

Computational Learning Theory 2: VC Dimension


Somak Aditya, Sudeshna Sarkar

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} $		

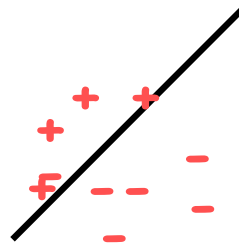




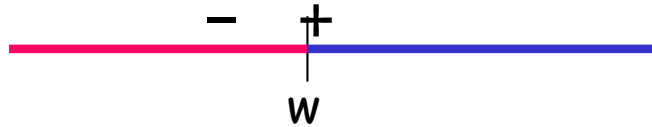
What if H is infinite?



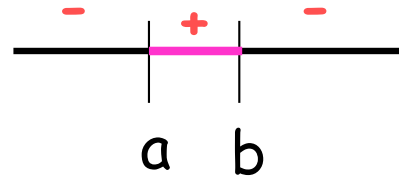
E.g., linear separators in \mathbb{R}^d



E.g., thresholds on the real line



E.g., intervals on the real line

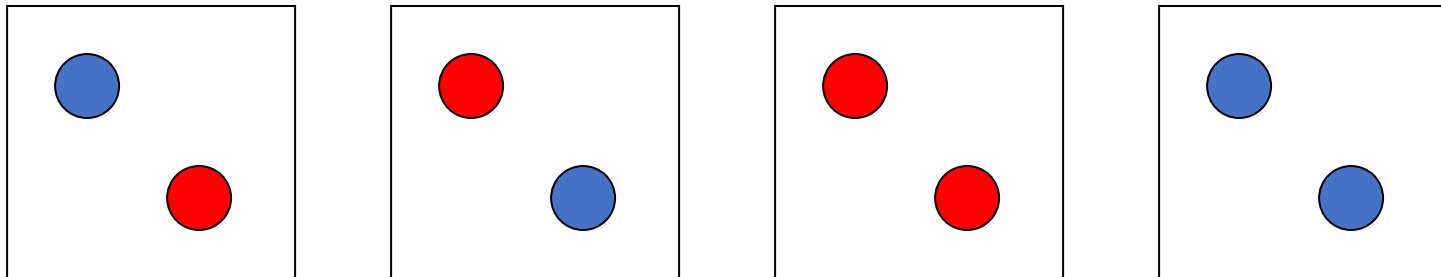


Sample Complexity: Infinite Hypothesis Spaces

- Need some measure of the expressiveness of infinite hypothesis spaces.
- The *Vapnik-Chervonenkis (VC) dimension* provides such a measure, denoted $VC(H)$.
 - Analogous to $\ln |H|$, there are bounds for sample complexity using $VC(H)$.

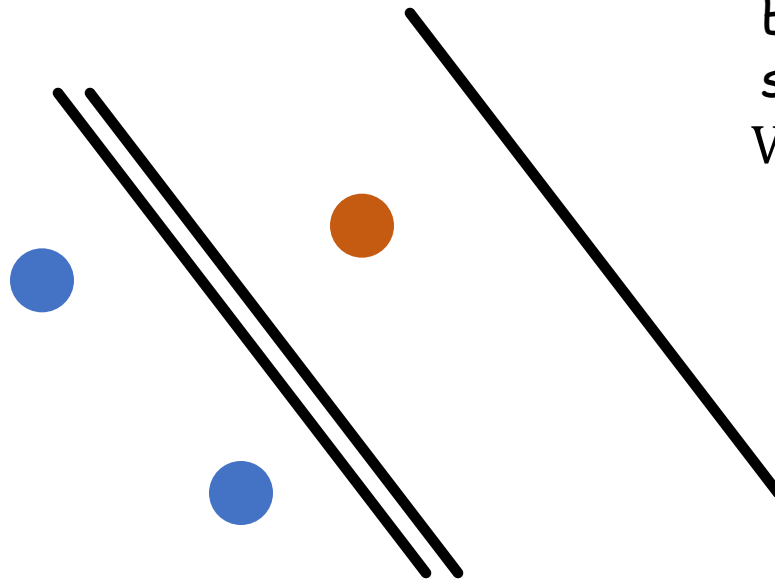
Shattering

- Consider a hypothesis for the 2-class problem.
- A set of N points (instances) can be labeled as $+$ or $-$ in 2^N ways.
- If for every such labeling a function can be found in \mathcal{H} consistent with this labeling, we state
 - that **the set of instances is shattered by \mathcal{H} .**



Three points in \mathbb{R}^2

- It is enough to find one set of three points that can be shattered.
- It is not necessary to be able to shatter **every possible set of three points** in 2 dimensions

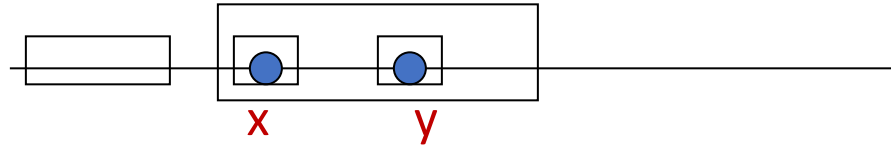


E.g., $H =$ linear
separators in \mathbb{R}^2
 $\text{VCdim } H \geq 3$



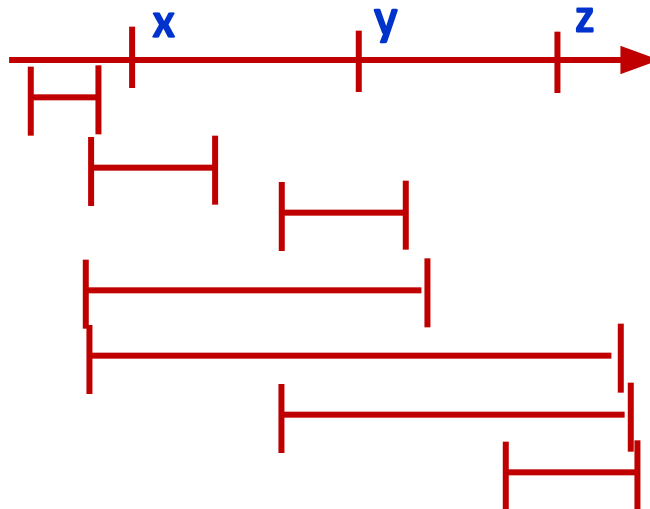
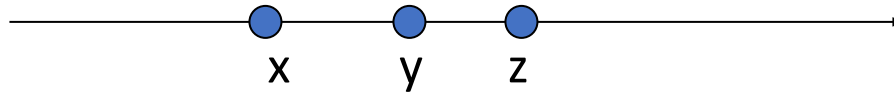
Shattering Instances

- Consider 2 instances described using a single real-valued feature being shattered by a single interval.



Shattering Instances (cont)

But 3 instances cannot be shattered by a single interval.



	+	-
-		x, y, z
x		y, z
y		x, z
x, y		z
x, y, z		
y, z		x
z		x, y
x, z		y

Shattering, VC Dimension (Formal)

- X is set of instances.
- S is a subset.
- H is a hypothesis class over instance space X
 - Set of functions from X to $\{0,1\}$
- For any $S \subseteq X$, $H(S)$ is set of all partitions induced by H . $H(S) \subseteq \{0,1\}^m$

$$H(S) = \{(c(x_1), \dots, c(x_m)) ; c \in H\}.$$

- For any m , $H[m]$ is maximum number of ways to split m points using concepts in H

$$H[m] = \max\{|H(S)| ; |S| = m, S \subseteq X\}$$

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 .

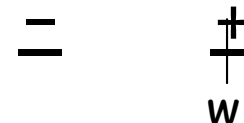
If arbitrarily large finite sets can be shattered by H , then

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

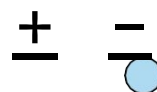
Shattering, VC-dimension

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.

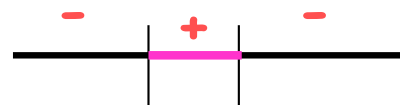
E.g., H = Thresholds on the real line



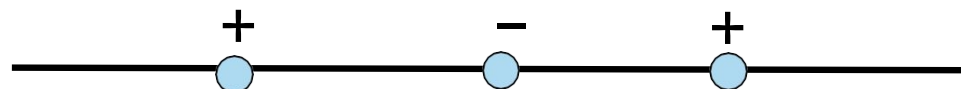
$VCdim(H) = 1$



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



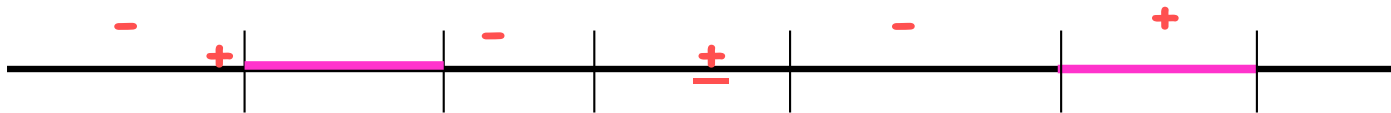
$VCdim(H) = 2$



Shattering, VC-dimension (Examples)

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)

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



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

What can we Infer?

- If there exists at least one subset of S of size d that can be shattered then $\text{VC}(H) \geq d$.
- If no subset of size d can be shattered, then $\text{VC}(H) < d$.
- For a single intervals on the real line, all sets of 2 instances can be shattered, but no set of 3 instances can, so $\text{VC}(H) = 2$.

VC Dimension – Inferences II

- An unbiased hypothesis space shatters the entire instance space.
- The larger the subset of S that can be shattered, the more expressive (and less biased) the hypothesis space is.
- The VC dimension of the set of oriented lines in 2-d is three.

PROOF OF $VC(H) \leq \log_2 |H|$

- Since there are 2^m partitions of m instances, in order for H to shatter instances: $|H| \geq 2^m$.
- Since $|H| \geq 2^m$, to shatter m instances, $VC(H) \leq \log_2 |H|$

What we have managed to prove

Concept class	VC Dimension	Why?
Half intervals	1	There is a dataset of size 1 that can be shattered No dataset of size 2 can be shattered
Intervals	2	There is a dataset of size 2 that can be shattered No dataset of size 3 can be shattered
Half-spaces in the plane	3	There is a dataset of size 3 that can be shattered No dataset of size 4 can be shattered

More VC dimensions

Concept class	VC Dimension
Linear threshold unit in d dimensions	$d + 1$
Neural networks	Number of parameters
1 nearest neighbors	infinite

What is the number of parameters
needed to specify a linear threshold unit
in d dimensions? $d + 1$

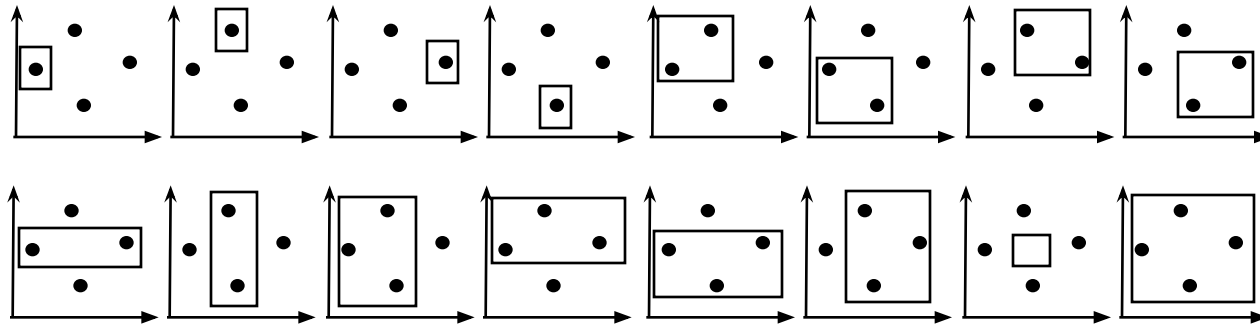
Local minima in learning means neural
networks may not find the best
parameters

Exercise: Try to prove this after we see
nearest neighbors

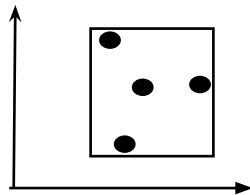
Intuition: A rich set of functions shatters large sets of points

VC Dimension Example

Consider axis-parallel rectangles in the real-plane, i.e. conjunctions of intervals on two real-valued features. Some 4 instances can be shattered.

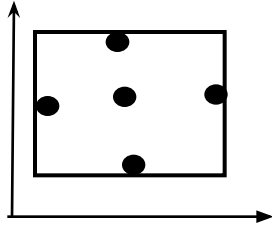


Some 4 instances cannot be shattered:



VC Dimension Example (cont)

- No five instances can be shattered since there can be at most 4 distinct extreme points (min and max on each of the 2 dimensions) and these 4 cannot be included without including any possible 5th point.



- Therefore $VC(H) = 4$
- Generalizes to axis-parallel hyper-rectangles (conjunctions of intervals in n dimensions): $VC(H)=2n$.

Upper Bound on Sample Complexity with VC

- Using VC dimension as a measure of expressiveness, the following number of examples have been shown to be sufficient for PAC Learning (Blumer *et al.*, 1989, 1986).

$$\frac{1}{\varepsilon} \left(4 \log_2 \left(\frac{2}{\delta} \right) + 8VC(H) \log_2 \left(\frac{13}{\varepsilon} \right) \right)$$

- (For finite H) Compared to the previous result using $\ln |H|$, this bound has some extra constants and an extra $\log_2(1/\varepsilon)$ factor. Since $VC(H) \leq \log_2 |H|$, this can provide a tighter upper bound on the number of examples needed for PAC learning.

Sample Complexity Lower Bound with VC

- There is also a general lower bound on the minimum number of examples necessary for PAC learning (Ehrenfeucht, *et al.*, 1989):
 - Consider any concept class C such that $VC(C) > 2$, any learner L and any $0 < \varepsilon < 1/8, 0 < \delta < 1/100$. Then there exists a distribution D and target concept in C such that if L observes fewer than:

$$\max\left(\frac{1}{\varepsilon} \log_2\left(\frac{1}{\delta}\right), \frac{VC(C)-1}{32\varepsilon}\right)$$

examples, then with probability at least δ , L outputs a hypothesis having error greater than ε .

- Ignoring constant factors, this lower bound is the same as the upper bound except for the extra $\log_2(1/\varepsilon)$ factor in the upper bound.

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$.

