Module 1

VC Dimention

Vapnik-Chervonenkis (VC) Dimension

- VC Dimension is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a space of functions that can be learned by a classification algorithm.

Shattering of a set

- Let D be a dataset containing N examples for a binary classification problem with class labels 0 and 1
- Let H be a hypothesis space for the problem
- Each hypothesis h in H partitions D into two disjoint subsets as follows:

$$\{x \in D \mid h(x) = 0\} \text{ and } \{x \in D \mid h(x) = 1\}.$$

Such a partition of S is called a “dichotomy” in D

Shattering of a set

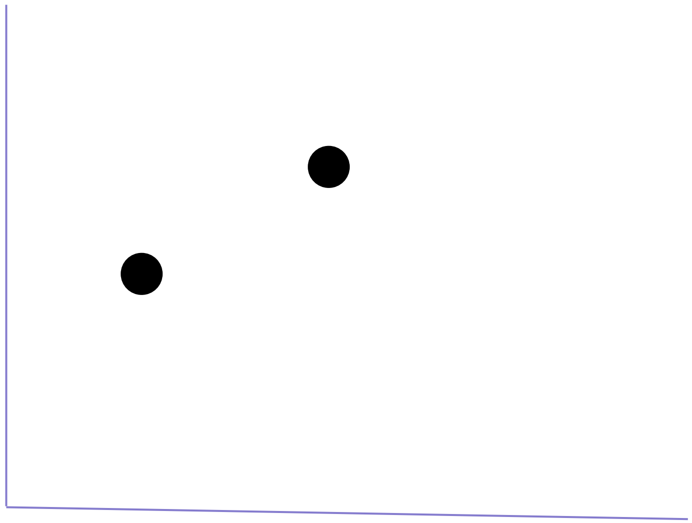
- There are 2^N possible dichotomies in D

- To each dichotomy of D there is a unique assignment of the labels "1" and "0" to the elements of D

- S is any subset of D then, S defines a unique hypothesis h as follows:

$$h(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

Vapnik-Chervonenkis (VC) Dimension



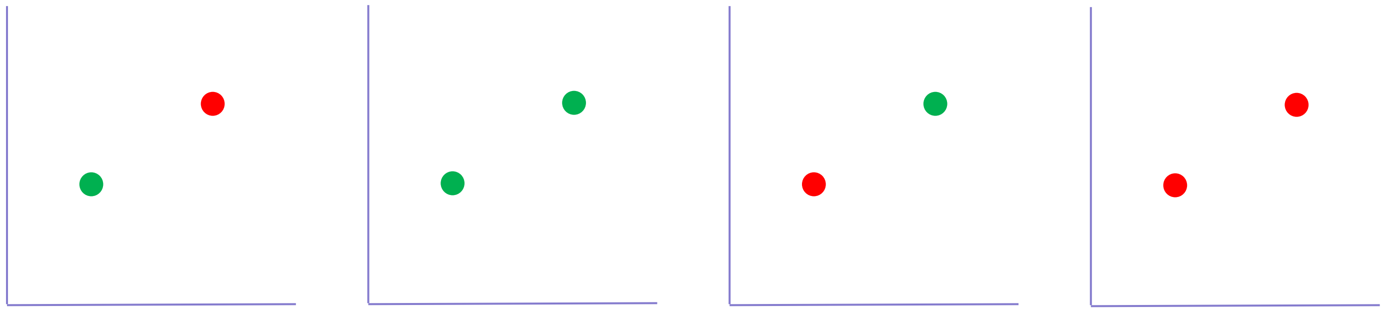
Total Data Points = 2

Classification

Class : A , true, 1, yes (green)

Class : B, false, 0, no (red)

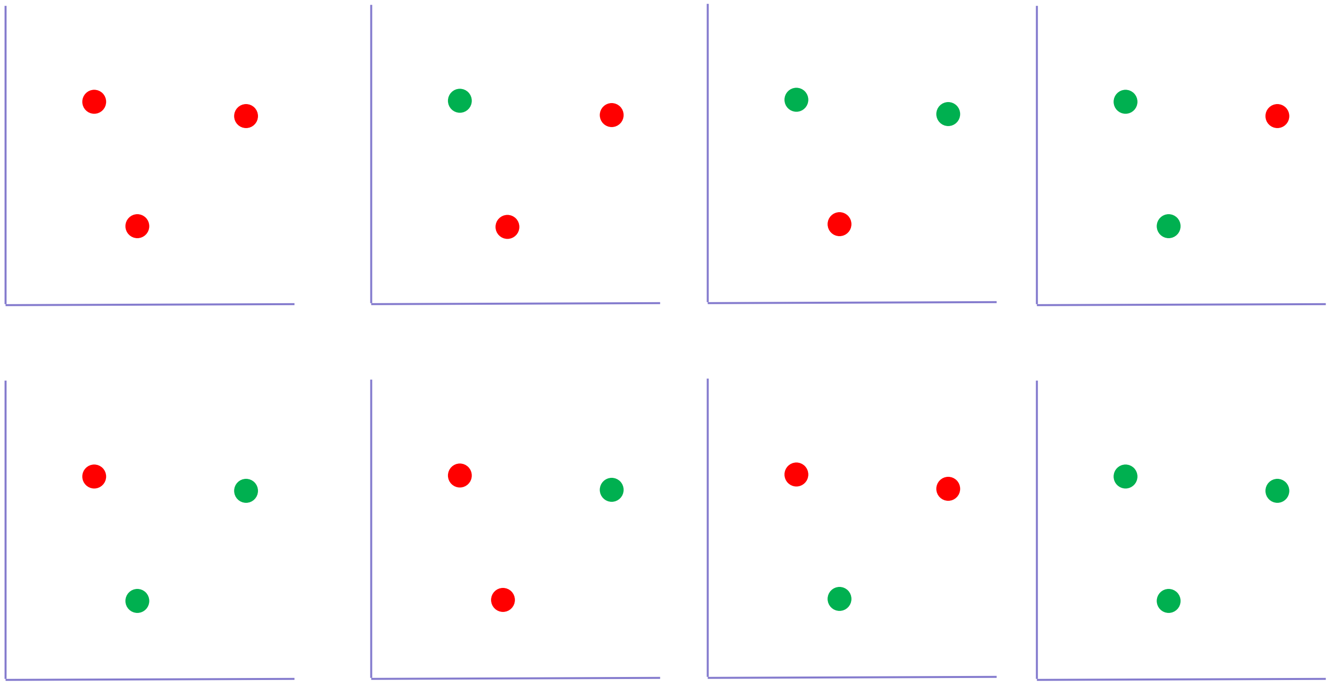
Vapnik-Chervonenkis (VC) Dimension



Two Numbers can be classified in four different ways

Vapnik-Chervonenkis (VC) Dimension

Total Data Points = 3



Three Numbers can be classified in eight different ways

Total data points N can be classified in 2^N different ways.

Vapnik-Chervonenkis (VC) Dimension

Shattering:

A hypothesis class H can shatter N data points if

- A hypothesis h∈H ->
- separates the positive examples from the negative ->
- for every problem

VC Dimension:

VC dimension of a hypothesis class H is the **maximum number of data points** which can be **shattered by H**

Vapnik-Chervonenkis dimension (VC dimension)

- Let H be the hypothesis space for some machine learning problem

- The Vapnik-Chervonenkis dimension of H
  - Also called the VC dimension of H
  - Denoted by V C(H)

- Measure of the complexity (or, capacity, expressive power, richness, or flexibility) of the space H

Example:

- Let the instance space X be the set of all real numbers
- Consider the hypothesis space defined by

$$H = \{h_m : m \text{ is a real number}\},$$

where

$$h_m \quad : \quad \text{IF } x \geq m \text{ THEN "1" ELSE "0".}$$

- Let D be a subset of X containing only a single number, say,

  $D = \{3.5\}$

- There are 2 dichotomies for this set

- These correspond to the following assignment of class labels:

x	3.25
Label	0

x	3.25
Label	1

- $h_4 \in H$ is consistent with the former dichotomy and $h_3 \in H$ is consistent with the latter.
- So, to every dichotomy in D there is a hypothesis in H consistent with the dichotomy.
- Therefore, the set D is shattered by the hypothesis space H.

x	3.25
Label	0

x	3.25
Label	1

$$H = \{h_m : m \text{ is a real number}\},$$

where

$$h_m \quad : \quad \text{IF } x \geq m \text{ THEN "1" ELSE "0"}.$$

- Let D be a subset of X containing two elements, say, D = {3.25; 4.75}

x	3.25	4.75
Label	0	0

(a)

x	3.25	4.75
Label	0	1

(b)

x	3.25	4.75
Label	1	0

(c)

x	3.25	4.75
Label	1	1

(d)

x	3.25	4.75
Label	0	0

(a)

x	3.25	4.75
Label	0	1

(b)

x	3.25	4.75
Label	1	0

(c)

x	3.25	4.75
Label	1	1

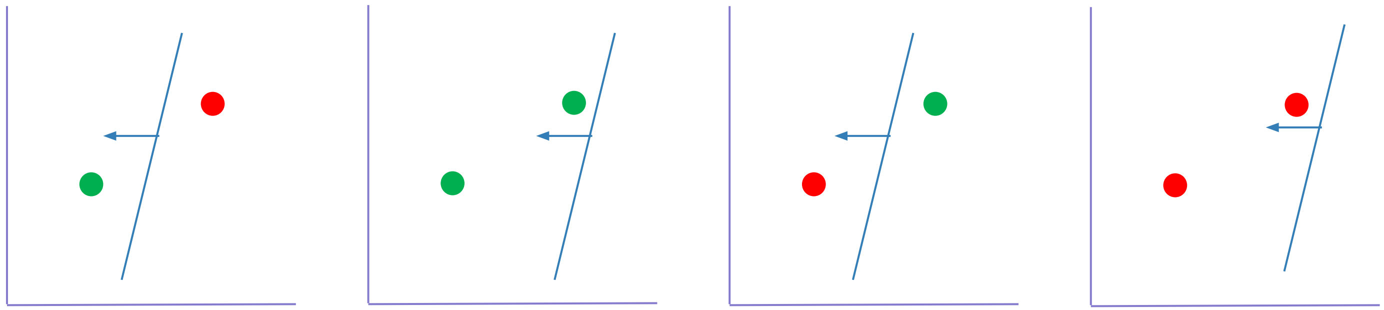
(d)

- In these dichotomies:
  - h5 is consistent with (a)
  - h4 is consistent with (b)
  - h3 is consistent with (d)
  - But there is no hypothesis hm > H consistent with (c)

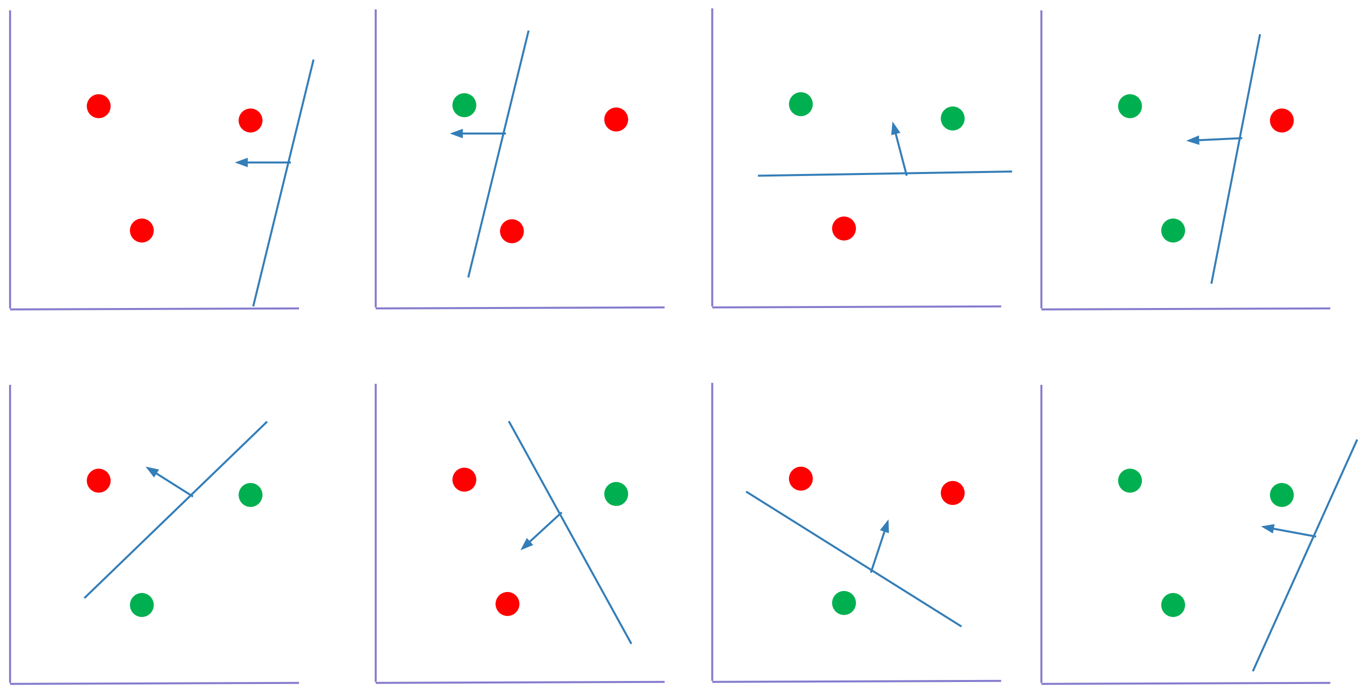
- Thus the two-element set D is not shattered by H
- The size of the largest finite subset of X shattered by H is **1**
- **This number** is the **VC dimension of H**

Vapnik-Chervonenkis (VC) Dimension

Can a line hypothesis class can shatter two data points?

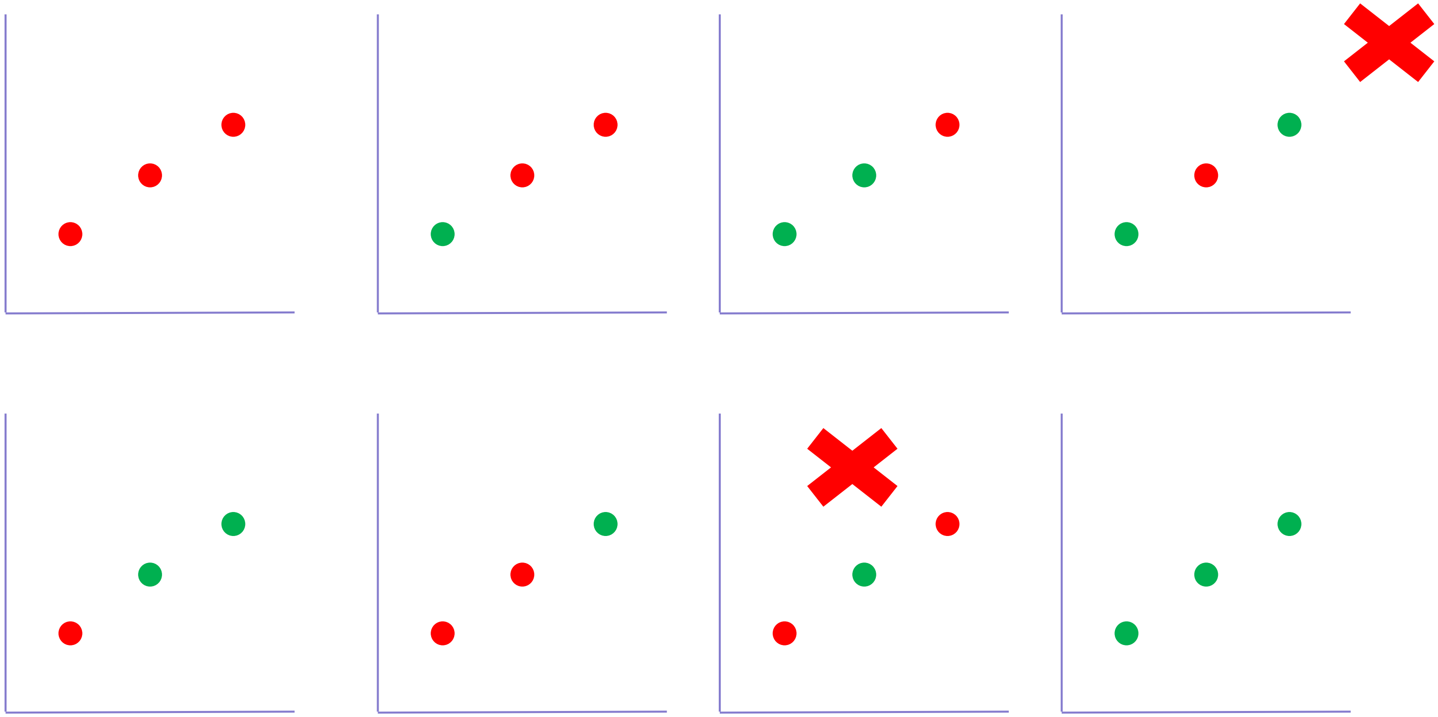


Vapnik-Chervonenkis (VC) Dimension



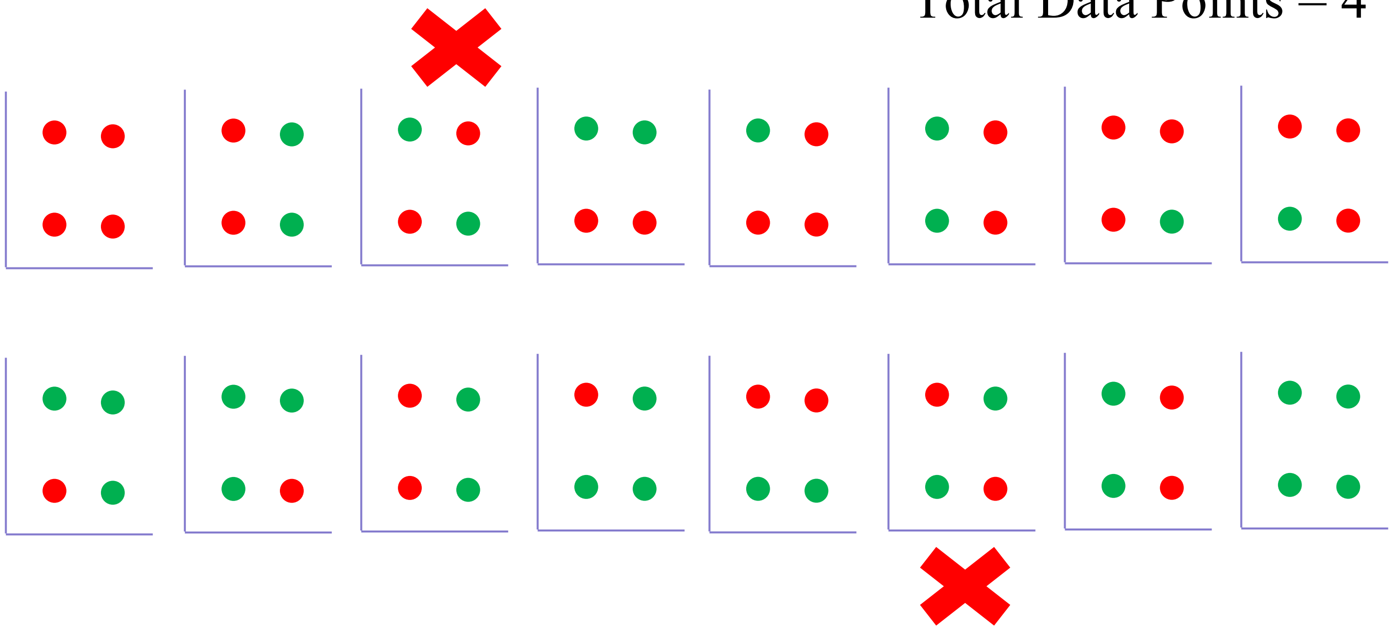
Can a line hypothesis class can shatter three data points?

Vapnik-Chervonenkis (VC) Dimension



Vapnik-Chervonenkis (VC) Dimension

Total Data Points = 4



We can’t find a dataset of 4 points which can be shattered by line class