# CS 760: Machine Learning
## **Learning Theory**

Josiah Hanna

University of Wisconsin-Madison

**November 16, 2023**

# Announcements

- HW 6 due Tuesday.
- Midterm regrade deadline is tonight.
    - For grading mistakes not arguing for partial credit.

# Outline

- **Learning Theory Motivation**
  - Questions to answer
- **PAC-learning**
  - Mistake bounds.
- **VC Dimension**
  - Definition, why useful

# Outline

- **Learning Theory Motivation**
  - Questions to answer
- **PAC-learning**
  - Mistake bounds.
- **VC Dimension**
  - Definition, why useful

# Why learning theory?

- Formal analysis of algorithms is important in all areas of computer science.
    - Example: binary search has time complexity $O(\log n)$.

- Desire a rigorous understanding of algorithms:
    - Be able to predict how an algorithm will work on new problems.
    - Understand when a problem is inherently hard (lower bounds).
    - Understand when a problem can be learned efficiently (time, space, training set size).
    - Provide guarantees on performance under certain conditions.

# Outline

- **Learning Theory Motivation**
  - Questions to answer
- **PAC-learning**
  - Mistake bounds.
- **VC Dimension**
  - Definition, why useful

# Formal Definition of Learning

- X: set of all possible inputs.
- $c : X \to \{0,1\}$ is the target **concept** to learn.
- $C$: a set of possible target concepts.
- $D$: a probability distribution over X.
  - $\sum_{x \in X} D(x) = 1$ and $\forall x, D(x) \geq 0$.
- $S$: a training sample of size m, $\{(x_i, c(x_i))\}_{i=1}^{m}$

# Formal Definition of Learning

- $H$ is a hypothesis class (e.g., the set of all linear classifiers).
- A learning algorithm receives sample $S$ and selects a hypothesis $h_S$ from $H$ with the goal of approximating $c$.
    - Note: we abstract away details of selection (e.g., linear regression).

# True vs Empirical Risk / Error

- How do we quantify our learning goal?
  - True Risk (unobservable, test error):

$$R(h) = E_{x \sim D}[\mathbf{1}\{h(x) \neq c(x)\}]$$

  - Empirical Risk (observable, training error):

$$\hat{R}_S(h) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{h(x_j) \neq c(x_j)\}$$

# PAC-Learning

- PAC learning: Probably approximately correct learning.
- Concept class C is PAC-learnable if there exists a learning algorithm such that, for all $c \in C$, $\epsilon > 0$, $\delta > 0$, and all distributions D,

$$\Pr(R(h_S) \leq \epsilon \,|\, S \sim D) \geq 1 - \delta$$

where $S$ has size $m$ which is a polynomial function of $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

**In words:** with probability $1 - \delta$, true error is less than $\epsilon$ with a polynomial sized training set.

# Sample Complexity Analysis: Consistent Case

- Goal: want to bound how poor a trained classifier could be after receiving $m$ samples.
- Theorem:
  - Let $H$ be a **finite** class of functions from $X$ to $\{0,1\}$.
  - Let $L$ be an algorithm that returns a consistent hypothesis, i.e., $\hat{R}_S(h_S) = 0$.
  - Then for any $\delta > 0$, we have with probability $1 - \delta$,

$$R(h_S) \leq \frac{1}{m}(\log|H| + \log\frac{1}{\delta})$$

# Sample Complexity Proof: Consistent Case

For any $\epsilon > 0$, define $H_\epsilon = \{h \in H \,|\, R(h) > \epsilon\}$. We want to proof with probability $1-\delta$ that a **consistent** $h_S$ will have low true error.

$$\mathbb{P}\Big[\widehat{R}_S(h_S) = 0 \Rightarrow R(h_S) \le \epsilon\Big] \ge 1-\delta \Leftrightarrow \mathbb{P}\Big[\widehat{R}_S(h_S) = 0 \wedge R(h_S) > \epsilon\Big] \le \delta$$

$$\Leftrightarrow \mathbb{P}\Big[\widehat{R}_S(h_S) = 0 \wedge h_S \in H_\epsilon\Big] \le \delta.$$

$$\mathbb{P}\Big[\exists h \in H \colon \widehat{R}_S(h) = 0 \wedge h \in H_\epsilon\Big]$$

$$= \mathbb{P}\Big[\widehat{R}_S(h_1) = 0 \vee \ldots \vee \widehat{R}_S(h_{|H_\epsilon|}) = 0\Big]$$

$$\le \sum_{h \in H_\epsilon} \mathbb{P}\Big[\widehat{R}_S(h) = 0\Big] \qquad \text{(union bound)}$$

$$\le \sum_{h \in H_\epsilon} (1-\epsilon)^m \le |H|(1-\epsilon)^m \le |H| e^{-m\epsilon}.$$

# Sample Complexity Proof: Consistent Case

We want to proof with probability $1 - \delta$ that a **consistent** $h_S$ will have low true error.

Set $\delta$ equal to upper bound from previous slide and solve for $\epsilon$:

$$\delta = |H| e^{-m\epsilon}$$

Obtain:

$$\epsilon = \frac{1}{m}(\log |H| + \log \frac{1}{\delta})$$

# Sample Complexity Analysis: Inconsistent Case

- Goal: still want to bound how poor a trained classifier could be after receiving $m$ samples.

- However, we want to drop the assumption that $\hat{R}(h_S) = 0$ for $h_S$ returned by our algorithm. **Why?**

# Sample Complexity Analysis: Inconsistent Case

- **Theorem**: let $H$ be a finite hypothesis set, then, for any $\delta > 0$, with probability at least $1-\delta$,

$$\forall h \in H,\ R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}.$$

- **Proof**: By the union bound,

$$\begin{aligned}
&\Pr\left[\max_{h \in H} \left|R(h) - \widehat{R}_S(h)\right| > \epsilon\right] \\
&= \Pr\left[\left|R(h_1) - \widehat{R}_S(h_1)\right| > \epsilon \vee \ldots \vee \left|R(h_{|H|}) - \widehat{R}_S(h_{|H|})\right| > \epsilon\right] \\
&\leq \sum_{h \in H} \Pr\left[\left|R(h) - \widehat{R}_S(h)\right| > \epsilon\right] \\
&\leq 2|H| \exp(-2m\epsilon^2).
\end{aligned}$$

(Hoeffding's Inequality)

From Mehryar Mohri lecture slides

# Outline

- **Learning Theory Motivation**
  - Questions to answer
- **PAC-learning**
  - Mistake bounds.
- **VC Dimension**
  - Definition, why useful

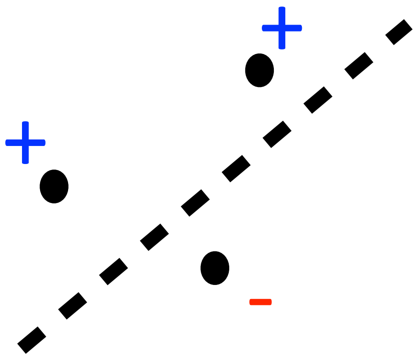
# VC-Dimension

- Formal measure of capacity for a function class.
  - i.e., flexibility, representational power, complexity

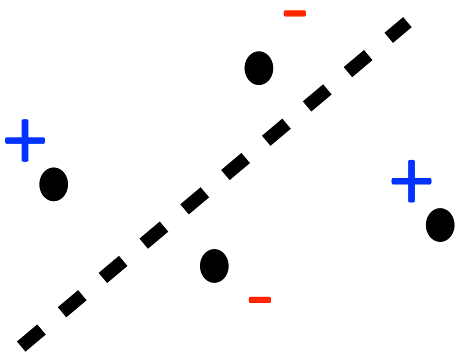
- A function class shatters a set of points if for all labeling of the points there is a function in the class that perfectly classifies the points.

- VC dimension of a function class is the size of the largest set of points that can be shattered by that class.

# VC-Dimension Example




Linear classifiers (in $R^2$) can shatter sets of three points




Linear classifiers (in $R^2$) cannot shatter sets of four points

In general, VC dimension of linear classifiers in $R^d$ is $d + 1$.

Only need one arrangement of points but must consider all possible labelings.

# Why VC-Dimension is Useful?

- Useful for characterizing infinite hypothesis classes.
  - Sample complexity bounds can depend on VC-dimension instead of size of hypothesis classes.

- Example Hardness Result (lower bound):

  - **Theorem**: let $H$ be a hypothesis set with VC-dimension $d > 1$. Then, for any learning algorithm $L$,

$$\exists D, \exists f \in H, \; \Pr_{S \sim D^m}\left[R_D(h_S, f) > \frac{d-1}{32m}\right] \geq 1/100.$$

See given reading for proof.

# Summary

- Learning theory enables rigorous understanding of machine learning problems and algorithms.

- (Some) key questions learning theory attempts to answer:
  - How hard is a problem?
  - Can we upper bound error for a given sample size?
  - Is a problem efficiently learnable?
    - In terms of space, time, and training set size.

# Thanks Everyone!

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mehryar Mohair