Review of Lecture 5

Instead of

$$\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 2 \quad M \quad e^{-2N\epsilon^2}$$

we seek to replace the growth function

$$\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \stackrel{?}{\leq} 2 \quad m_{\mathcal{H}}(N) \quad e^{-2N\epsilon^2}$$

We proved $m_{\mathcal{H}}(N)$ can be bounded by a polynomial

$$m_{\mathcal{H}}(N) \leq B(N,k) \leq \text{polynomial}(N)$$

$B(N,k)$ Maximum number of dichotomies on N points, with break point k

Review of Lecture 5

- $B(N,k)$ for N=3, k=2

‘Puzzle’

Start here

x_1	x_2	x_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

- $B(N,k)$ for N=3, k=2

$$B(3,2) = 4$$

x_1	x_2	x_3
~~0~~	~~0~~	~~0~~
~~0~~	~~0~~	~~1~~
~~0~~	~~1~~	~~0~~
0	1	1
~~1~~	~~0~~	~~0~~
1	0	1
1	1	0
1	1	1

Start here

Review of Lecture 5

- Recursive bound on $B(N,k)$

$B(N,k) = \alpha + 2\beta$

$\alpha + \beta \le B(N-1,k)$

$\beta \le B(N-1,k-1)$

$B(N,k) \le$

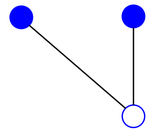
$B(N-1,k) + B(N-1,k-1)$

	# of rows	x_1	x_2	...	x_{N-1}	x_N
S_1	α	0	0	...	0	0
		1	0	...	0	1
		⋮	⋮	⋮	⋮	⋮
		0	1	...	1	1
		1	0	...	1	0
S_2^-	β	0	1	...	0	0
		1	1	...	0	0
		⋮	⋮	⋮	⋮	⋮
		0	1	...	0	0
		1	1	...	1	0
S_2^+	β	0	1	...	0	1
		1	1	...	0	1
		⋮	⋮	⋮	⋮	⋮
		0	1	...	0	1
		1	1	...	1	1

- $m_{\mathcal{H}}(N)$ is polynomial

if $\mathcal{H}$ has a break point k

		k						
		1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1	3	4	4	4	4	..
	3	1	4	7	8	8	8	..
N	4	1	5	..	..			
	5	1	6					
	6	1	7					
	:	:	:					



$$m_{\mathcal{H}}(N) \le B(N,k) \le B(N-1,k) + B(N-1,k-1) \le \sum_{i=0}^{k-1} \binom{N}{i}$$

maximum power is N^{k-1}

ECE 6254
Statistical Machine Learning

Professor: Amirali Aghazadeh
Office: Coda S1209
Georgia Institute of Technology

Lecture 6: The VC Dimension

Georgia Institute of Technology

Outline

- Prove that $m_{\mathcal{H}}(N)$ is polynomial 
- Prove that $m_{\mathcal{H}}(N)$ can replace M 
- The VC dimension
- VC dimension of Perceptrons
- Interpreting the VC dimension

A (tighter) bound with the growth function

Instead of $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 2 \quad M \quad e^{-2N\epsilon^2}$

we want $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \overset{?}{\leq} 2 \quad m_{\mathcal{H}}(N) \quad e^{-2N\epsilon^2}$

With probability at least $1-\delta$: $R(h^*) \overset{?}{\leq} \hat{R}_N(h^*) + \sqrt{\frac{1}{2N}\ln\frac{2m_{\mathcal{H}}(N)}{\delta}}$

The VC Generalization Bound

Instead of $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 2 \quad M \quad e^{-2N\epsilon^2}$

we want $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \overset{?}{\leq} 2 \quad m_{\mathcal{H}}(N) \quad e^{-2N\epsilon^2}$

$$\boxed{\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}N\epsilon^2}}$$

With probability at least $1 - \delta$: $R(h^*) \overset{?}{\leq} \hat{R}_N(h^*) + \sqrt{\frac{1}{2N}\ln\frac{2m_{\mathcal{H}}(N)}{\delta}}$

The VC Generalization Bound

Instead of $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 2 \quad M \quad e^{-2N\epsilon^2}$

we want $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \overset{?}{\leq} 2 \quad m_{\mathcal{H}}(N) \quad e^{-2N\epsilon^2}$

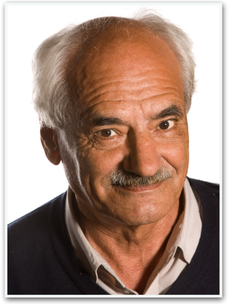
$$\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}N\epsilon^2}$$

With probability at least $1-\delta$: $R(h^*) \overset{?}{\leq} \hat{R}_N(h^*) + \sqrt{\frac{1}{2N}\ln\frac{2m_{\mathcal{H}}(N)}{\delta}}$

With probability at least $1-\delta$: $R(h^*) \leq \hat{R}_N(h^*) + \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$

The VC Generalization Bound


Vladimir Vapnik


Alexey Chervonenkis

1960-1990

Instead of $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 2 \quad M \quad e^{-2N\epsilon^2}$

we want $\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \overset{?}{\leq} 2 \quad m_{\mathcal{H}}(N) \quad e^{-2N\epsilon^2}$

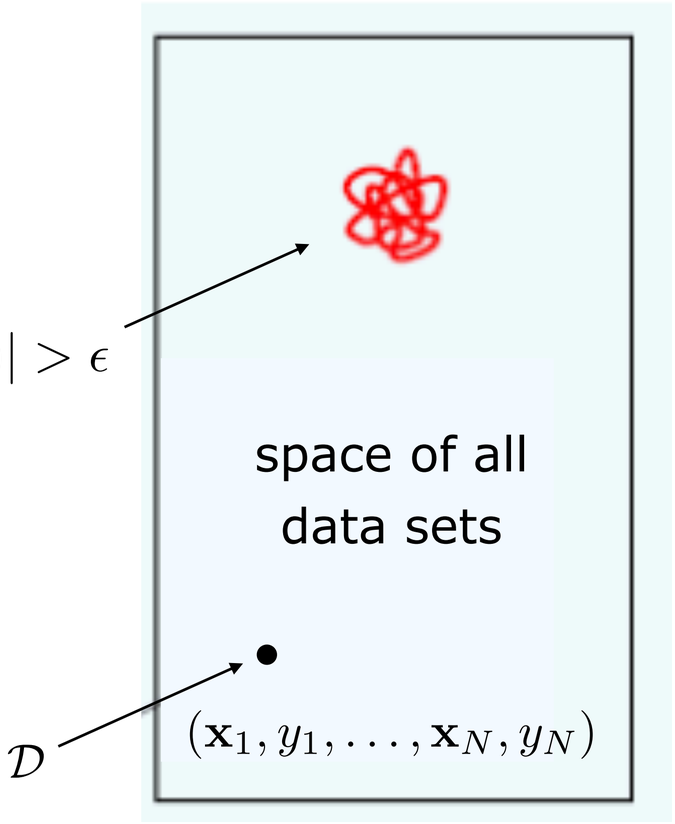
$$\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}N\epsilon^2}$$

With probability at least $1-\delta$: $R(h^*) \overset{?}{\leq} \hat{R}_N(h^*) + \sqrt{\frac{1}{2N}\ln\frac{2m_{\mathcal{H}}(N)}{\delta}}$

With probability at least $1-\delta$: $R(h^*) \leq \hat{R}_N(h^*) + \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$

VC Bound: Intuition

Hoeffding Inequality

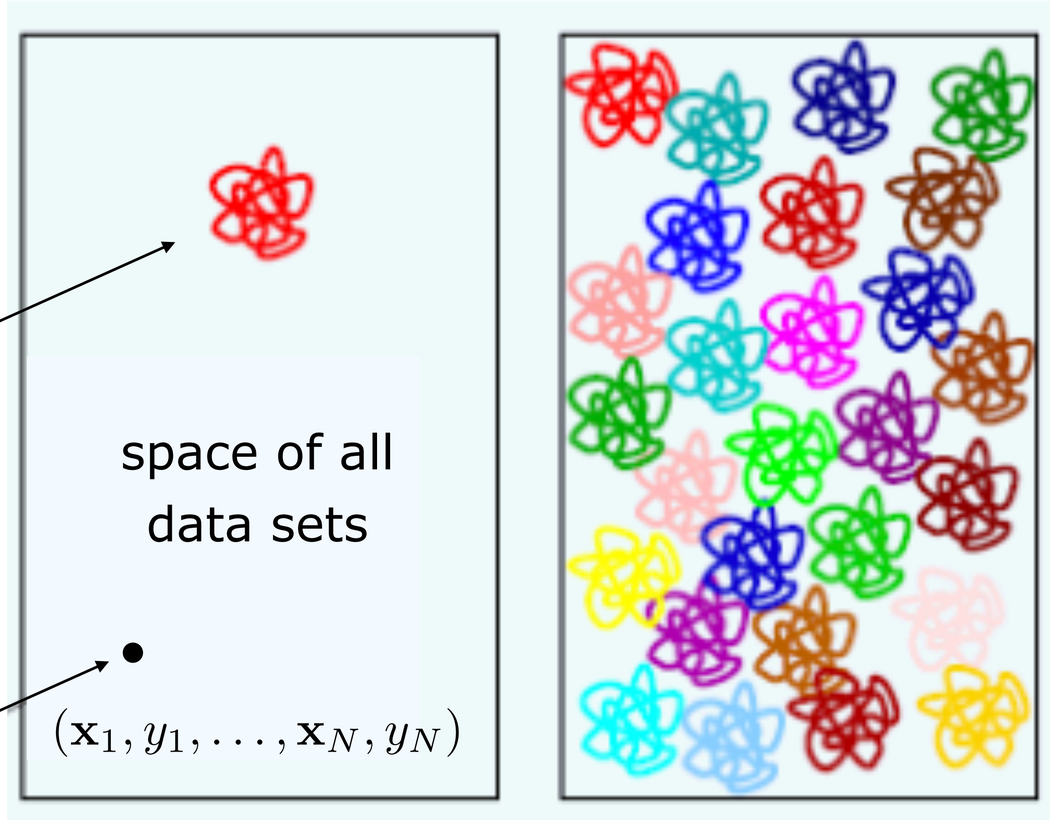


- choose a fixed h

VC Bound: Intuition

Hoeffding Inequality

Union Bound



- choose a fixed h

- consider all h in H
- assume no overlap

VC Bound: Intuition

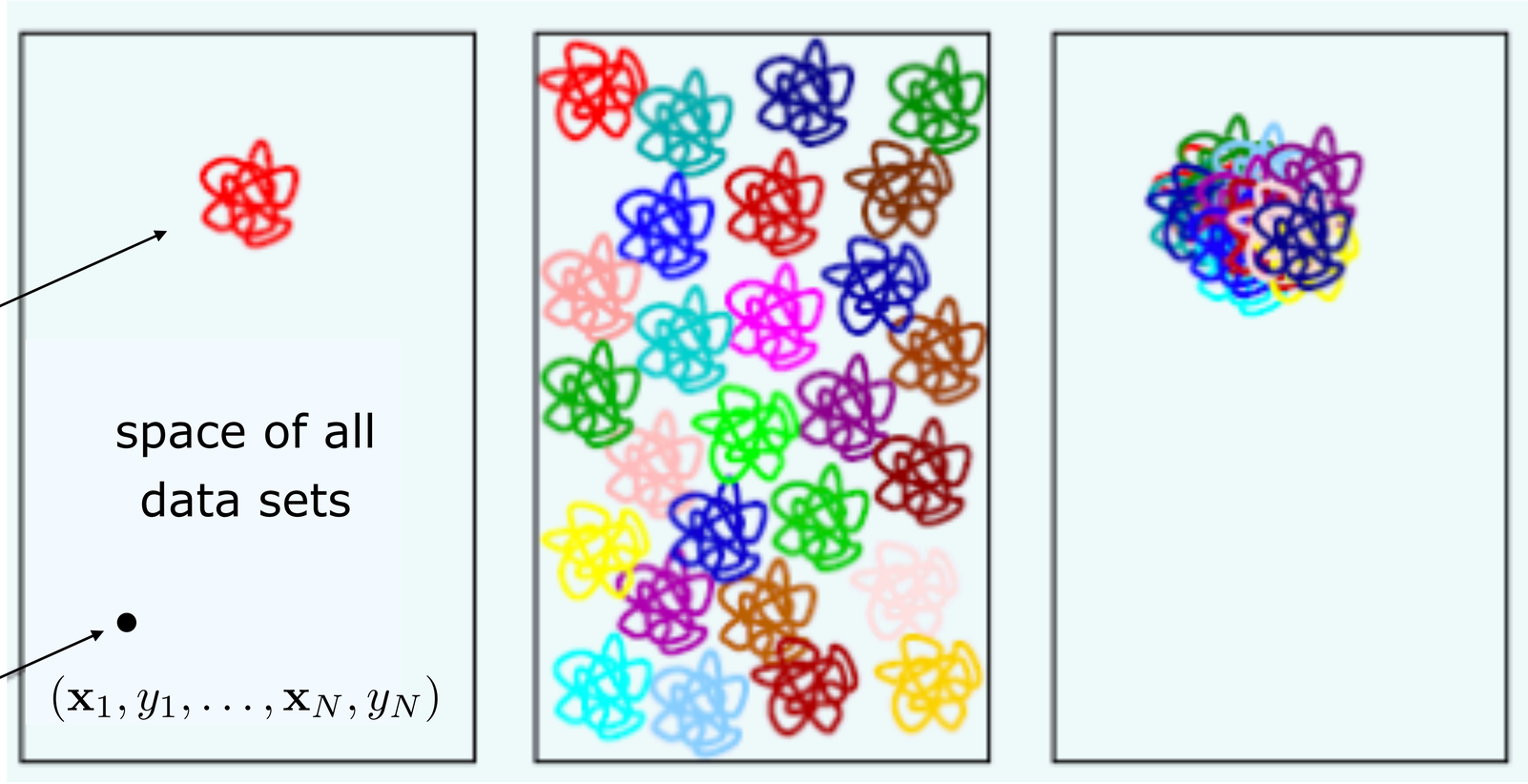
Hoeffding Inequality

Union Bound

VC Bound

$|\hat{R}_N(h) - R(h)| > \epsilon$

(bad events)



- choose a fixed h

- consider all h in H
- assume no overlap

- keeps track of overlaps

VC Bound: Proof Challenges

$$\mathbb{P}(\sup_{h\in\mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}N\epsilon^2}$$

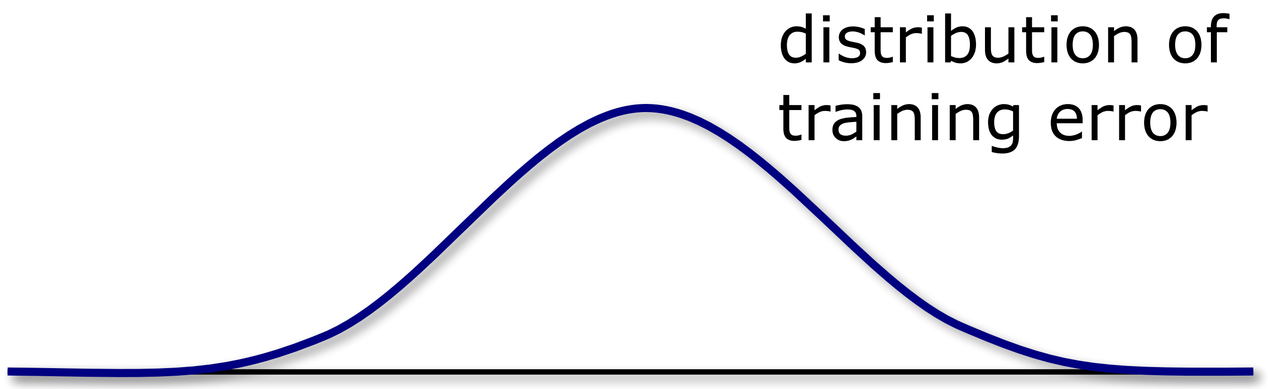
- How does $m_{\mathcal{H}}(N)$ relates to overlaps?
- $R(h)$ is difficult to manipulate compared to $\hat{R}_N(h)$
  - Focusing on finite many dichotomies $m_{\mathcal{H}}(N)$
  - $\hat{R}_N(h)$ is finite but $R(h)$ can still take infinitely many values

What to do with $R(h)$?

- Key insight is to sample a second data set
- In addition to the training data with draw a **ghost data** (only for proof)

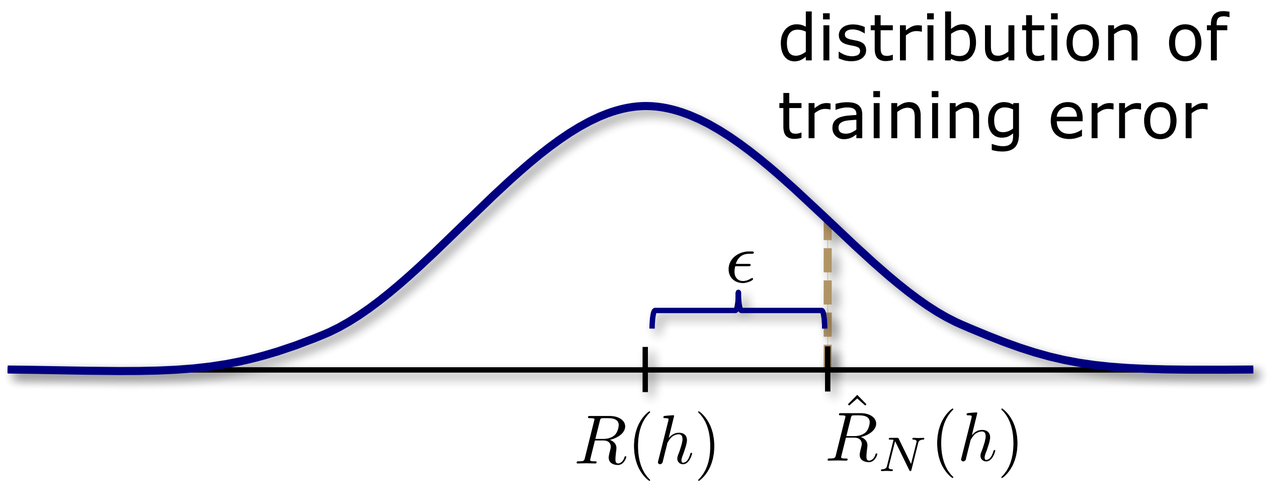
What to do with $R(h)$?

- Key insight is to sample a second data set
- In addition to the training data with draw a **ghost data** (only for proof)



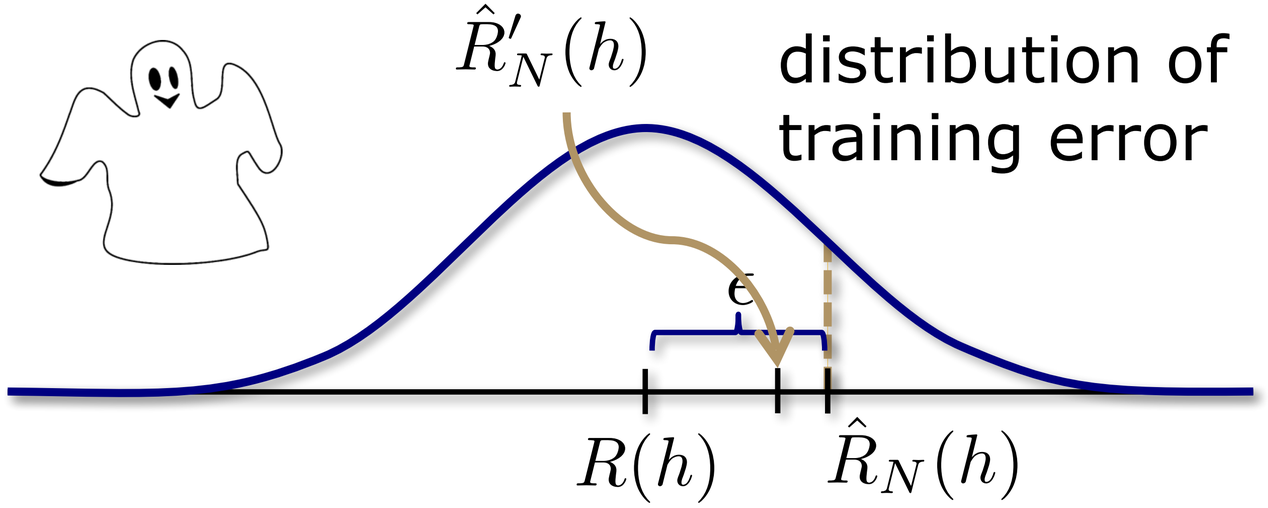
What to do with $R(h)$?

- Key insight is to sample a second data set
- In addition to the training data with draw a **ghost data** (only for proof)



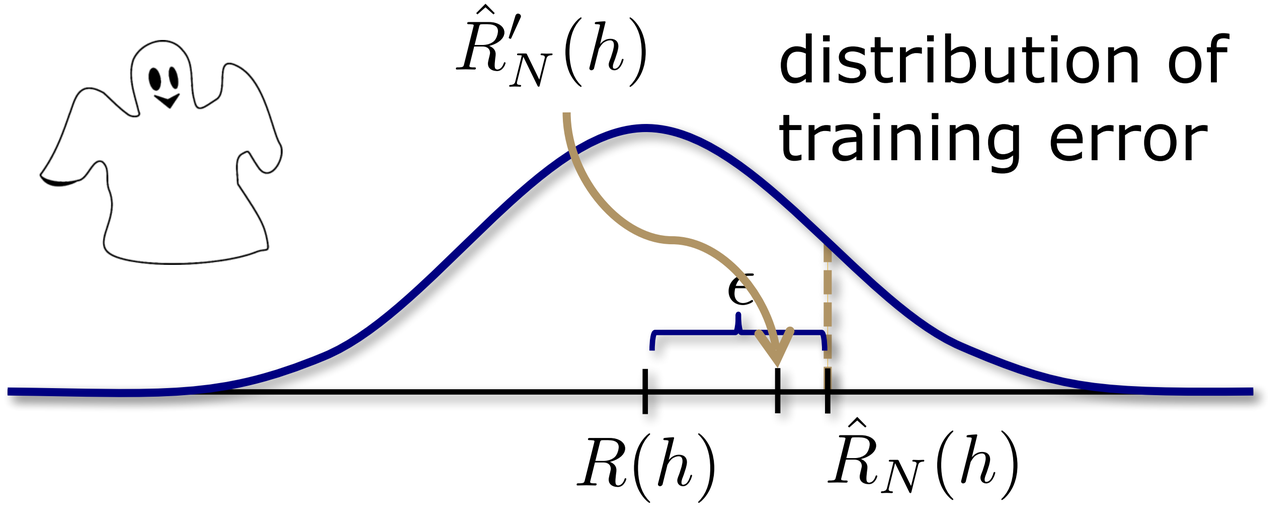
What to do with $R(h)$?

- Key insight is to sample a second data set
- In addition to the training data with draw a **ghost data** (only for proof)



What to do with $R(h)$?

- Key insight is to sample a second data set
- In addition to the training data with draw a **ghost data** (only for proof)

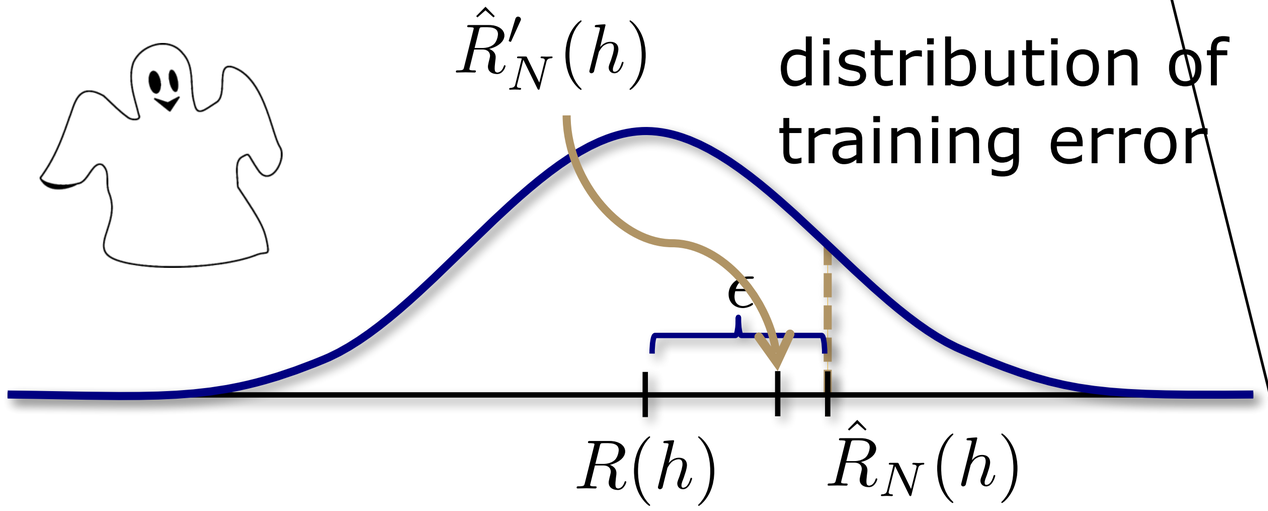


$$(1 - e^{-\frac{1}{2}\epsilon^2 N})\mathbb{P}[\sup_{h\in\mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon] \leq \mathbb{P}[\sup_{h\in\mathcal{H}} |\hat{R}_N(h) - \hat{R}'_N(h)| > \frac{\epsilon}{2}]$$

See the posted notes for full proof

What to do with $R(h)$?

$$\mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}N\epsilon^2}$$



$$(1 - e^{-\frac{1}{2}\epsilon^2 N})\mathbb{P}[\sup_{h\in\mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon] \leq \mathbb{P}[\sup_{h\in\mathcal{H}} |\hat{R}_N(h) - \hat{R}'_N(h)| > \frac{\epsilon}{2}]$$

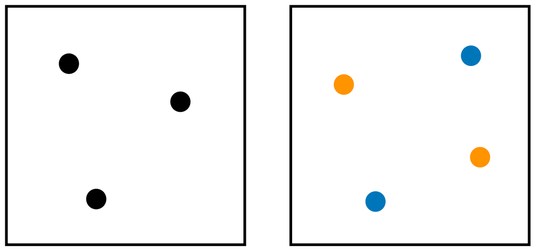
The VC Dimension

The VC dimension $d_{VC}(\mathcal{H})$ of a hypothesis set $\mathcal{H}$ is

- the largest value of N where $m_{\mathcal{H}}(N) = 2^N$

  "the most points $\mathcal{H}$ can shatter"

- $d_{VC} = k - 1$ (where k is the breakpoint)



$m_{\mathcal{H}}(3) = 8$ $m_{\mathcal{H}}(4) = 14$

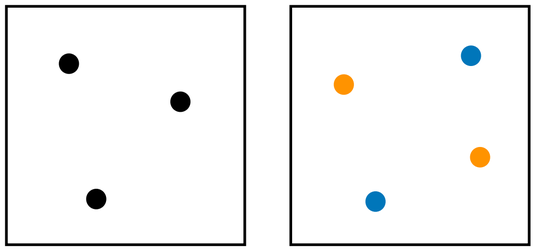
The VC Dimension

The VC dimension $d_{VC}(\mathcal{H})$ of a hypothesis set $\mathcal{H}$ is

- the largest value of N where $m_{\mathcal{H}}(N) = 2^N$

  "the most points $\mathcal{H}$ can shatter"

- $d_{VC} = k - 1$ (where k is the breakpoint)



$m_{\mathcal{H}}(3) = 8$

$m_{\mathcal{H}}(4) = 14$

$$d_{VC} \geq N \iff \text{there exists } \mathcal{D} \text{ of size } N \text{ such that } \mathcal{H} \text{ shatters } \mathcal{D}$$

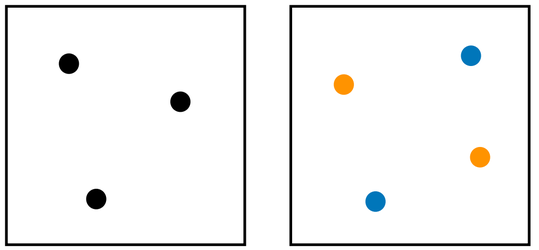
The VC Dimension

The VC dimension $d_{VC}(\mathcal{H})$ of a hypothesis set $\mathcal{H}$ is

- the largest value of N where $m_{\mathcal{H}}(N) = 2^N$

  "the most points $\mathcal{H}$ can shatter"

- $d_{VC} = k - 1$ (where k is the breakpoint)



$m_{\mathcal{H}}(3) = 8$ $m_{\mathcal{H}}(4) = 14$

$$d_{VC} \geq N \iff \text{there exists } \mathcal{D} \text{ of size } N \text{ such that } \mathcal{H} \text{ shatters } \mathcal{D}$$

$$d_{VC} < N \iff \text{no set of } N \text{ points can be shattered by } \mathcal{H}$$

The growth function

As a function of the break point k

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

As a function of the VC dimension d_{VC}

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

$$\underbrace{\qquad\qquad}_{\text{maximum power is } N^{d_{VC}}}$$

Examples

$\mathcal{H}$ is positive rays:

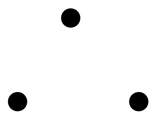
Examples

$\mathcal{H}$ is positive rays: $d_{VC} = 1$

Examples

$\mathcal{H}$ is positive rays: $d_{VC} = 1$

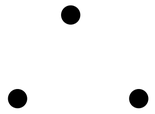
$\mathcal{H}$ is 2D perceptrons: $d_{VC} = 3$



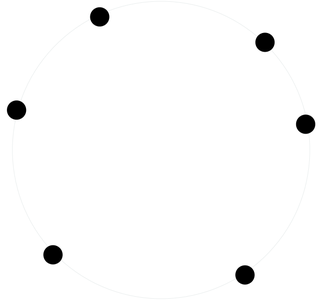
Examples

$\mathcal{H}$ is positive rays: $d_{VC} = 1$

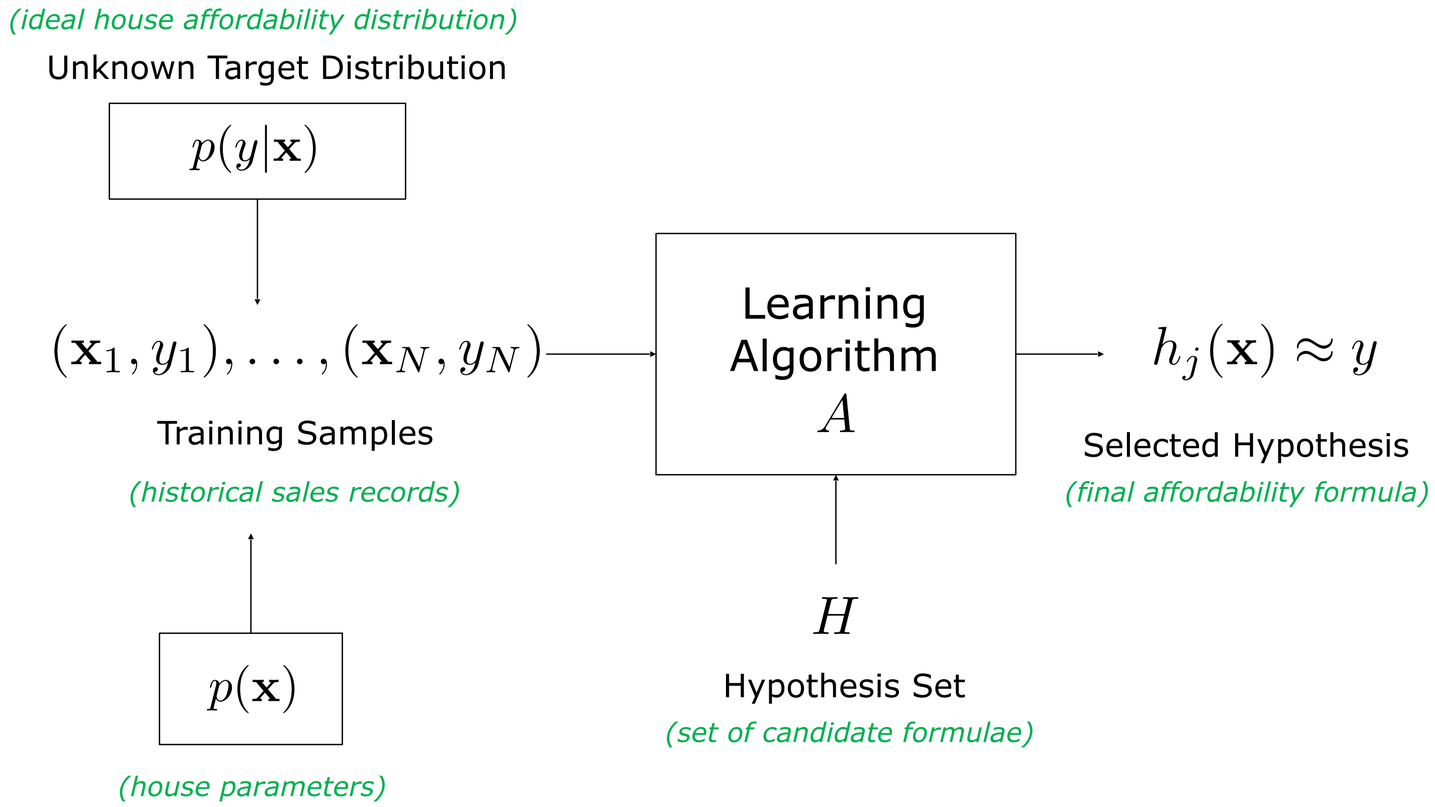
$\mathcal{H}$ is 2D perceptrons: $d_{VC} = 3$



$\mathcal{H}$ is convex sets: $d_{VC} = \infty$



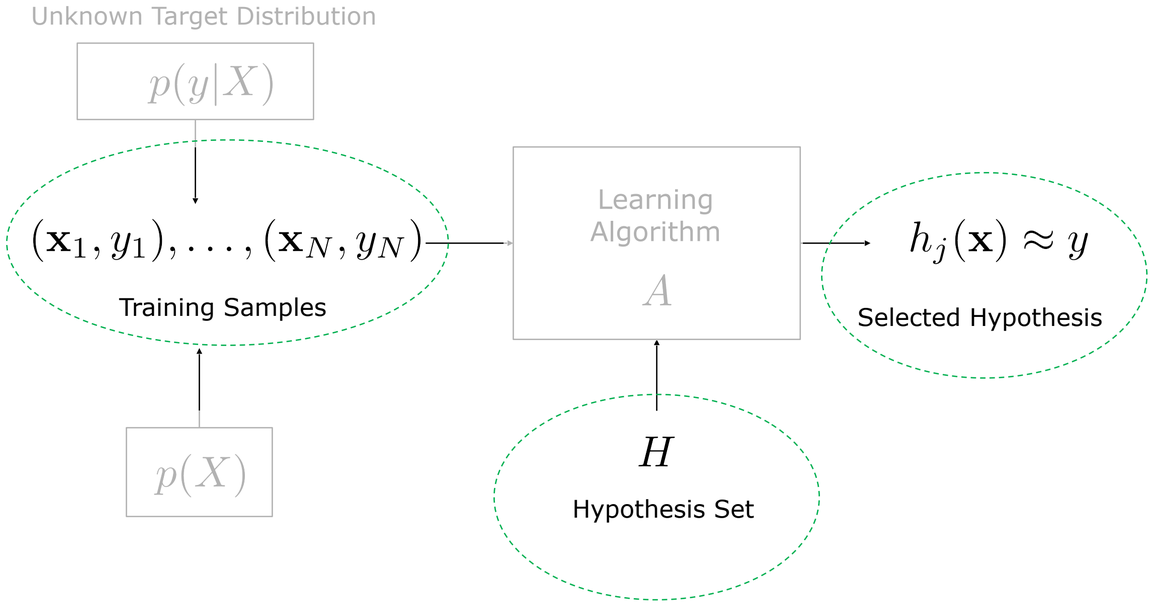
VC dimension and learning



VC dimension and learning

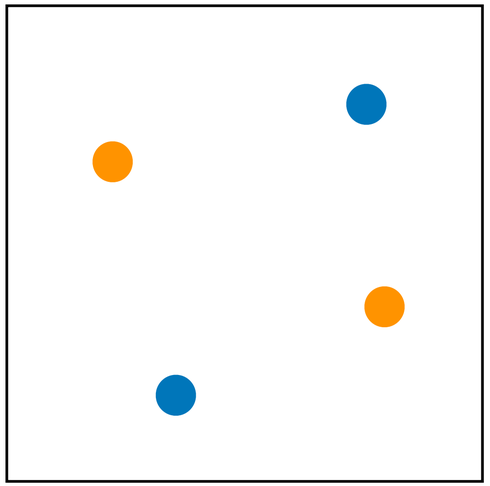
$d_{VC}(\mathcal{H})$ is finite $\implies h^* \in \mathcal{H}$ will generalize

- Independent of the *learning algorithm*
- Independent of the *input distribution*
- Independent of the *target distribution*



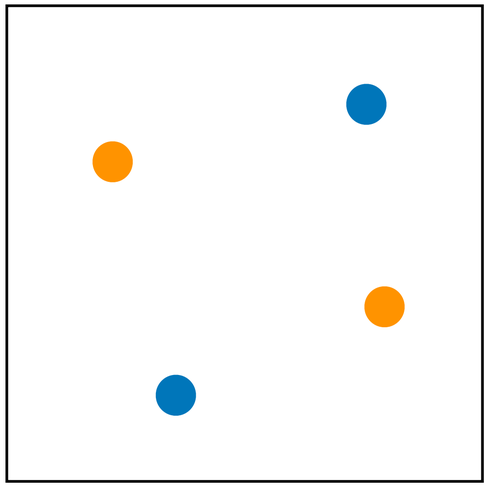
VC dimension of perceptrons

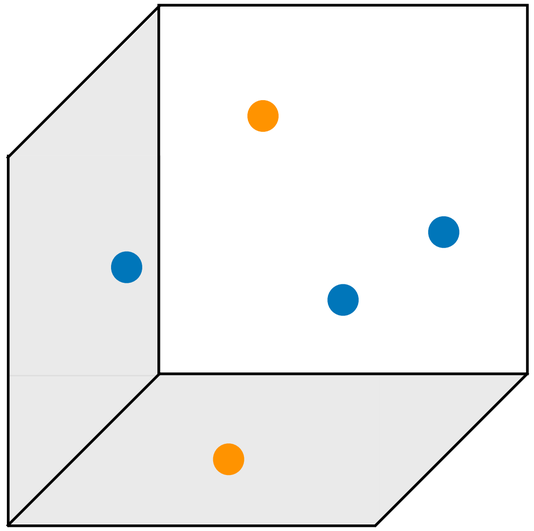
$d = 2, d_{VC} = 3$



VC dimension of perceptrons

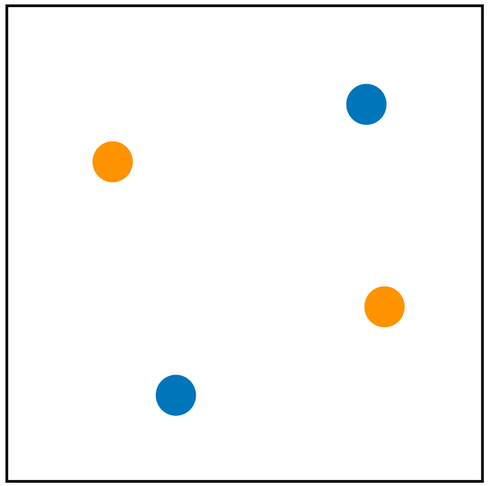
$d = 2, d_{VC} = 3$



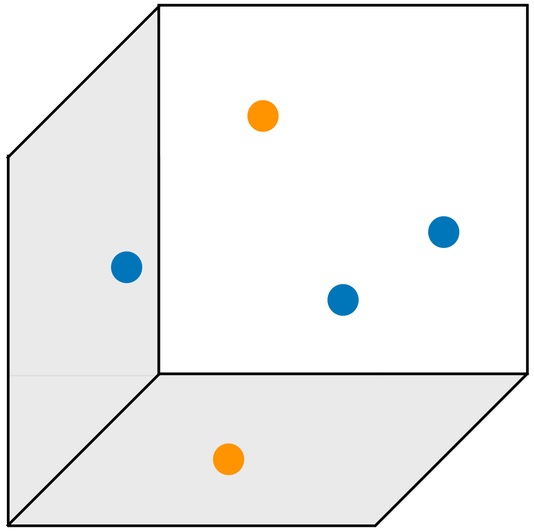


VC dimension of perceptrons

$d = 2, d_{VC} = 3$

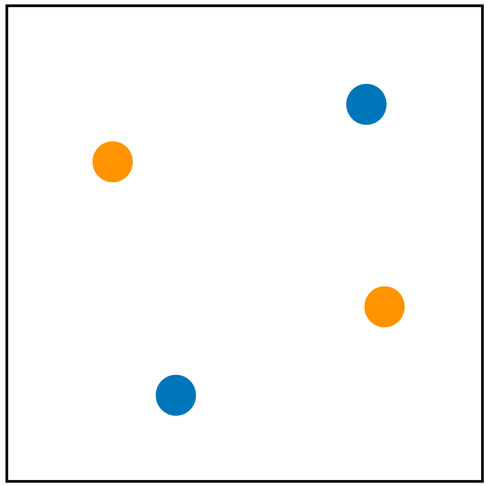


$d = 3, d_{VC} = 4$

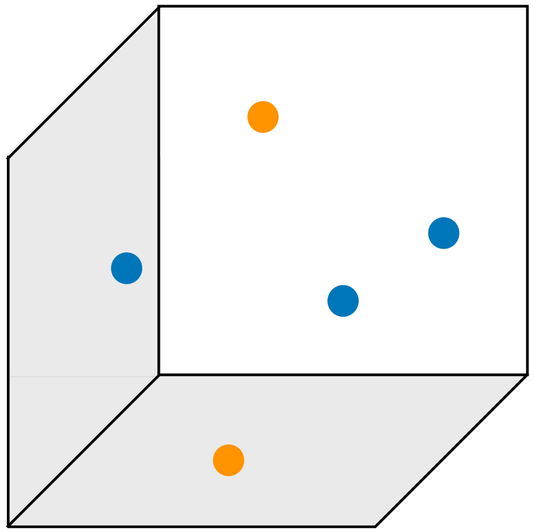


VC dimension of perceptrons

$d = 2, d_{VC} = 3$



$d = 3, d_{VC} = 4$



$d_{VC} = d + 1$

need to prove

$d_{VC} \geq d + 1$

$d_{VC} \leq d + 1$

One direction: Goal $d_{VC} \geq d+1$

$$d_{VC} \geq N \iff \text{there exists } \mathcal{D} \text{ of size } N \text{ such that } \mathcal{H} \text{ shatters } \mathcal{D}$$

Select a set of $N = d+1$ points in $\mathbb{R}^d$ such that:

$$\mathbf{X} = \underbrace{\begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_{d+1}^T- \end{bmatrix}}_{d+1} = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$\mathbf{X} \in \mathbb{R}^{N \times (d+1)}$ is invertible (i.e., full rank).

One direction: Goal $d_{VC} \geq d+1$

$$d_{VC} \geq N \iff \text{there exists } \mathcal{D} \text{ of size } N \text{ such that } \mathcal{H} \text{ shatters } \mathcal{D}$$

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ can we find a hypothesis $\mathbf{w}$ satisfying $\mathbf{y} = \text{sign}(\mathbf{X}\mathbf{w})$?

We can do much more using the full rank setup: $\mathbf{y} = \mathbf{X}\mathbf{w}$

$$\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$$

To finish the proof

To show $d_{\mathrm{VC}} \leq d + 1$, we need to show

a) There are $d + 1$ points we cannot shatter

b) There are $d + 2$ points we cannot shatter

c) We cannot shatter any set of $d + 1$ points

d) We cannot shatter any set of $d + 2$ points ✔

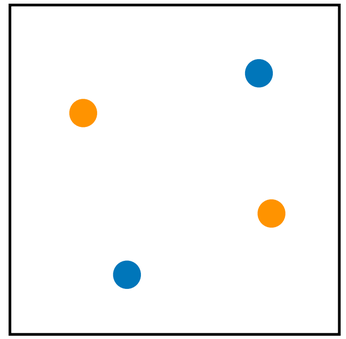
Other direction: Goal $d_{VC} \leq d + 1$

$$d_{VC} \leq N \iff \text{no set of } N+1 \text{ points can be shattered by } \mathcal{H}$$

Take any $d + 2$ points $\mathbf{x}_1, \ldots, \mathbf{x}_{d+2}$

More points than dimensions $\implies$ we must have $\mathbf{x}_j = \sum_{i \neq j} \alpha_i \mathbf{x}_i$

where not all $\alpha_i = 0$



Other direction: Goal $d_{VC} \leq d + 1$

$$d_{VC} \leq N \iff \text{no set of } N+1 \text{ points can be shattered by } \mathcal{H}$$

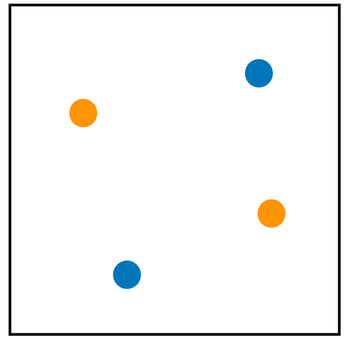
$$\mathbf{x}_j = \sum_{i \neq j} \alpha_i \mathbf{x}_i$$

Consider the following dichotomy:

$$\mathbf{x}_i\text{'s with non-zero } \alpha_i \text{ get } y_i = \text{sign}(\alpha_i)$$

$$\text{and } \mathbf{x}_j \text{ gets } y_j = -1$$

No perceptron can create such a dichotomy!



Why not?

$$\mathbf{x}_j = \sum_{i \neq j} \alpha_i \mathbf{x}_i \quad \Rightarrow \quad \mathbf{w}^{\mathrm{T}} \mathbf{x}_j = \sum_{i \neq j} \alpha_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i$$

If $y_i = \text{sign}(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i) = \text{sign}(\alpha_i)$ $\Rightarrow$ $\alpha_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i > 0$

This means $\mathbf{w}^{\mathrm{T}} \mathbf{x}_j = \sum_{i \neq j} \alpha_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i > 0$

Thus $y_j = \text{sign}(\mathbf{w}^{\mathrm{T}} \mathbf{x}_j) = +1$

Outline

- Prove that $m_{\mathcal{H}}(N)$ is polynomial
- Prove that $m_{\mathcal{H}}(N)$ can replace M
- The VC dimension
- VC dimension of Perceptrons
- Interpreting the VC dimension

Interpretation the VC dimension

We have just shown that for a perceptron in $\mathbb{R}^d$

$$d_{\mathsf{VC}} \geq d + 1$$

$$d_{\mathsf{VC}} \leq d + 1$$

$$d_{\mathsf{VC}} = d + 1$$

Interpretation the VC dimension

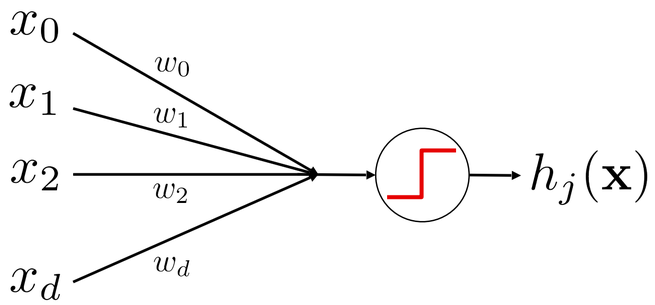
We have just shown that for a perceptron in $\mathbb{R}^d$

$$d_{\mathsf{VC}} \geq d + 1$$

$$d_{\mathsf{VC}} \leq d + 1$$

$$\Rightarrow \quad d_{\mathsf{VC}} = d + 1$$

How many parameters does a perceptron in $\mathbb{R}^d$ have?

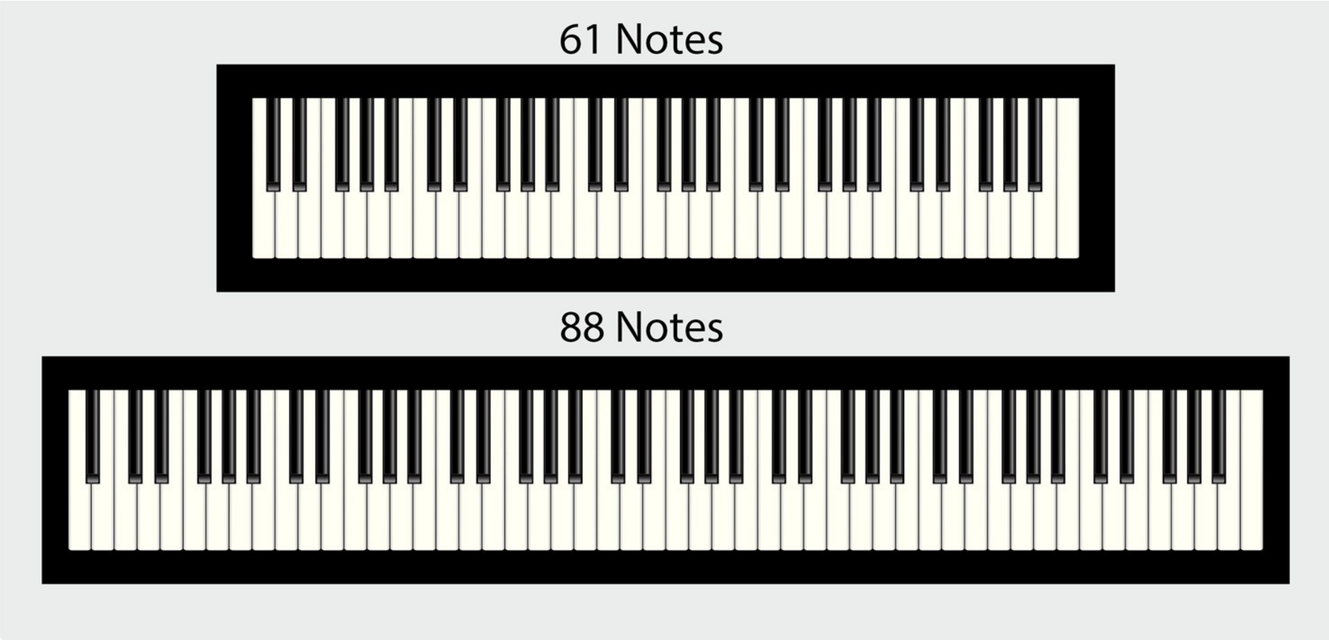


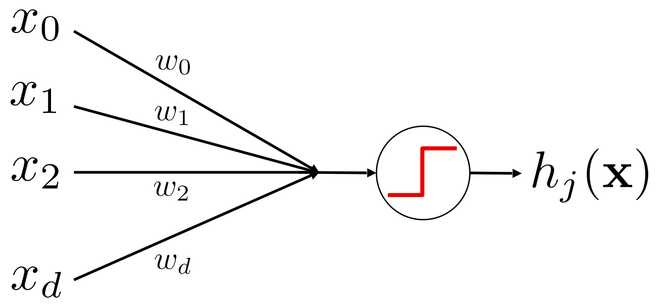
$$\Rightarrow \quad d + 1$$

Degrees of freedom

Hypothesis set determines the parameters

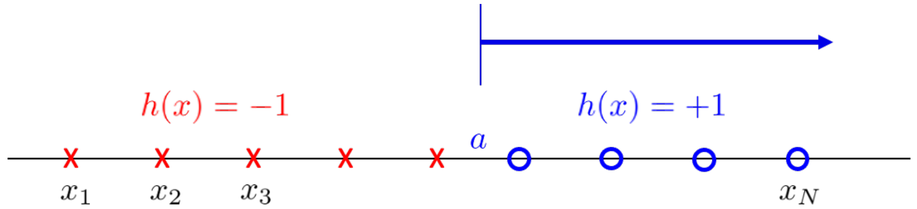
Parameters create ‘degrees of freedom’ ~ d_{VC}





The usual examples

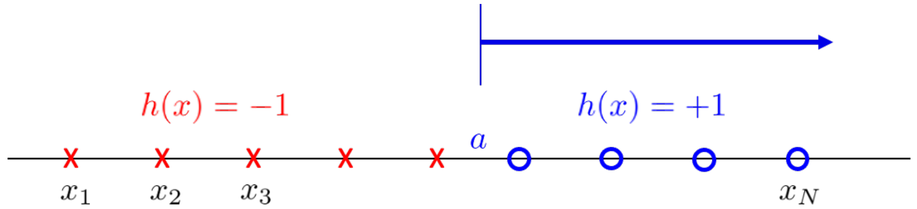
- Positive rays
  - $d_{\mathrm{VC}} = 1$
  - 1 parameter



$$h(x) = \mathrm{sign}(x - a)$$

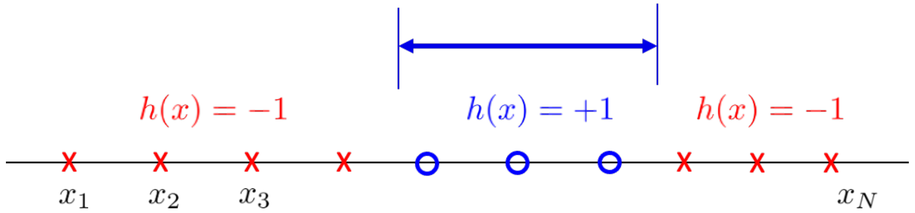
The usual examples

- Positive rays
  - $d_{\mathsf{VC}} = 1$
  - 1 parameter



$$h(x) = \text{sign}(x - a)$$

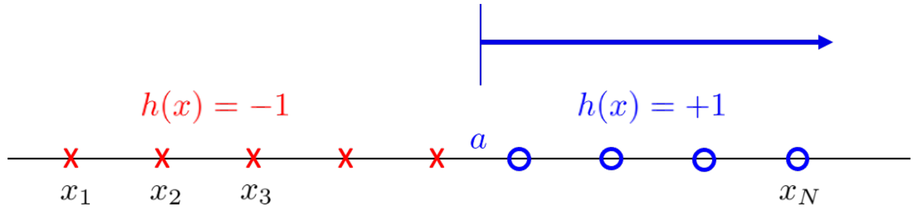
- Positive intervals
  - $d_{\mathsf{VC}} = 2$
  - 2 parameters



$$h(x) = \text{sign}(x - a) - \text{sign}(x - b)$$

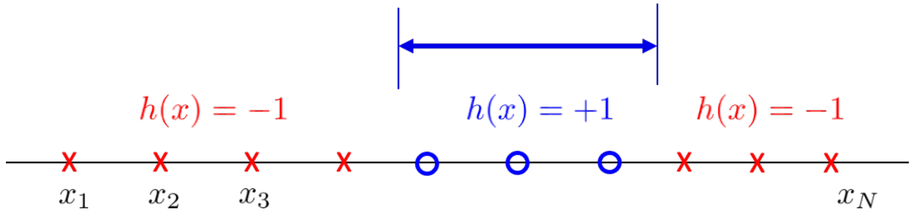
The usual examples

- Positive rays
  - $d_{\mathsf{VC}} = 1$
  - 1 parameter



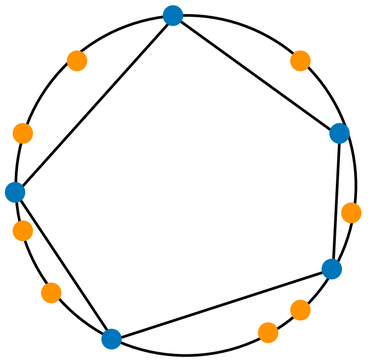
$$h(x) = \text{sign}(x - a)$$

- Positive intervals
  - $d_{\mathsf{VC}} = 2$
  - 2 parameters



$$h(x) = \text{sign}(x - a) - \text{sign}(x - b)$$

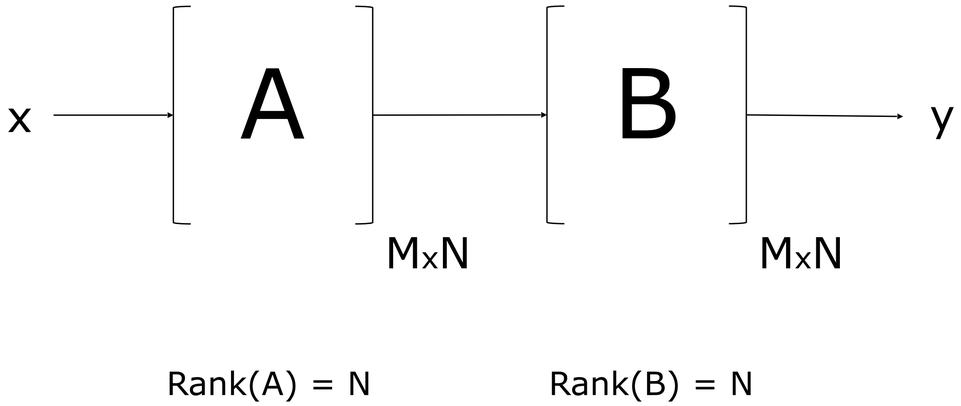
- Convex sets
  - $d_{\mathsf{VC}} = \infty$
  - as many parameters as you want



Effective number of parameters

More parameters do not always yield more degrees of freedom

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$



d_{VC} measures the effective number of parameters

VC ‘Rule-of-Thumb’: Number of data points needed

With probability at least $1-\delta$

$$R(h^*) \leq \hat{R}_N(h^*) + \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Plug in the bound for growth function

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1}\binom{N}{i} \leq 1 + N^{d_{VC}}$$

VC ‘Rule-of-Thumb’: Number of data points needed

With probability at least $1-\delta$

$$R(h^*) \leq \hat{R}_N(h^*) + \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Plug in the bound for growth function

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1}\binom{N}{i} \leq 1 + N^{d_{VC}}$$

$$R(h^*) \leq \hat{R}_N(h^*) + \sqrt{\frac{8}{N}\ln\frac{4((2N)^{d_{VC}}+1)}{\delta}}$$

$$R(h^*) \lessapprox \hat{R}_N(h^*) + \sqrt{\frac{8d_{VC}}{N}\ln\frac{8N}{\delta}}$$

VC ‘Rule-of-Thumb’: Number of data points needed

How big does our training set need to be (sample complexity)?

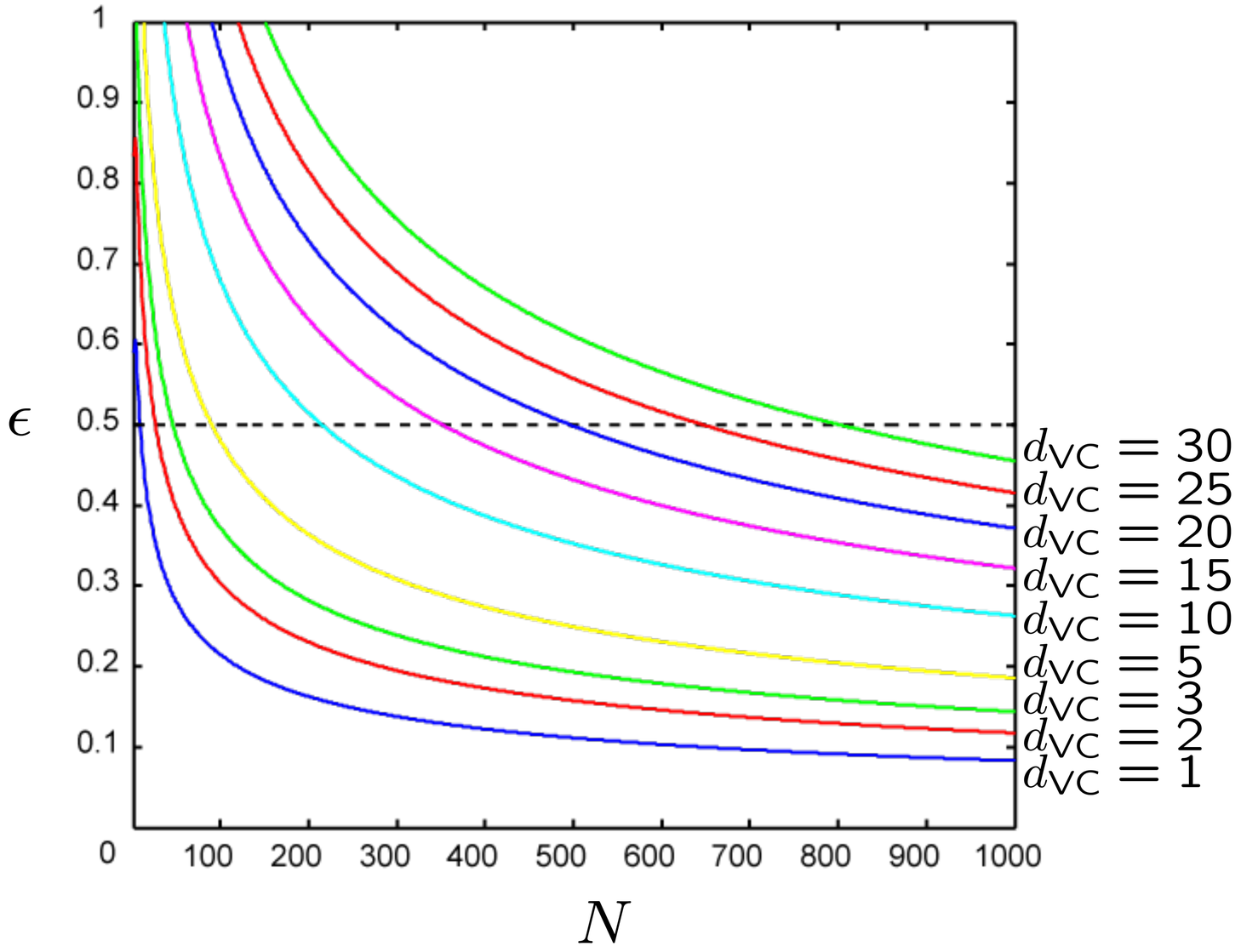
$$R(h^*) \lessapprox \hat{R}_N(h^*) + \underbrace{\sqrt{\frac{8d_{VC}}{N}\ln\frac{8N}{\delta}}}_{\epsilon}$$

To see this, let’s ignore the constants and suppose that

$$\epsilon \sim \sqrt{\frac{d_{VC}}{N}\ln N}$$

VC bound in action

$$\epsilon \sim \sqrt{\frac{d_{VC}}{N} \ln N}$$



RULE OF THUMB: $N \geq 20 d_{\text{VC}}$