



10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

PAC Learning

Learning Theory Readings:

Murphy --
Bishop --
HTF --
Mitchell 7

Matt Gormley
Lecture 28
May 1, 2016

Reminders

- **Homework 9: Applications of ML**
 - **Release: Mon, Apr. 24**
 - **Due: Wed, May 3 at 11:59pm**

Outline

- **Statistical Learning Theory**
 - True Error vs. Train Error
 - Function Approximation View (aka. PAC/SLT Model)
 - Three Hypotheses of Interest
- **Probably Approximately Correct (PAC) Learning**
 - PAC Criterion
 - PAC Learnable
 - Consistent Learner
 - Sample Complexity
- **Generalization and Overfitting**
 - Realizable vs. Agnostic Cases
 - Finite vs. Infinite Hypothesis Spaces
 - VC Dimension
 - Sample Complexity Bounds
 - Empirical Risk Minimization
 - Structural Risk Minimization
- **Excess Risk**

LEARNING THEORY

Questions For Today

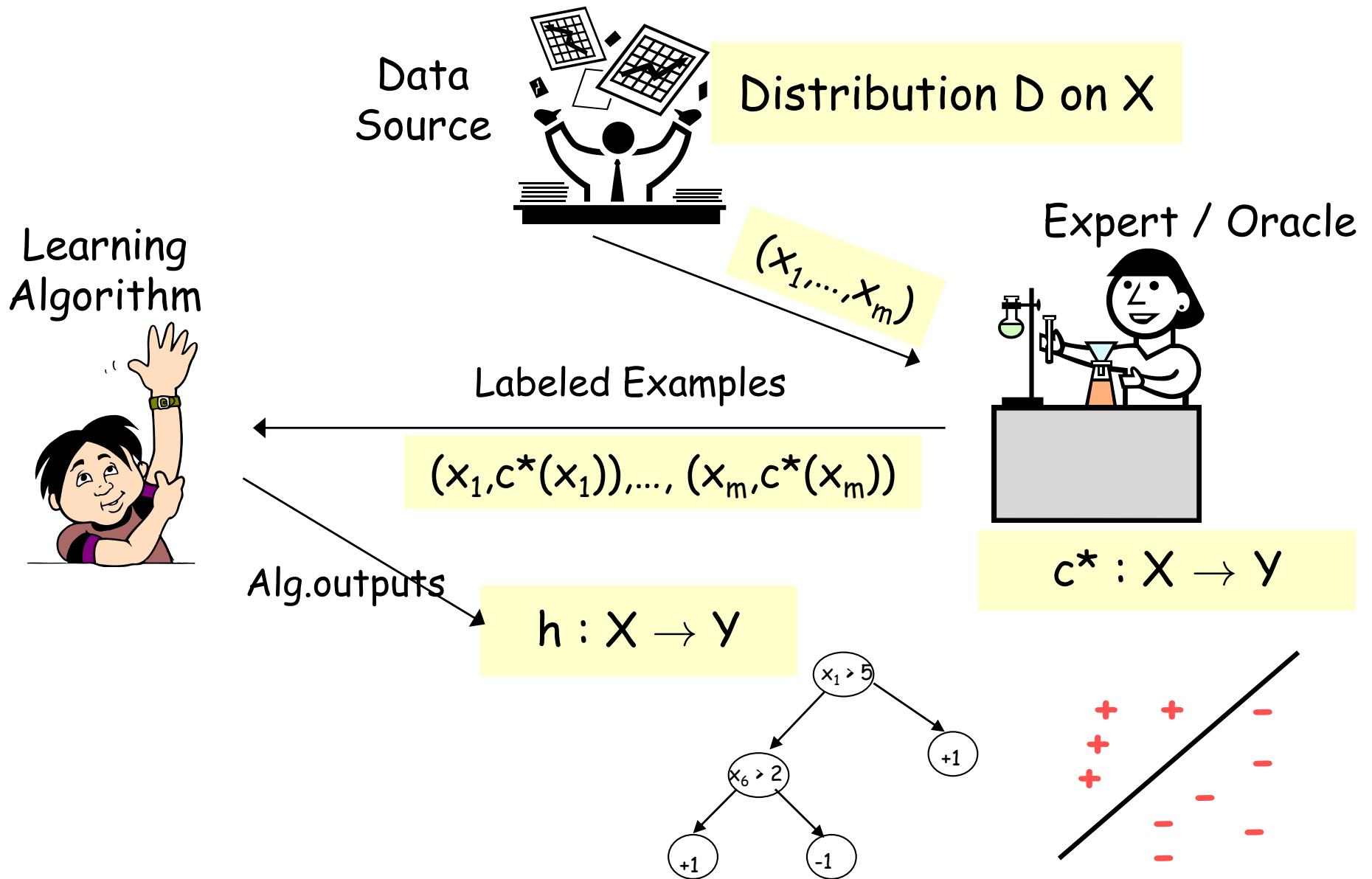
1. Given a classifier with zero training error, what can we say about generalization error?
(Sample Complexity, Realizable Case)
2. Given a classifier with low training error, what can we say about generalization error?
(Sample Complexity, Agnostic Case)
3. Is there a theoretical justification for regularization to avoid overfitting?
(Structural Risk Minimization)

Statistical Learning Theory

Whiteboard:

- Function Approximation View (aka. PAC/SLT Model)
- True Error vs. Train Error
- Three Hypotheses of Interest

PAC/SLT models for Supervised Learning



PAC / SLT Model

We've also referred to this as the "Function Approximation View"

1. Generate instances from *unknown* distribution p^*

$$\mathbf{x}^{(i)} \sim p^*(\mathbf{x}), \forall i \tag{1}$$

2. Oracle labels each instance with *unknown* function c^*

$$y^{(i)} = c^*(\mathbf{x}^{(i)}), \forall i \tag{2}$$

3. Learning algorithm chooses hypothesis $h \in \mathcal{H}$ with low(est) training error, $\hat{R}(h)$

$$\hat{h} = \underset{h}{\operatorname{argmin}} \hat{R}(h) \tag{3}$$

4. Goal: Choose an h with low generalization error $R(h)$

Two Types of Error

True Error (aka. **expected risk**)

$$R(h) = P_{\mathbf{x} \sim p^*(\mathbf{x})}(c^*(\mathbf{x}) \neq h(\mathbf{x}))$$

Train Error (aka. **empirical risk**)

$$\begin{aligned}\hat{R}(h) &= P_{\mathbf{x} \sim \mathcal{S}}(c^*(\mathbf{x}) \neq h(\mathbf{x})) \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{1}(c^*(\mathbf{x}^{(i)}) \neq h(\mathbf{x}^{(i)})) \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{1}(y^{(i)} \neq h(\mathbf{x}^{(i)}))\end{aligned}$$

where $\mathcal{S} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}_{i=1}^N$ is the training data set, and $\mathbf{x} \sim \mathcal{S}$ denotes that \mathbf{x} is sampled from the empirical distribution.

This quantity
is always
unknown

We can
measure this
on the training
data

Three Hypotheses of Interest

The **true function** c^* is the one we are trying to learn and that labeled the training data:

$$y^{(i)} = c^*(\mathbf{x}^{(i)}), \forall i \quad (1)$$

The **expected risk minimizer** has lowest true error:

$$h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h) \quad (2)$$

The **empirical risk minimizer** has lowest training error:

$$\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h) \quad (3)$$

PAC LEARNING

Probably Approximately Correct (PAC) Learning

Whiteboard:

- PAC Criterion
- Meaning of “Probably Approximately Correct”
- PAC Learnable
- Consistent Learner
- Sample Complexity

PAC Learning

The **PAC criterion** is that our learner produces a high accuracy learner with high probability:

$$P(|R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta \quad (1)$$

Suppose we have a learner that produces a hypothesis $h \in \mathcal{H}$ given a sample of N training examples. The algorithm is called **consistent** if for every ϵ and δ , there exists a positive number of training examples N such that for any distribution p^* , we have that:

$$P(|R(h) - \hat{R}(h)| > \epsilon) < \delta \quad (2)$$

The **sample complexity** is the minimum value of N for which this statement holds. If N is finite for some learning algorithm, then \mathcal{H} is said to be **learnable**. If N is a polynomial function of $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ for some learning algorithm, then \mathcal{H} is said to be **PAC learnable**.

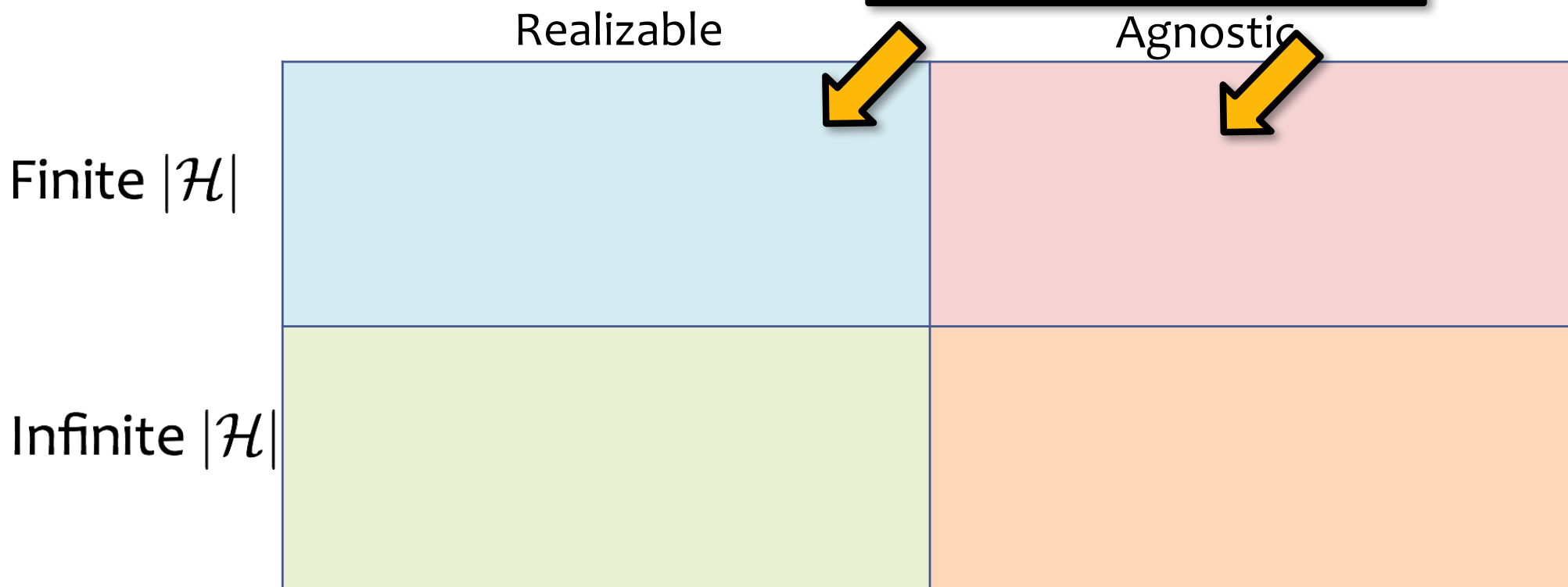
SAMPLE COMPLEXITY RESULTS

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

We'll start with the finite case...



Generalization and Overfitting

Whiteboard:

- Realizable vs. Agnostic Cases
- Finite vs. Infinite Hypothesis Spaces
- Sample Complexity Bounds (Finite Case)

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	
Infinite $ \mathcal{H} $		

Example: Conjunctions

In-Class Quiz:

Suppose H = class of conjunctions over x in $\{0,1\}^M$

If $M = 10$, $\epsilon = 0.1$, $\delta = 0.01$, how many examples suffice?

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	
Infinite $ \mathcal{H} $		

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).



Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) < \epsilon$.
Infinite $ \mathcal{H} $		

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) > 0$.	$N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$, $ R(h) - \hat{R}(h) \leq \epsilon$.
Infinite $ \mathcal{H} $		

We need a new definition of "complexity" for a Hypothesis space for these results (see VC Dimension)

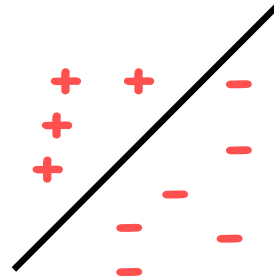
VC DIMENSION



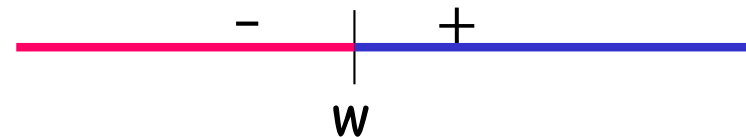
What if H is infinite?



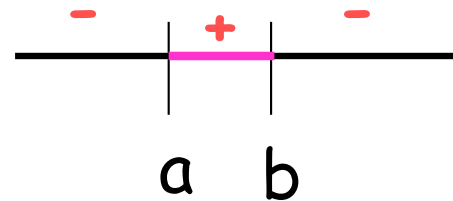
E.g., linear separators in \mathbb{R}^d



E.g., thresholds on the real line



E.g., intervals on the real line



Shattering, VC-dimension

Definition:

$H[S]$ - the set of splittings of dataset S using concepts from H .

H shatters S if $|H[S]| = 2^{|S|}$.

A set of points S is shattered by H if there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in S are achievable using concepts in H .

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The **VC-dimension** of a hypothesis space H is the cardinality of the largest set S that can be shattered by H .

If arbitrarily large finite sets can be shattered by H , then

$$\text{VCdim}(H) = \infty$$

Shattering, VC-dimension

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The **VC-dimension** of a hypothesis space H is the cardinality of the largest set S that can be shattered by H .

If arbitrarily large finite sets can be shattered by H , then $\text{VCdim}(H) = \infty$

To show that VC-dimension is d :

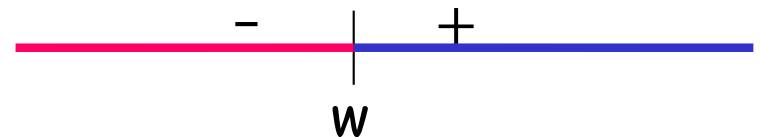
- **there exists** a set of **d points** that can be shattered
- there is **no set of $d+1$ points** that can be shattered.

Fact: If H is finite, then $\text{VCdim}(H) \leq \log(|H|)$.

Shattering, VC-dimension

If the VC-dimension is d , that means **there exists** a set of d points that can be shattered, but there is **no** set of $d+1$ points that can be shattered.

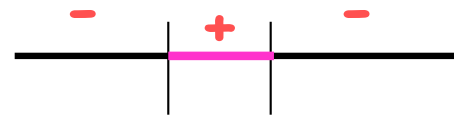
E.g., $H =$ Thresholds on the real line



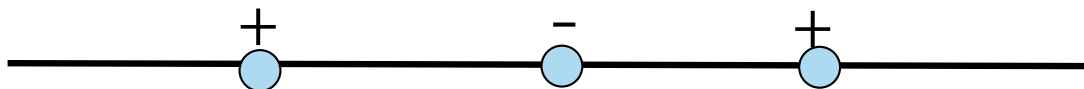
$$\text{VCdim}(H) = 1$$



E.g., $H =$ Intervals on the real line



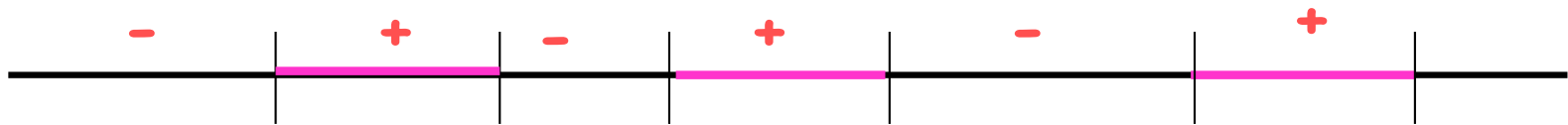
$$\text{VCdim}(H) = 2$$



Shattering, VC-dimension

If the VC-dimension is d , that means **there exists** a set of d points that can be shattered, but there is **no** set of $d+1$ points that can be shattered.

E.g., $H = \text{Union of } k \text{ intervals on the real line}$ $\text{VCdim}(H) = 2k$



$$\text{VCdim}(H) \geq 2k$$

A sample of size $2k$ shatters
(treat each pair of points as a
separate case of intervals)

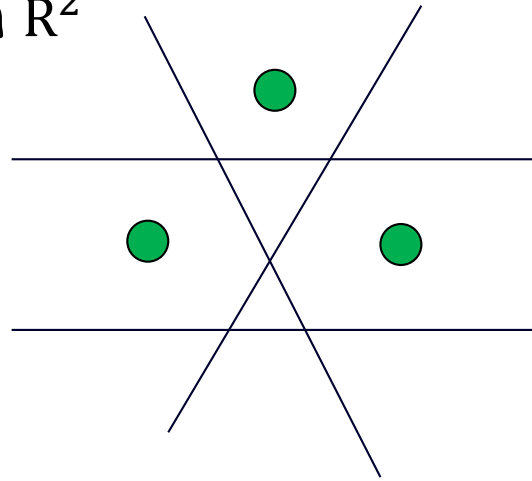
$$\text{VCdim}(H) < 2k + 1$$



Shattering, VC-dimension

E.g., H = linear separators in \mathbb{R}^2

$\text{VCdim}(H) \geq 3$

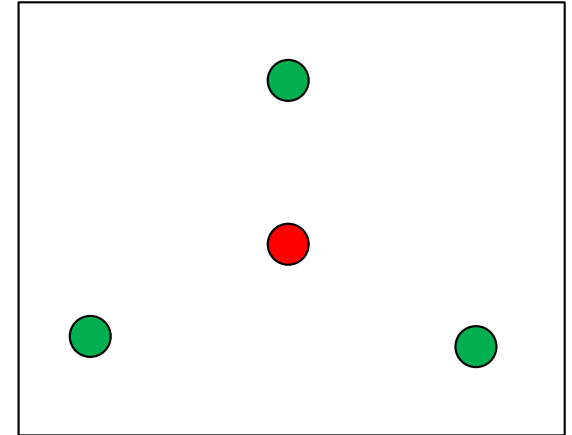


Shattering, VC-dimension

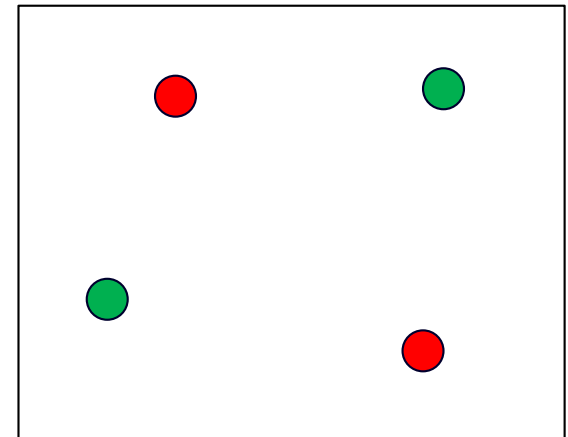
E.g., H = linear separators in \mathbb{R}^2

$\text{VCdim}(H) < 4$

Case 1: one point inside the triangle formed by the others. Cannot label inside point as positive and outside points as negative.



Case 2: all points on the boundary (convex hull). Cannot label two diagonally as positive and other two as negative.





Fact: VCdim of linear separators in \mathbb{R}^d is $d+1$

SAMPLE COMPLEXITY RESULTS

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) > 0$.	$N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have $ \hat{R}(h) - R(h) \leq \epsilon$.
Infinite $ \mathcal{H} $		

We need a new definition of "complexity" for a Hypothesis space for these results (see VC Dimension)

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) < \epsilon$.
Infinite $ \mathcal{H} $	$N = O(\frac{1}{\epsilon} [\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N = O(\frac{1}{\epsilon^2} [\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.

Generalization and Overfitting

Whiteboard:

- Sample Complexity Bounds (Infinite Case)
- Empirical Risk Minimization
- Structural Risk Minimization

EXCESS RISK

Excess Risk

There are two common quantities to consider based on the:

empirical risk minimizer $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h)$ and

expected risk minimizer $h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$.

1. We can bound the difference between the expected risk and empirical risk $R(\hat{h}) - \hat{R}(\hat{h})$. Note that both of these quantities are functions of the ERM hypothesis \hat{h} .
2. The **excess risk** $R(\hat{h}) - R(h^*)$ is the difference in *true* error between the ERM hypothesis \hat{h} and the expected risk minimizer h^* .

We aim to prove that $P(R(\hat{h}) - R(h^*) \leq \epsilon) \geq (1 - \delta)$ or equivalently that $P(R(\hat{h}) - R(h^*) > \epsilon) < \delta$.

Excess Risk Results

Bounds on the excess risk $R(\hat{h}) - R(h^*)$:

- realizable case, finite $|\mathcal{H}|$: $O\left(\frac{\log(|\mathcal{H}|)}{N}\right)$
- agnostic case, finite $|\mathcal{H}|$: $O\left(\sqrt{\frac{\log(|\mathcal{H}|)}{N}}\right)$
- infinite $|\mathcal{H}|$: $O\left(\sqrt{\frac{\text{VC}(\mathcal{H}) \log(N)}{N}}\right)$

Questions For Today

1. Given a classifier with zero training error, what can we say about generalization error?
(Sample Complexity, Realizable Case)
2. Given a classifier with low training error, what can we say about generalization error?
(Sample Complexity, Agnostic Case)
3. Is there a theoretical justification for regularization to avoid overfitting?
(Structural Risk Minimization)