Announcements

Assignments

- HW5
  - Due Mon, 10/26, 11:59 pm
- HW6
  - Out tomorrow
  - Due Mon, 11/2, 11:59 pm

Midterm 2

- Mon, 11/9, during lecture

Fireside Chat about the CMU ML PhD Program

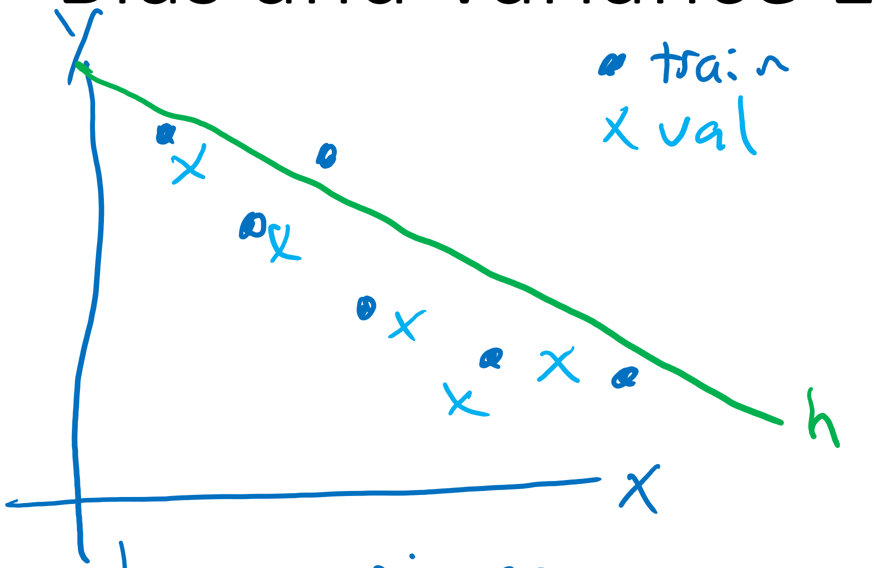
- Fri, 10/30, 8:00 pm
- See Piazza for details, including form to show interest

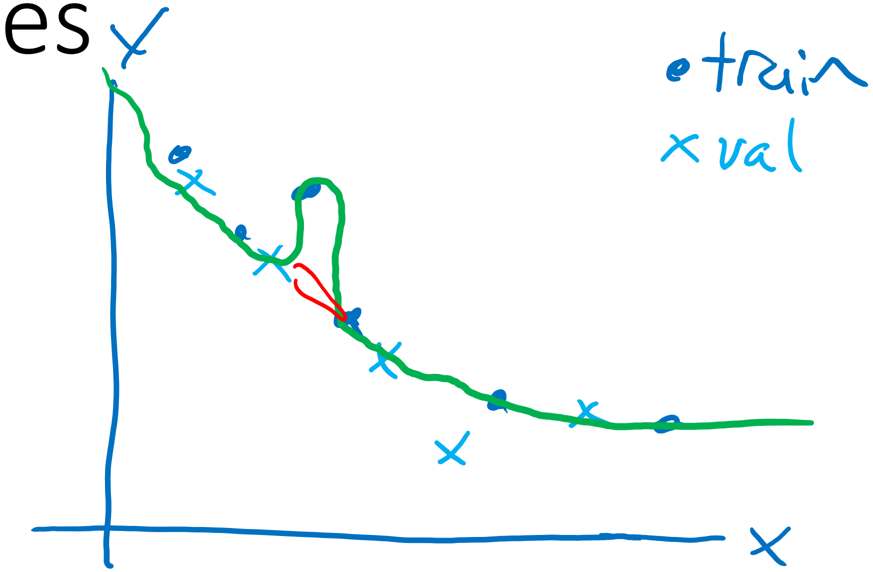
Introduction to Machine Learning

Learning Theory

Instructor: Pat Virtue

Bias and Variance Examples



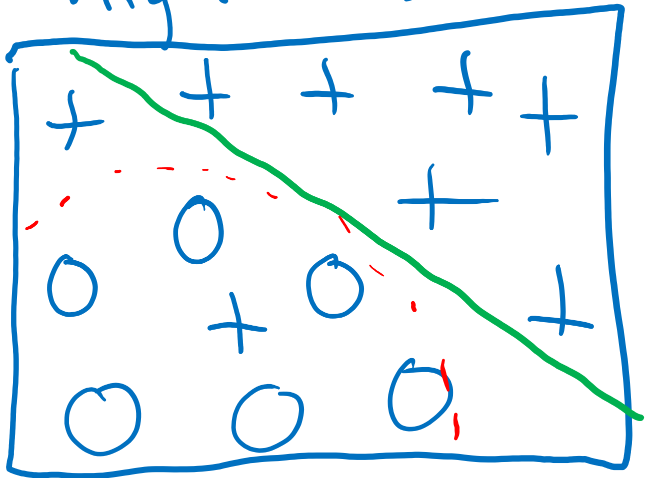


low variance

high bias

med variance

low bias

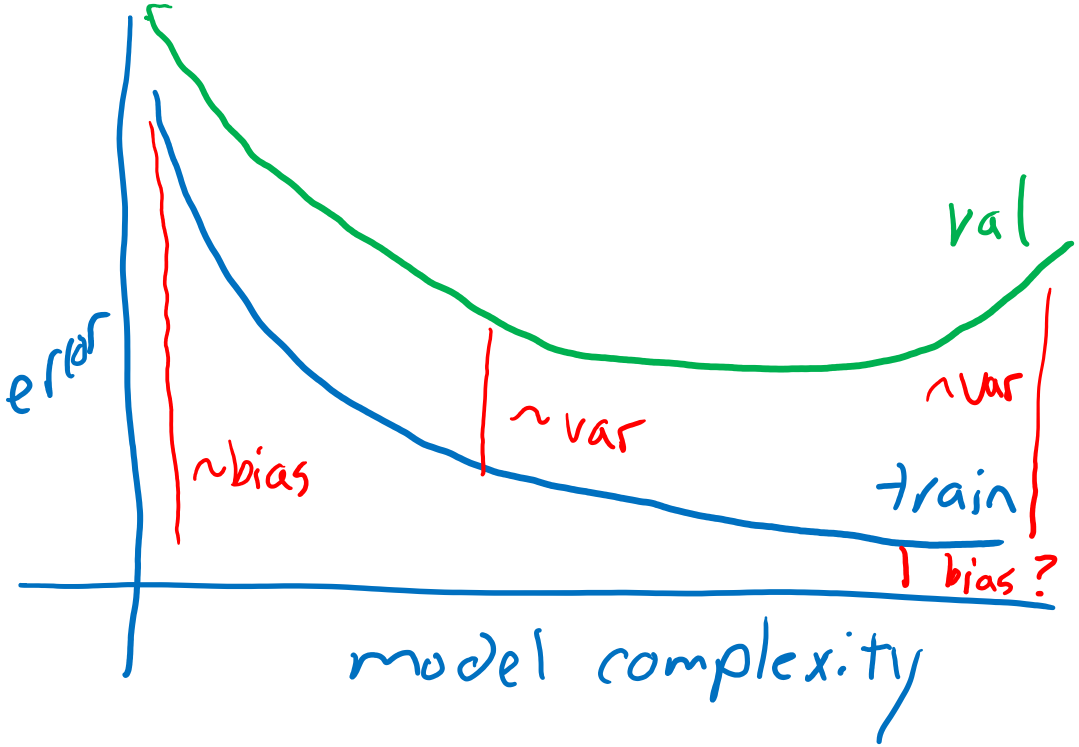


Slide credit: Andrew Ng, Stanford

Bias and Variance Examples

Assume that $h^* \leq 0.1$

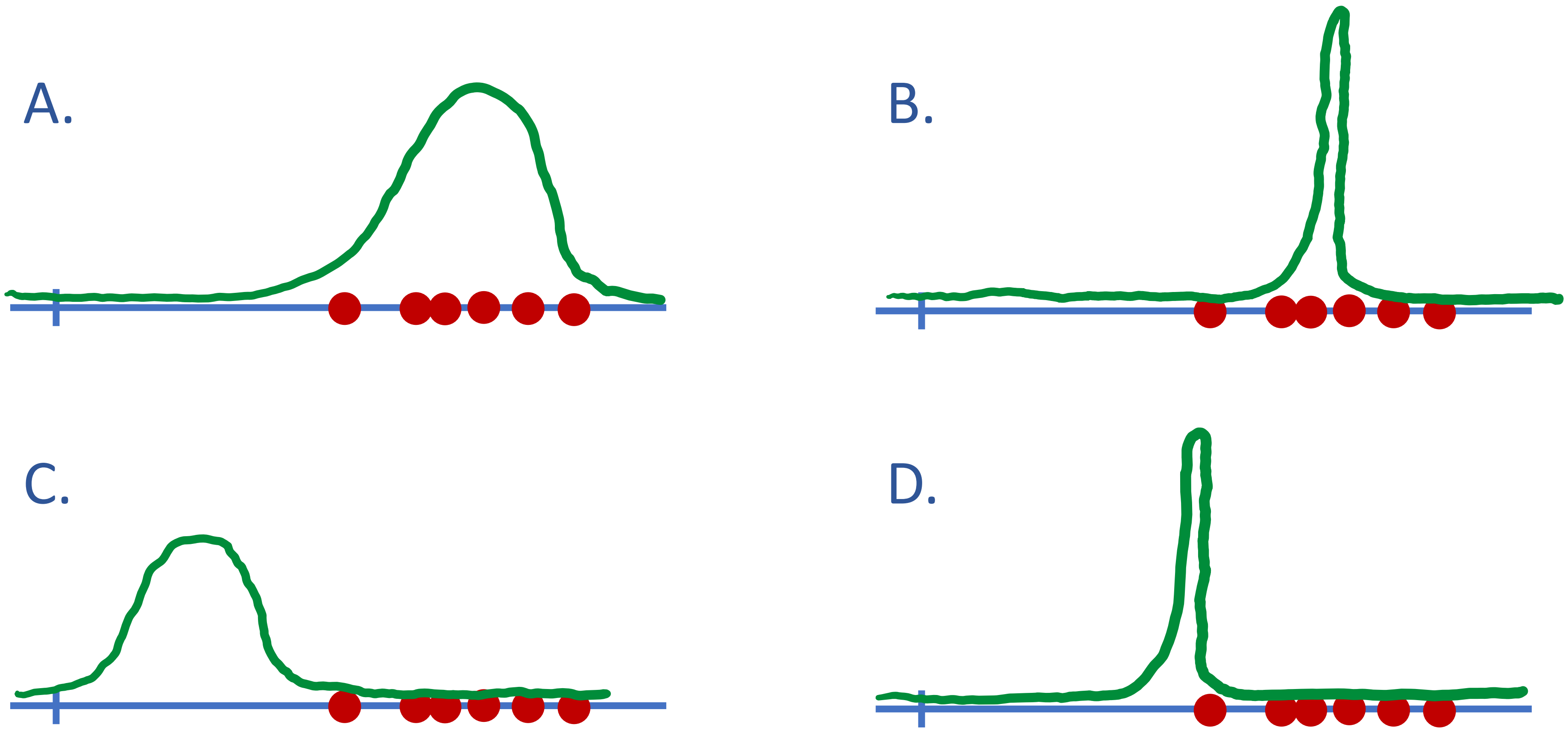
	h_A	h_B
train err	0.25	0.1
val err	0.3	0.3
	high bias	low bias?
	low var	high var



Previous Piazza Polls

Poll 1: [SELECT TWO] Which have high variance?

Poll 2: [SELECT TWO] Which have high bias?



Questions

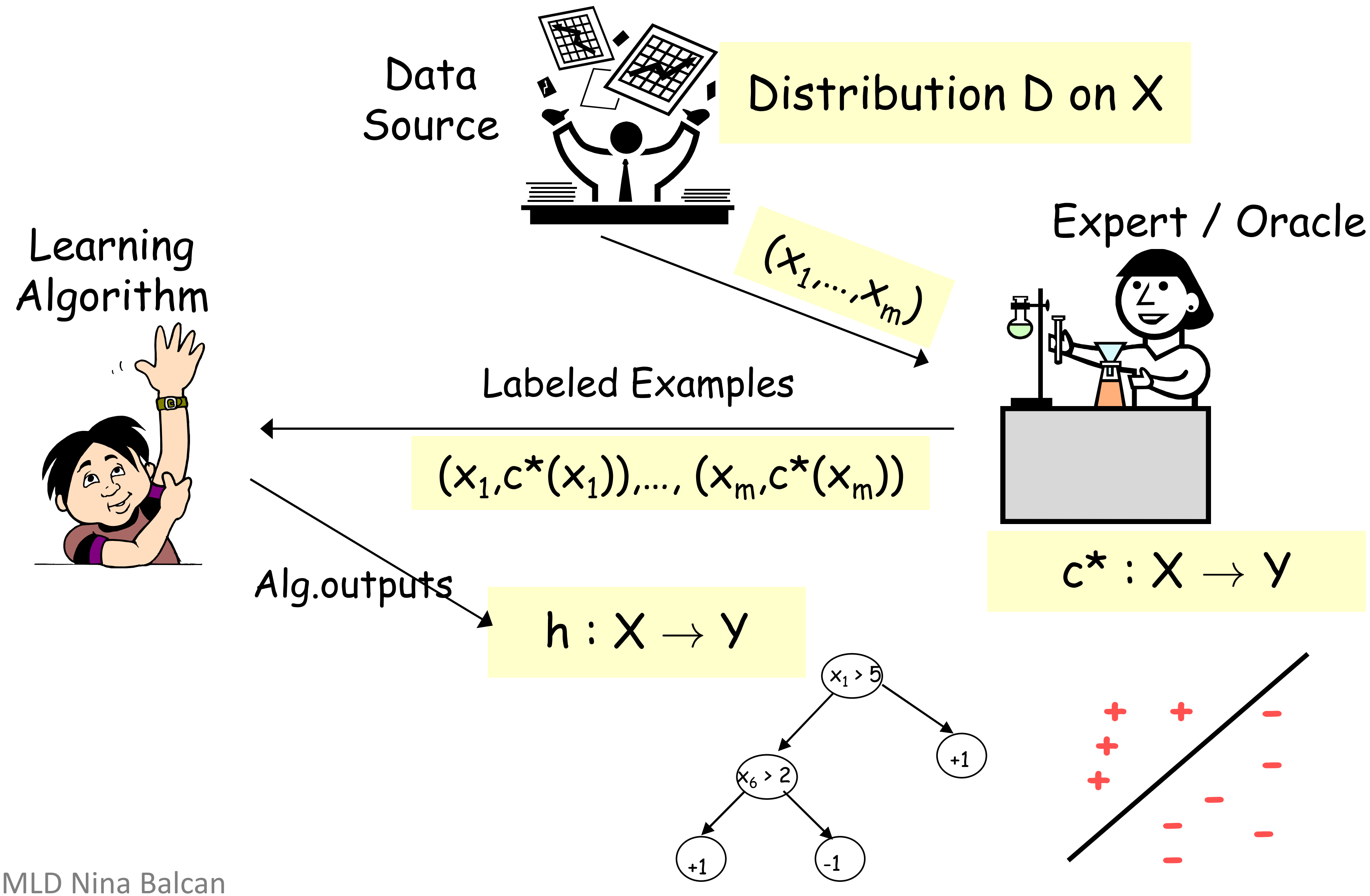
1. Given a classifier with **zero training error**, what can we say about **true error** (aka. generalization error)?
   (Sample Complexity, Realizable Case)

2. Given a classifier with **low training error**, what can we say about **true error** (aka. generalization error)?
   (Sample Complexity, Agnostic Case)

3. Is there a **theoretical justification for regularization** to avoid overfitting?
   (Structural Risk Minimization)

Slide credit: CMU MLD Matt Gormley

Model for Supervised Learning



Slide credit: CMU MLD Nina Balcan

Risk, 0-1 Loss, and Error Rate

Risk is the expected loss over data points

0-1 loss means we have a cost of one when classify a point wrong

Risk for 0-1 loss is simply error rate

Two Types of Error

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

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

This quantity is always **unknown**

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

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

We can **measure** this on the training data

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

Slide credit: CMU MLD Matt Gormley

PAC / SLT Model

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, $\hat{R}(h)$

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

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

Slide credit: CMU MLD Matt Gormley

Three Hypotheses of Interest

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

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

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

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

Question:
True or False:
h* and c* are always equal.

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

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

Slide credit: CMU MLD Matt Gormley

Piazza Poll 1

True or False: h* and c* are always equal.

PAC Learning

Can we bound $R(h)$ in terms of $\hat{R}(h)$?

Definition: PAC Criterion:

PAC Learning

PAC Criterion

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

Definition: sample complexity

Definition: consistent hypothesis

Slide credit: CMU MLD Matt Gormley

PAC Learning

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

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

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:

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

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.**

Slide credit: CMU MLD Matt Gormley

PAC Learning

Four types of problems

Slide credit: CMU MLD Matt Gormley

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...

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$		
Infinite $\|\mathcal{H}\|$		

PAC Learning

PAC Criterion

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

Theorem 1: Sample Complexity (Realizable, Finite $|\mathcal{H}|$)

PAC Learning

Proof of Theorem 1

See PAC Learning: Theorem 1 notes and video

Slide credit: CMU MLD Matt Gormley

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 $\lvert\mathcal{H}\rvert$	**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 $R(h) \leq \epsilon$.	
Infinite $\lvert\mathcal{H}\rvert$		

Slide credit: CMU MLD Matt Gormley

Piazza Poll 2

Question:

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

Example hypotheses:

$h(\mathbf{x}) = x_1\,(1-x_3)\,x_5$

$h(\mathbf{x}) = x_1\,(1-x_2)\,x_4\,(1-x_5)$

If $M = 10$, $\varepsilon = 0.1$, $\delta = 0.01$, how many examples suffice according to Theorem 1?

Answer:

A. $10*(2*\ln(10)+\ln(100)) \approx 92$
B. $10*(3*\ln(10)+\ln(100)) \approx 116$
C. $10*(10*\ln(2)+\ln(100)) \approx 116$
D. $10*(10*\ln(3)+\ln(100)) \approx 156$
E. $100*(2*\ln(10)+\ln(10)) \approx 691$
F. $100*(3*\ln(10)+\ln(10)) \approx 922$
G. $100*(10*\ln(2)+\ln(10)) \approx 924$
H. $100*(10*\ln(3)+\ln(10)) \approx 1329$

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 $R(h) \leq \epsilon$.

Piazza Poll 2

Question:

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

Example hypotheses:

$h(\mathbf{x}) = x_1 (1-x_3) x_5$

$h(\mathbf{x}) = x_1 (1-x_2) x_4 (1-x_5)$

If M = 10, ε = 0.1, δ = 0.01, how many examples suffice according to Theorem 1?

Answer:

A. $10*(2*\ln(10)+\ln(100)) \approx 92$
B. $10*(3*\ln(10)+\ln(100)) \approx 116$
C. $10*(10*\ln(2)+\ln(100)) \approx 116$
D. $10*(10*\ln(3)+\ln(100)) \approx 156$
E. $100*(2*\ln(10)+\ln(10)) \approx 691$
F. $100*(3*\ln(10)+\ln(10)) \approx 922$
G. $100*(10*\ln(2)+\ln(10)) \approx 924$
H. $100*(10*\ln(3)+\ln(10)) \approx 1329$

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 $R(h) \leq \epsilon$.

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 $\lvert\mathcal{H}\rvert$	**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 $\lvert\mathcal{H}\rvert$		

Slide credit: CMU MLD Matt Gormley

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 $\lvert\mathcal{H}\rvert$		

Slide credit: CMU MLD Matt Gormley

Using a PAC bound

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

- Given ε and δ, yields sample complexity

#training data, $m \geq \dfrac{\ln|H| + \ln\frac{1}{\delta}}{\epsilon}$

- Given m and δ, yields error bound

error, $\epsilon \geq \dfrac{\ln|H| + \ln\frac{1}{\delta}}{m}$

Summary of PAC bounds for finite model classes

With probability $\geq 1-\delta$,

1) For all $h \in H$ s.t. $\text{error}_{\text{train}}(h) = 0$,

$$\text{error}_{\text{true}}(h) \leq \varepsilon = \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

Haussler's bound

2) For all $h \in H$

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \leq \varepsilon = \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1-\delta$

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- Fixed |H|

Training size		
m small	small	large
m large	large	small

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

PAC bound and Bias-Variance tradeoff

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- Fixed m

Model class	$\downarrow$	$\downarrow$
|H| large (complex)	small	large
|H| small (simple)	large	small

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

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 $\|\mathcal{H}\|$		

Slide credit: CMU MLD Matt Gormley

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…

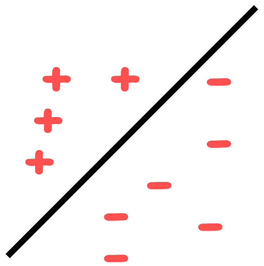
	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq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

Slide credit: CMU MLD Matt Gormley

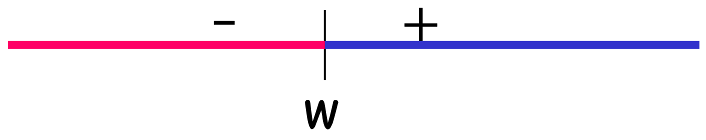
VC DIMENSION

What if H is infinite?

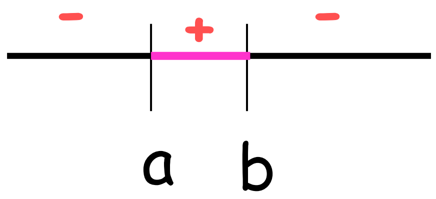
E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



Shattering, VC-dimension

Definition:

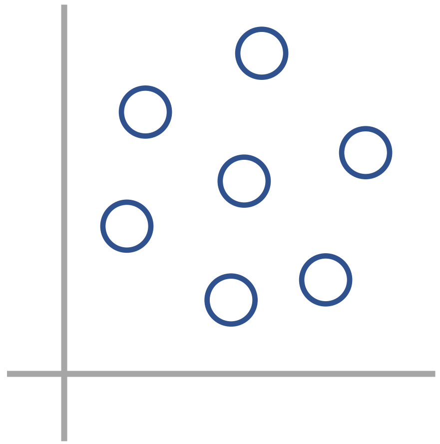
$H[S]$ – the set of splittings of dataset S using concepts from H.

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

A set of points S is shattered by H 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 credit: CMU MLD Nina Balcan

Example: Shattering for Binary Classification

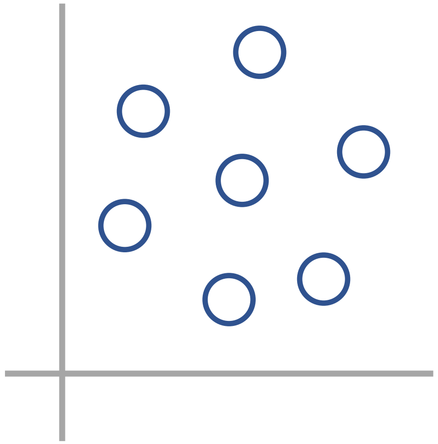


Piazza Poll 3

Does $\mathcal{H}$ shatter $\mathcal{S}$, where $\mathcal{H}$ = set of circular decision boundaries and $\mathcal{S}$ = set of 2D points?

i.e. Does the number of splittings, $|\mathcal{H}[\mathcal{S}]|$, equal $2^{|\mathcal{S}|}$?

i.e. Can a circular decision boundary perfectly separate any labelling of $\mathcal{S}$?



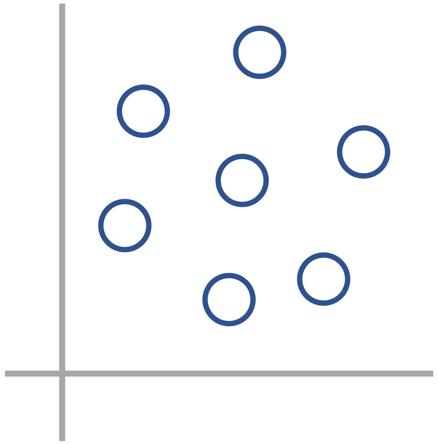
Piazza Poll 3

Does $\mathcal{H}$ shatter $\mathcal{S}$, where $\mathcal{H}$ = set of circular decision boundaries and $\mathcal{S}$ = set of 2D points?

i.e. Does the number of splittings, $|\mathcal{H}[\mathcal{S}]|$, equal $2^{|\mathcal{S}|}$?

i.e. Can a circular decision boundary perfectly separate any labelling of $\mathcal{S}$?

No.



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 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.

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 credit: CMU MLD Nina Balcan

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

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 $\mathrm{VCdim}(H) \leq \log(|H|)$.

Slide credit: CMU MLD Nina Balcan

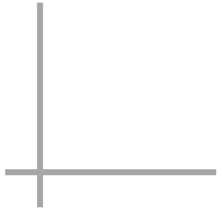
Example: VC Dimension for Linear Separators

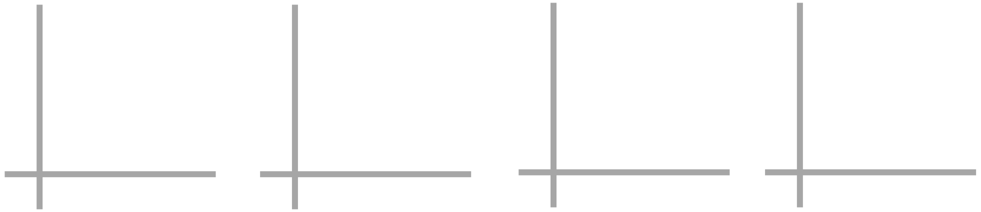
Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

1. $\exists\ \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $!\exists\ \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$

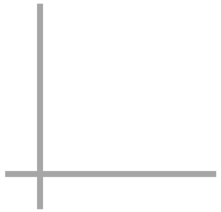
1. Pick one (unlabeled) dataset for d | List all possible labelings of $\mathcal{S}$ | Show that we can shatter $\mathcal{S}$

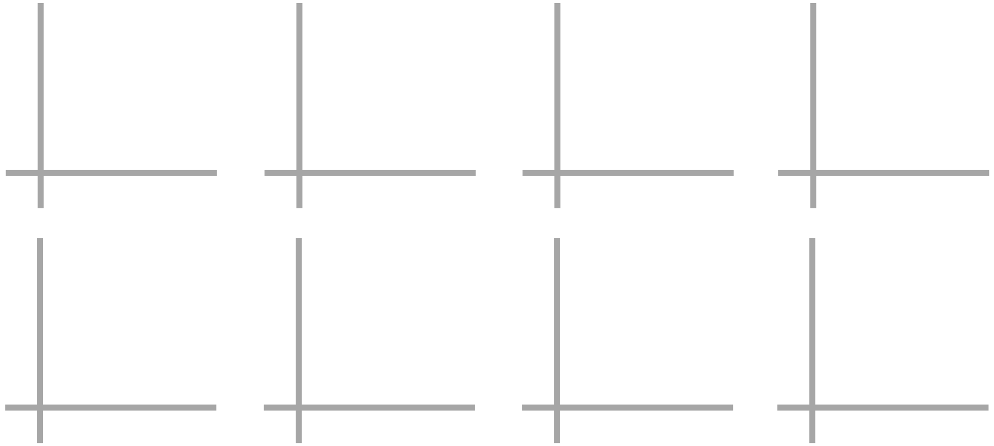
d=2





d=3





Slide credit: CMU MLD Matt Gormley

Example: VC Dimension for Linear Separators

Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

1. $\exists\ \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $!\exists\ \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $\forall\ \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ $\mathcal{H}$ cannot shatter $\mathcal{S}$

Slide credit: CMU MLD Matt Gormley

Example: VC Dimension for Linear Separators

Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

1. $\exists\, \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $!\exists\, \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$

But…

Isn't there a dataset of size d=3 that can't be shattered?

∃ 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

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 $\lvert\mathcal{H}\rvert$	**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 $\lvert\mathcal{H}\rvert$	**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 $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

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…

Slide credit: CMU MLD Matt Gormley

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 $\mathcal{H}$,

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

Slide credit: CMU MLD Matt Gormley

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 $\hat{R}(h) = 0$),

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

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

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

Should these corollaries inform how we do model selection?

PAC Bounds and Model Selection

Is Corollary 4 useful?

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

PAC Bounds and Regularization

Example: Linear separator in $\mathbb{R}^M$

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

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)

PAC Learning 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