10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

PAC Learning

Matt Gormley
Lecture 25
Apr. 28, 2021

Reminders

- Homework 7: Graphical Models
  - Out: Mon, Apr. 19
  - Due: Fri, Apr. 30 at 11:59pm
- Homework 8: Learning Paradigms
  - Out: Fri, Apr. 30
  - Due: Fri, May. 7 at 11:59pm

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

1. Bound is **inversely linear in epsilon** (e.g. halving the error requires double the examples)
2. Bound is **only logarithmic in |H|** (e.g. quadrupling the hypothesis space only requires double the examples)

1. Bound is **inversely quadratic in epsilon** (e.g. halving the error requires 4x the examples)
2. Bound is **only logarithmic in |H|** (i.e. same as Realizable case)

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

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 $\|\mathcal{H}\|$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\|\mathcal{H}\|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that or all $h \in \mathcal{H}$ we $\leq \epsilon$.
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).

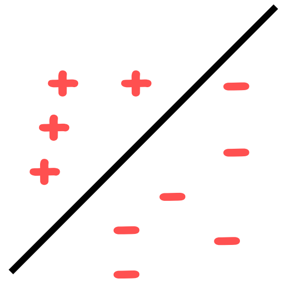
Four Cases we care about…

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\|\mathcal{H}\|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.
Infinite $\|\mathcal{H}\|$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.

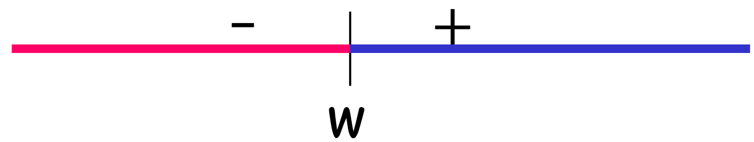
VC DIMENSION

What if H is infinite?

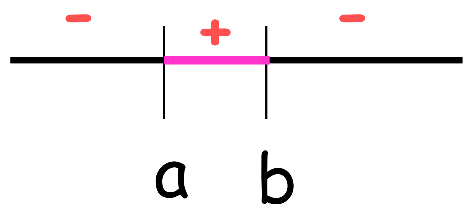
E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



Slide from Nina Balcan

Shattering, VC-dimension

Definition:

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

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

A set of points S is shattered by H is 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.

Slide from Nina Balcan

VC Dimension

Whiteboard:

- Shattering example: binary classification

Shattering, VC-dimension

Definition:

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

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

A set of points S is shattered by H is 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 $\mathrm{VCdim(H)} = \infty$

Slide from Nina Balcan

VC Dimension

Whiteboard:

- VC Dimension Example: linear separators
- Proof sketch of VCDim for linear separators in 2D

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

To show that VC-dimension is d:

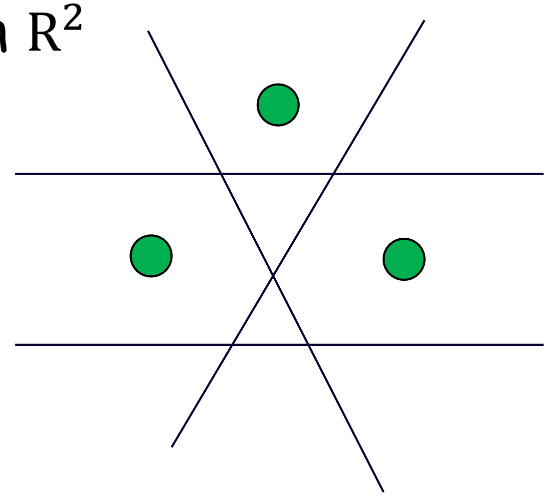
- there exists a set of d points that can be shattered
- there is no set of d+1 points that can be shattered.

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

Slide from Nina Balcan

Shattering, VC-dimension

E.g., **H=** linear separators in R^2



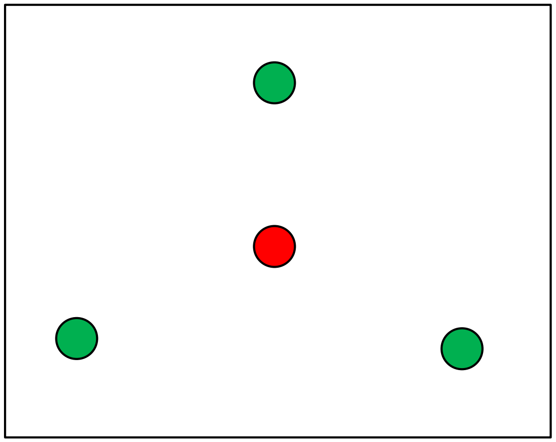
$\mathrm{VCdim(H)} \geq 3$

Shattering, VC-dimension

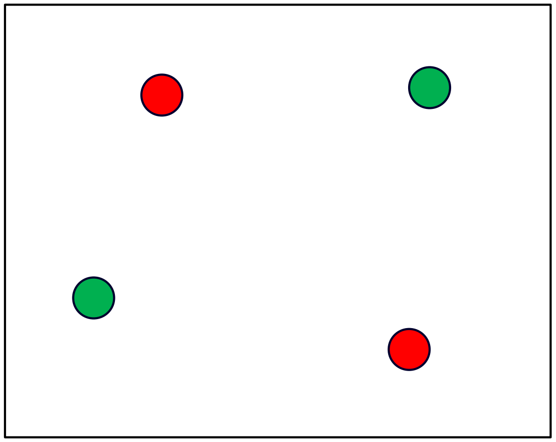
E.g., **H=** linear separators in R^2

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

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 R^d is d+1

∃ vs. ∀

VCDim

– Proving **VC Dimension** requires us to show that **there exists (∃)** a dataset of size d that can be shattered and that **there does not exist (∄)** a dataset of size d+1 that can be shattered

Shattering

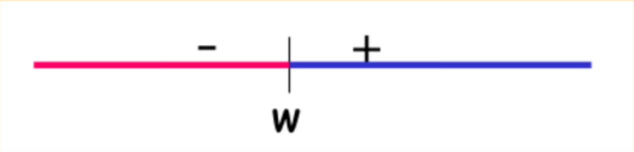
– Proving that a particular dataset can be **shattered** requires us to show that **for all (∀)** labelings of the dataset, our hypothesis class contains a hypothesis that can correctly classify it

VC-Dimension Examples

- *Definition*: If VC(H) = d, then **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Question:

What is the VC Dimension of H = **thresholds on the real line.** That is for a threshold w, everything to the right of w is labeled as +1, everything else is labeled -1.



Answer:

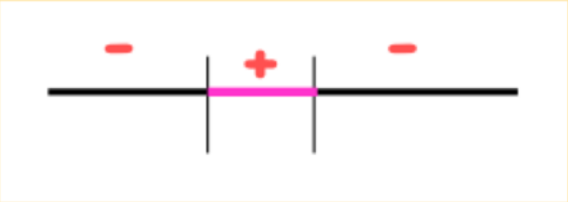
Figures from Nina Balcan

VC-Dimension Examples

- *Definition*: If VC(H) = d, then **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Question:

What is the VC Dimension of H = **intervals on the real line.** That is for an interval (w_1, w_2), everything inside the interval is labeled as +1, everything else is labeled -1.



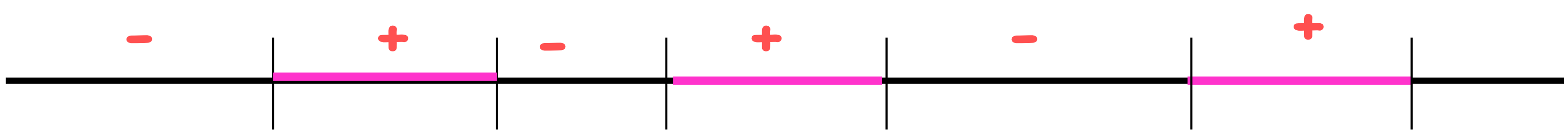
Answer:

Figures from Nina Balcan

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= Union of k intervals on the real line $\mathrm{VCdim}(\mathrm{H}) = 2\mathrm{k}$



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

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

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

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 $\|\mathcal{H}\|$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\|\mathcal{H}\|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.
Infinite $\|\mathcal{H}\|$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.

SLT-style Corollaries

Thm. 1 $N \geq \frac{1}{\epsilon}\left[\log(|\mathcal{H}|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

Solve the inequality in Thm.1 for epsilon to obtain Corollary 1

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

We can obtain similar corollaries for each of the theorems…

SLT-style Corollaries

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

Corollary 2 (Agnostic, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right]}$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O\left(\frac{1}{N}\left[\text{VC}(\mathcal{H}) \ln\left(\frac{N}{\text{VC}(\mathcal{H})}\right) + \ln\left(\frac{1}{\delta}\right)\right]\right) \tag{1}$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\text{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \tag{2}$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O\left(\frac{1}{N}\left[\text{VC}(\mathcal{H}) \ln\left(\frac{N}{\text{VC}(\mathcal{H})}\right) + \ln\left(\frac{1}{\delta}\right)\right]\right) \qquad (1)$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\text{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \qquad (2)$$

Should these corollaries inform how we do model selection?

Generalization and Overfitting

Whiteboard:

- Model Selection
- Empirical Risk Minimization
- Structural Risk Minimization
- Motivation for Regularization

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)

Learning Theory Objectives

You should be able to…

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world learning examples
- Distinguish between a large sample and a finite sample analysis
- Theoretically motivate regularization