_2 \wedge x_3$ $\quad$ $h = x_1 \wedge x_4$

Probably Approximately Correct (PAC) Learning

- There is an *arbitrary, unknown* distribution $\mathcal{D}$ over X.
- Learn from *poly* $\left(\frac{1}{\varepsilon}, \frac{1}{\delta}\right)$ many examples $(x, \mathrm{c}(x))$, where $x \sim \mathcal{D}$.
- $\mathrm{Risk}_{\mathcal{D}}(\mathrm{h}, \mathrm{c}) = \mathbf{Pr}_{x\sim\mathcal{D}}(\mathrm{h}(x) \neq \mathrm{c}(x))$.



Goal ([Valiant, 1984])

$$\mathbf{Pr}\left(\mathrm{Risk}_{\mathcal{D}}(\mathrm{h}, \mathrm{c}) \leqslant \varepsilon\right) \geqslant 1 - \delta\ .$$

Efficiently PAC Learning Conjunctions

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

- Request m examples and look on the positive ones.

example	hypothesis h
	$x_1 \wedge \overline{x}_1 \wedge x_2 \wedge \overline{x}_2 \wedge x_3 \wedge \overline{x}_3 \wedge x_4 \wedge \overline{x}_4 \wedge x_5 \wedge \overline{x}_5$
$((11010), +)$	$x_1 \wedge x_2 \wedge \overline{x}_3 \wedge x_4 \wedge \overline{x}_5$
$((10010), +)$	$x_1 \wedge \overline{x}_3 \wedge x_4 \wedge \overline{x}_5$
$((10011), +)$	$x_1 \wedge \overline{x}_3 \wedge x_4$

Theorem (PAC Learning of Finite Concept Classes)

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

- For conjunctions $|\mathcal{C}| = 3^n + 1$.
- 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 $\Pi_{\mathcal{H}}(\mathbf{x})$ of $\mathbf{x}$ by $\mathcal{H}$: Distinct vectors $(\mathrm{h}(x_1), \mathrm{h}(x_2), \ldots, \mathrm{h}(x_m))$ as h runs through $\mathcal{H}$.

- $\Pi_{\mathcal{H}}(\mathbf{x}) \leqslant 2^m$.

Different Classifications and the Growth Function

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

Number of Classifications $\Pi_{\mathcal{H}}(\mathbf{x})$ of $\mathbf{x}$ by $\mathcal{H}$: Distinct vectors $(\mathrm{h}(x_1), \mathrm{h}(x_2), \ldots, \mathrm{h}(x_m))$ as h runs through $\mathcal{H}$.

- $\Pi_{\mathcal{H}}(\mathbf{x}) \leqslant 2^m$.

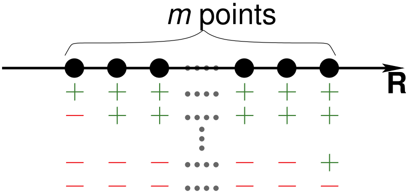
Growth Function: $\Pi_{\mathcal{H}}(m) = \max\{\Pi_{\mathcal{H}}(\mathbf{x}) \ : \ \mathbf{x} \in X^m\}$.

Example

Rays on a line:

$$\mathrm{h}_\vartheta(x) = \begin{cases} + & , \quad \text{if } x \geqslant \vartheta \\ - & , \quad \text{otherwise} \end{cases}$$

$\Pi_{\mathcal{H}}(m) = m + 1$.



The Vapnik-Chervonenkis Dimension

Definition

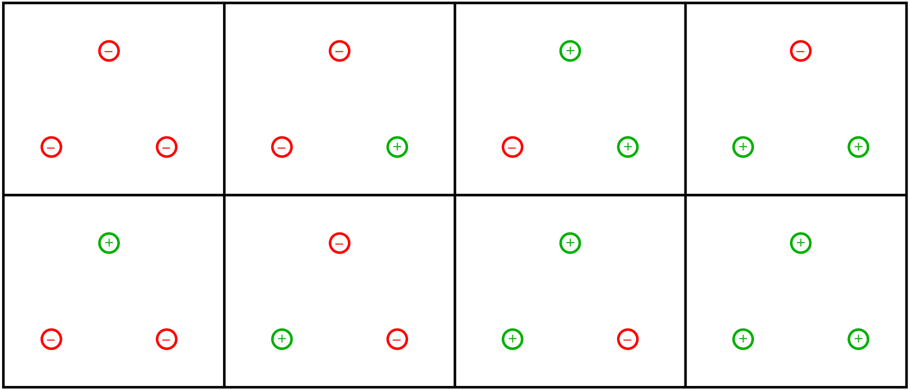
A sample $\mathbf{x}$ of size m is *shattered* by $\mathcal{H}$, or $\mathcal{H}$ *shatters* $\mathbf{x}$, if $\mathcal{H}$ can give all 2^m possible classifications of $\mathbf{x}$.

Definition (*VC* dimension)

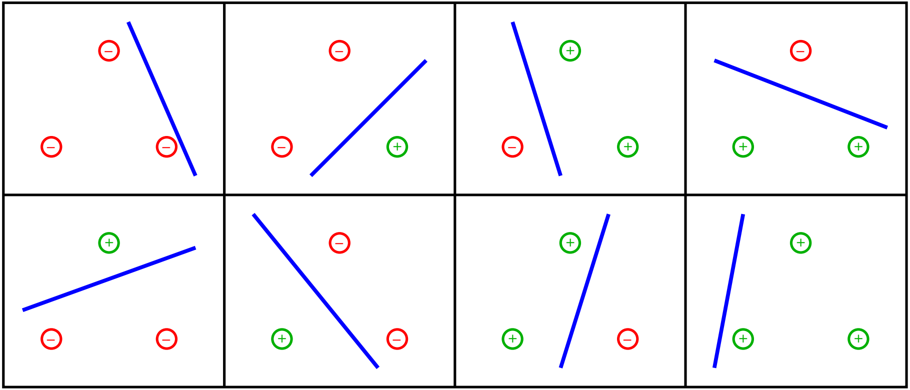
$$\textit{VC-dim}\,(\mathcal{C}) = \max\{m \;:\; \Pi_{\mathcal{C}}(m) = 2^m\}$$

- Our ray example has $\textit{VC-dim}\,(\text{Rays}) = 1$.
  - One point is shattered.
  - Two points are not shattered $(+, -)$
- Lower Bound $\Longrightarrow$ Explicit construction that achieves 2^m.
- Upper Bound $\Longrightarrow$ For *any* sample $\mathbf{x}$ of length m we can not achieve 2^m.

Configurations of 3 Points in 2D



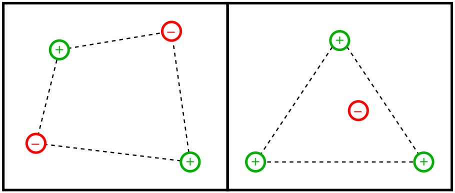
Halfspaces Shatter 3 Points in 2D



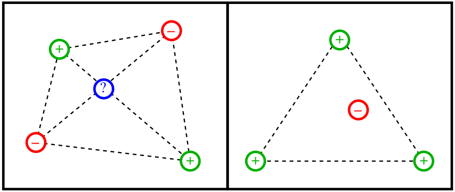
Question

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 $\mathbf{R}^d$ can be partitioned into two (disjoint) sets whose convex hulls intersect.

Corollary

- *VC-dim* $(HALFSPACES) = 3$ *in* 2 *dimensions.*
- *VC-dim* $(HALFSPACES) = d + 1$ *in* $d \geqslant 1$ *dimensions.*

Sauer's Lemma

Lemma (Sauer's Lemma)

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

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

Proposition

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

VC-Dimension

Theorem

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

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

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.