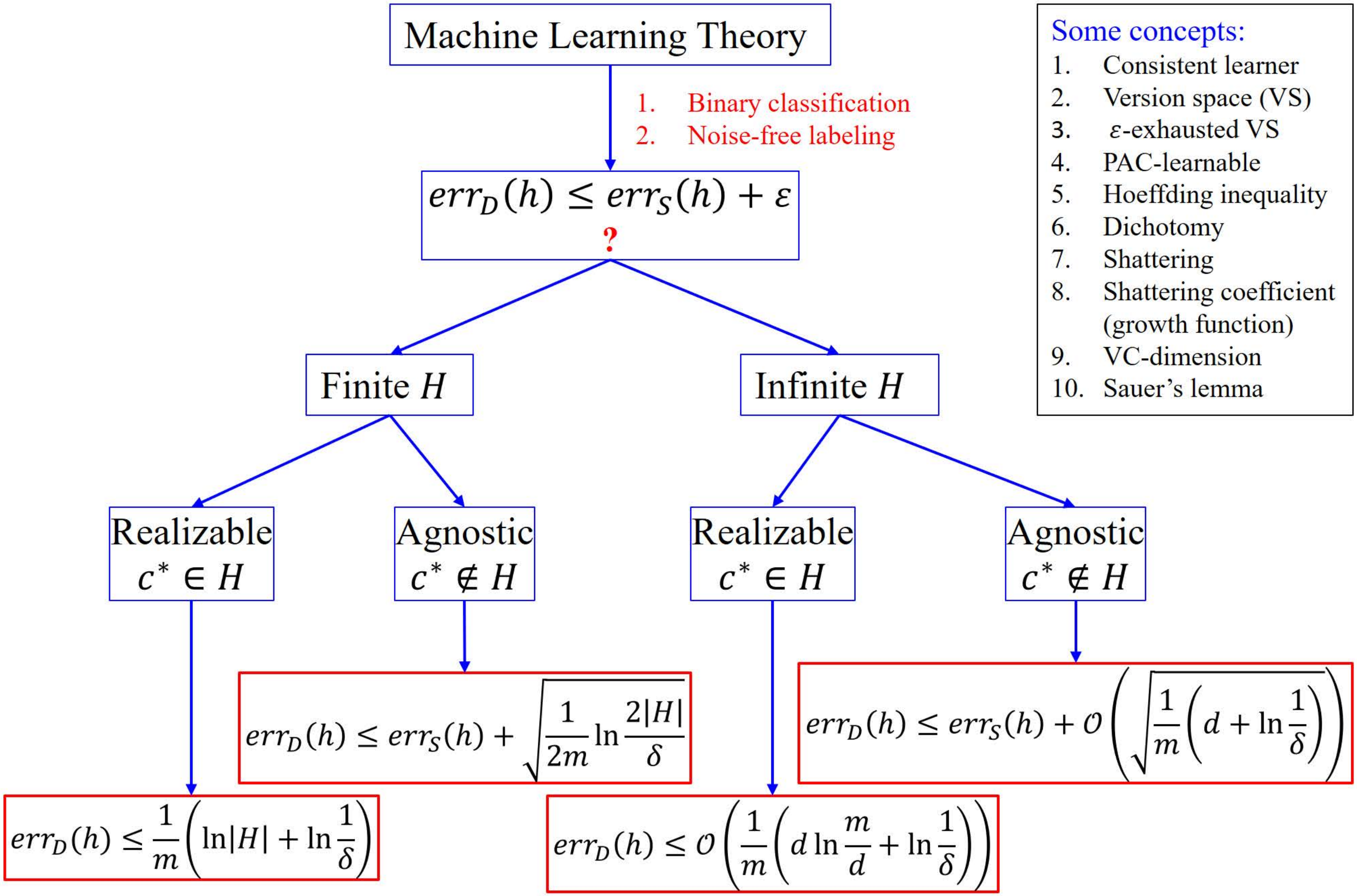
Discussion 7

Machine Learning Theory

VC dimension

田鹏超

tianpch@shanghaitech.edu.cn


Machine Learning Theory
1. Binary classification
2. Noise-free labeling
$err_D(h) \le err_S(h) + \varepsilon$
?
Finite H
Infinite H
Realizable
$c^* \in H$
Agnostic
$c^* \notin H$
Realizable
$c^* \in H$
Agnostic
$c^* \notin H$
$err_D(h) \le \frac{1}{m}\left(\ln|H| + \ln\frac{1}{\delta}\right)$
$err_D(h) \le err_S(h) + \sqrt{\frac{1}{2m}\ln\frac{2|H|}{\delta}}$
$err_D(h) \le \mathcal{O}\left(\frac{1}{m}\left(d\ln\frac{m}{d} + \ln\frac{1}{\delta}\right)\right)$
$err_D(h) \le err_S(h) + \mathcal{O}\left(\sqrt{\frac{1}{m}\left(d + \ln\frac{1}{\delta}\right)}\right)$
Some concepts:
1. Consistent learner
2. Version space (VS)
3. ε-exhausted VS
4. PAC-learnable
5. Hoeffding inequality
6. Dichotomy
7. Shattering
8. Shattering coefficient (growth function)
9. VC-dimension
10. Sauer's lemma

Effective number of hypotheses

- H[S] – the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

$\forall S \subseteq X$

Definition: H shatters S if $|H[S]| = 2^{|S|} \stackrel{\cdot}{=} 2^m$

specific

Shattering, VC-dimension

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

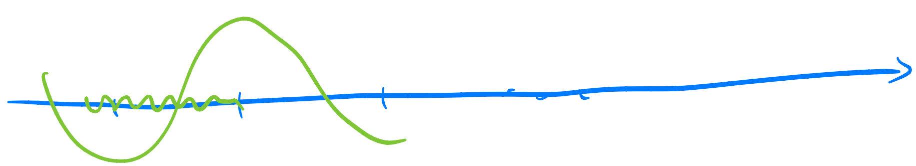
A set of points S is shattered by H is there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways, all possible ways of classifying points in S are achievable using concepts in H.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

$|S| = m$

The VC-dimension of a hypothesis space H is the cardinality of the largest set S that can be shattered by H.

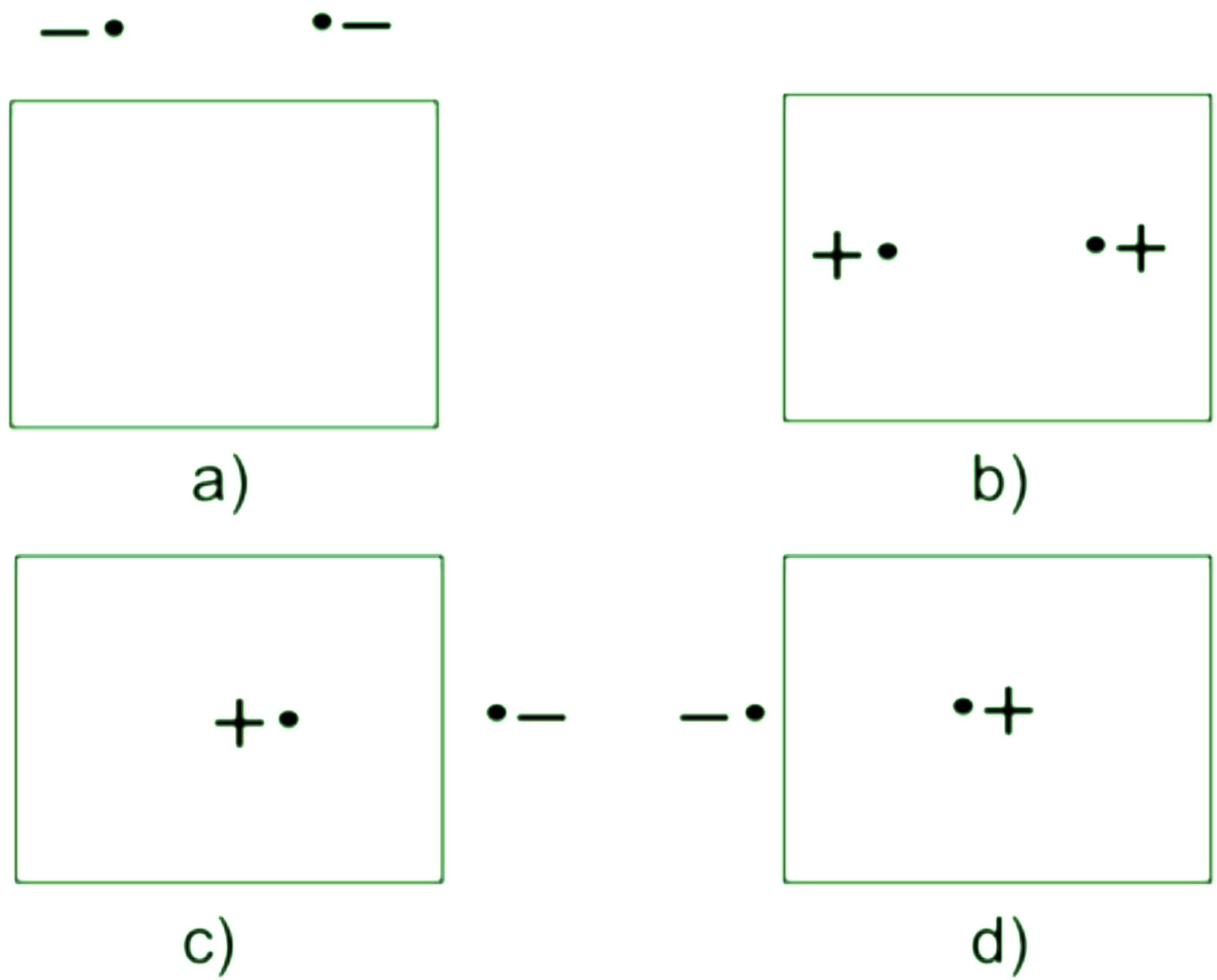
If arbitrarily large finite sets can be shattered by H, then $\mathrm{VCdim}(\mathrm{H}) = \infty$



$\sin \theta x$

VC dimension: rectangles

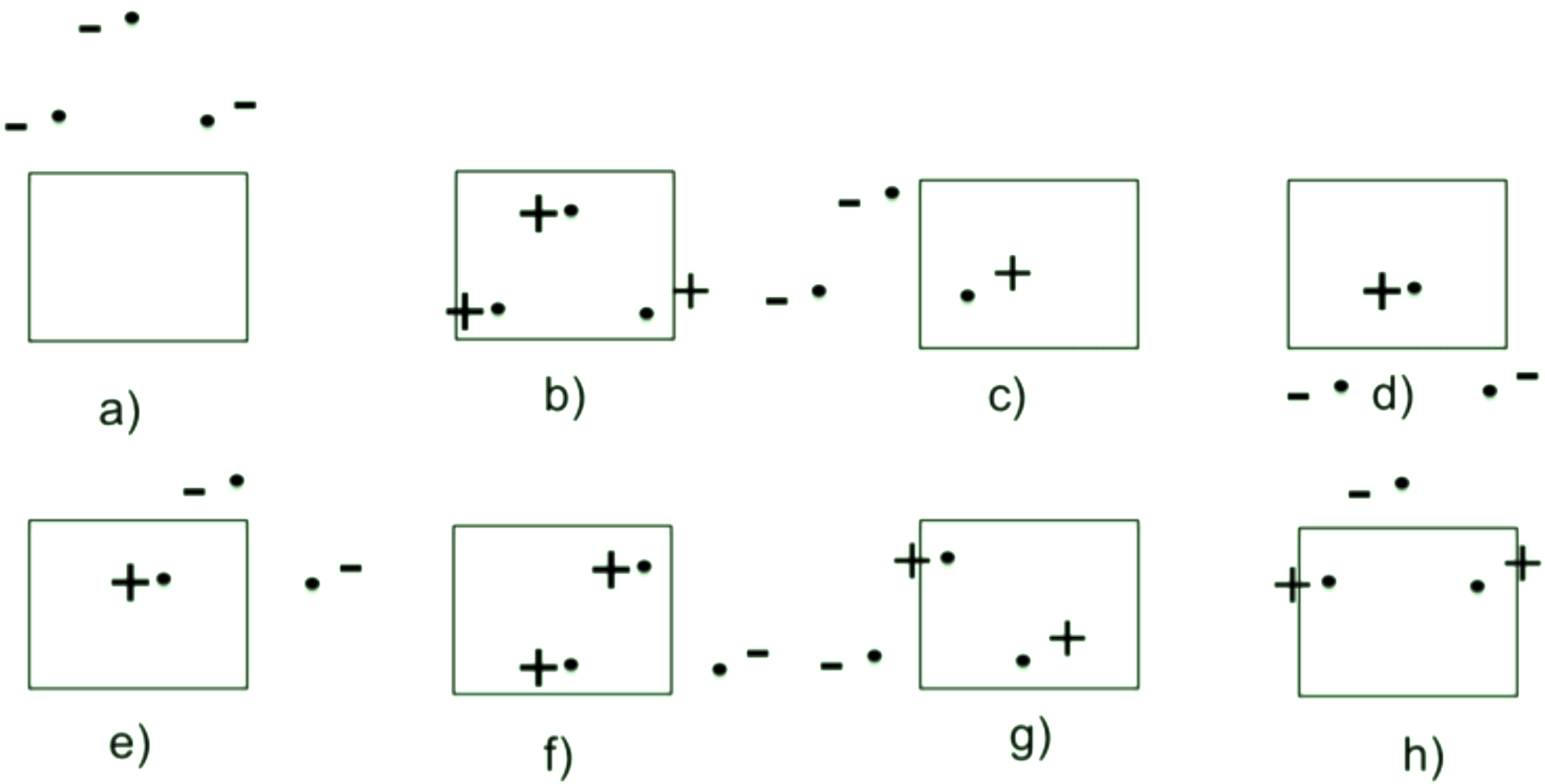
Now we look at another example, where the hypothesis labels the point inside the rectangle decided by the two points $(x_1, y_1), (x_2, y_2)$ positive, and otherwise negative. The case for two points:



A 2
B 3
C 4
D 5
E 6

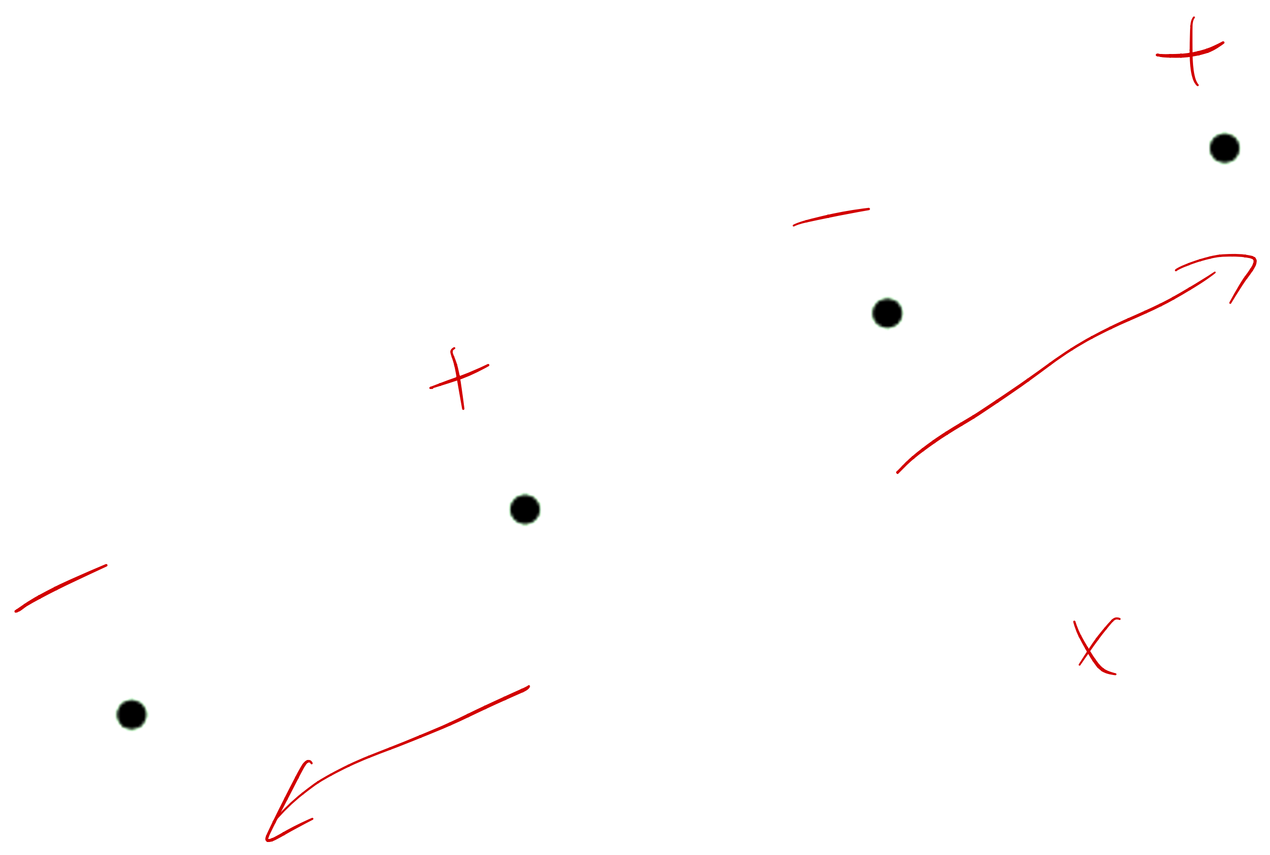
VC dimension: rectangles

For three points:



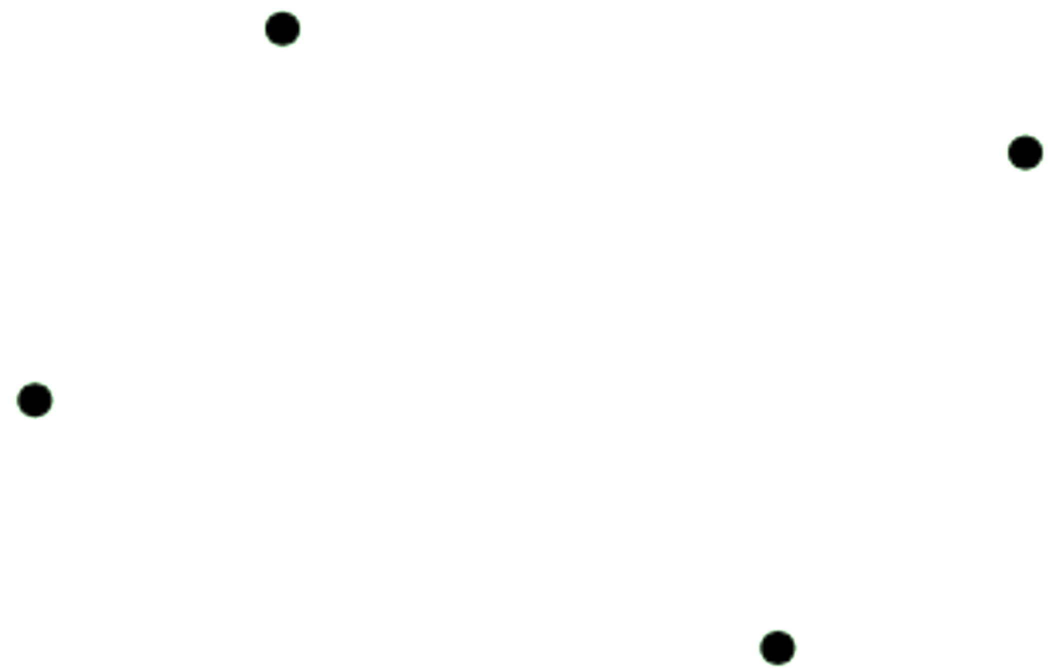
VC dimension: rectangles

The case for four points is a little different; it is not possible to produce all the dichotomies for certain situations, one of them is presented below:



VC dimension: rectangles

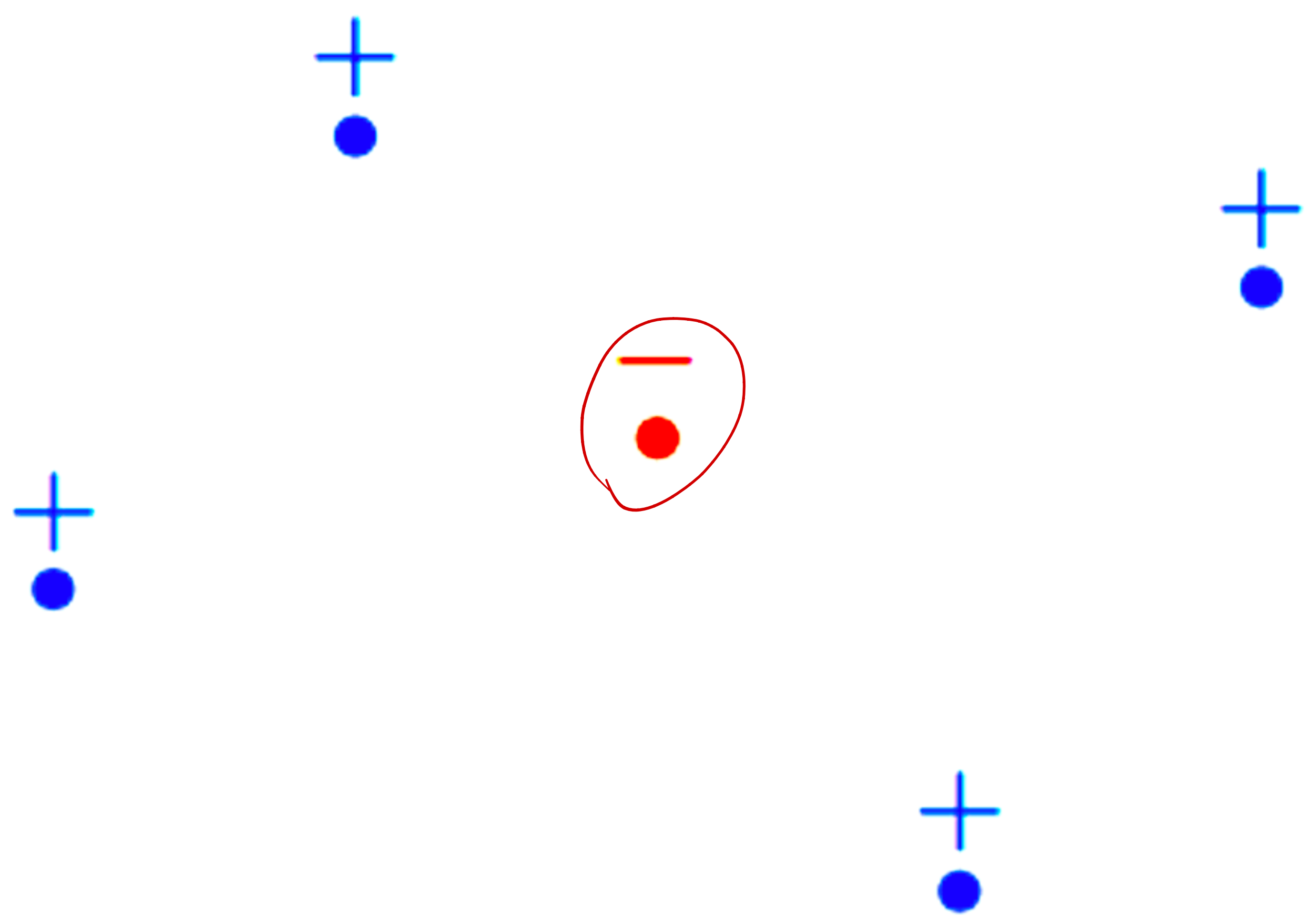
However, this configuration can be shattered!



Therefore, the VC dimension is at least 4

VC dimension: rectangles

But not this one:



VC dimension: rectangles

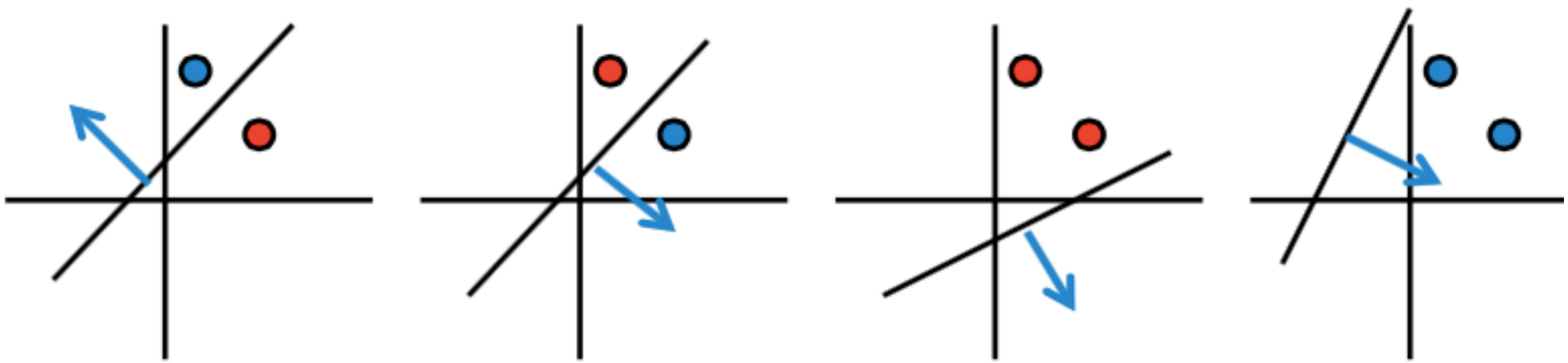
The VC dimension of rectangles is the cardinality of the maximum set of points that can be shattered by a rectangle

The VC dimension of rectangles is 4 because there exists a set of 4 points that can be shattered by a rectangle and any set of 5 points cannot be shattered by a rectangle

VC dimension: linear separator

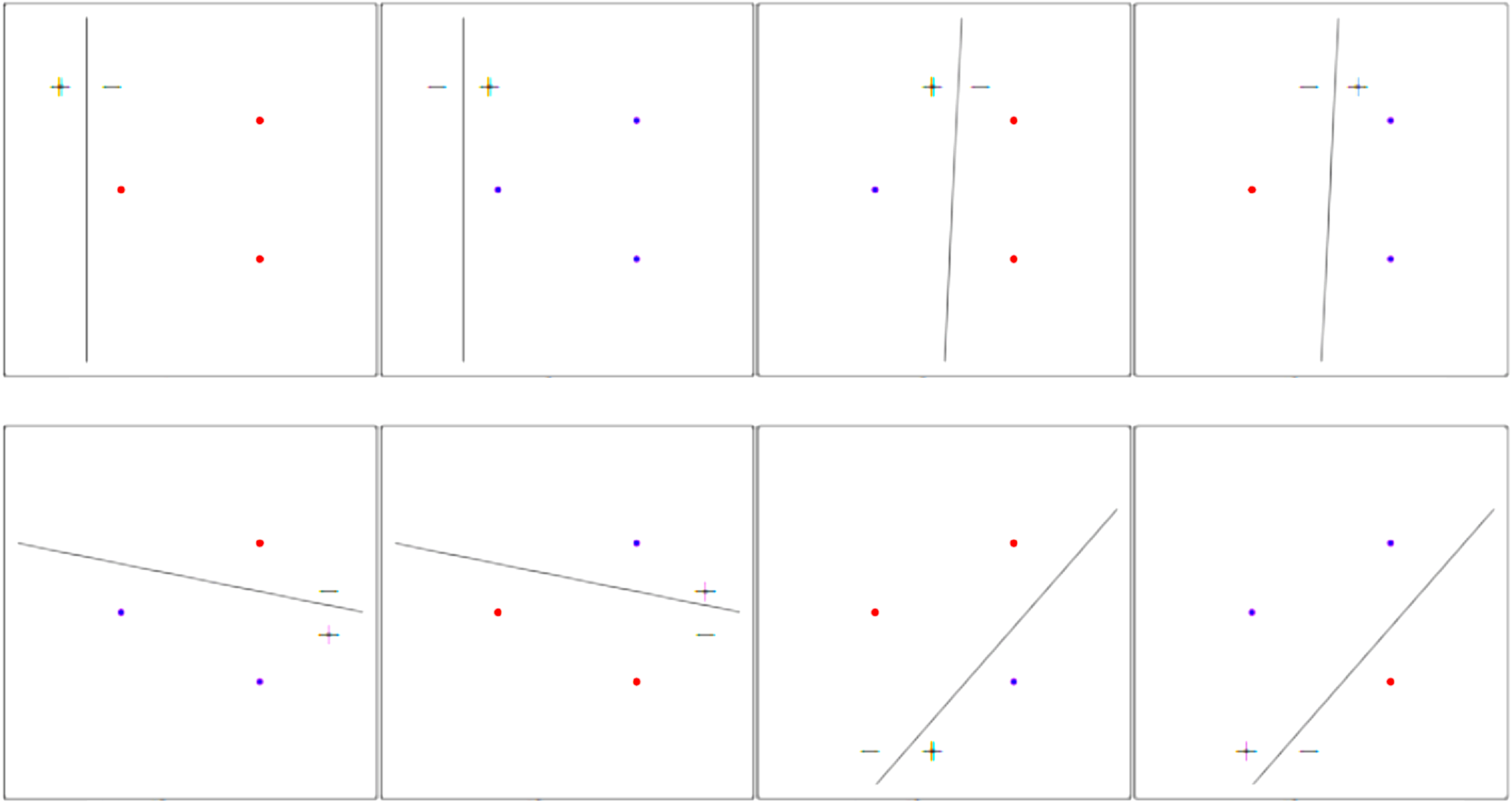
Can $h_\theta(\mathbf{x}) = \text{sign}(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$ shatter these points?

sign $(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3)$



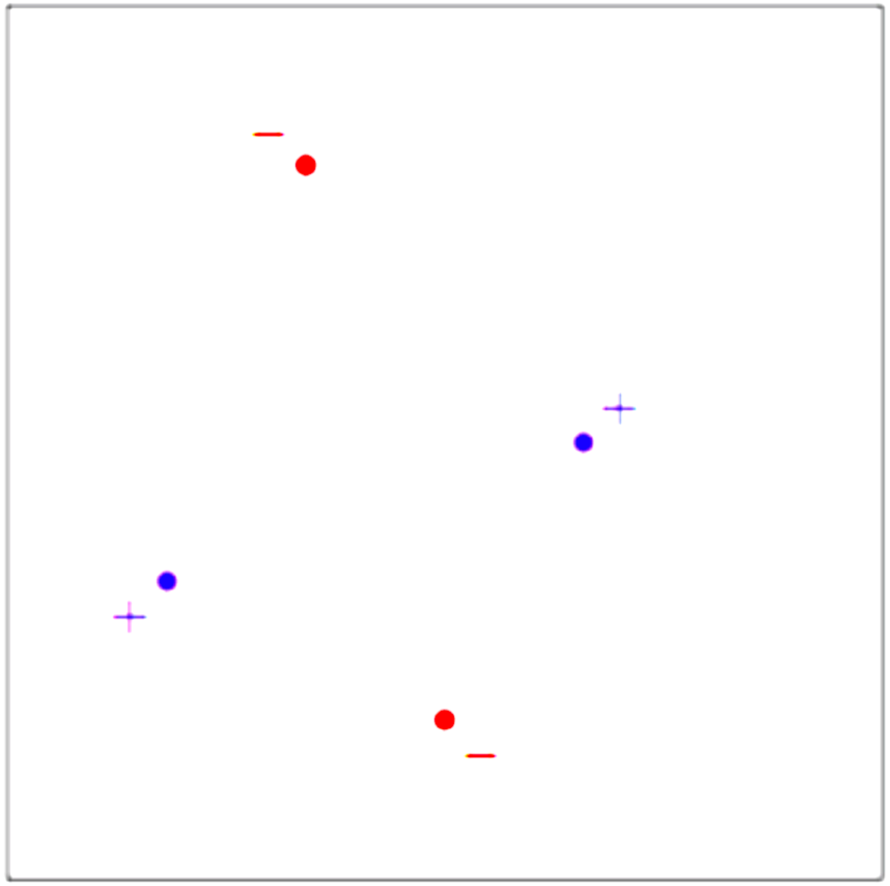
- $d_{vc} = 2?, \ d_{vc} \leq 2?, \ d_{vc} \geq 2?$

VC dimension: linear separator



VC dimension

However, things are a little different with the case of 4 points. For the case of 4 points, there are $2^4 - 2 = 14$ kinds of labeling. As the usual 2^m number of labelings, this time there are two labeling that is not achievable by linear classifiers. Below presents one of them:



- In $\mathbb{R}^2$, linear separater has $d_{vc} = 3$

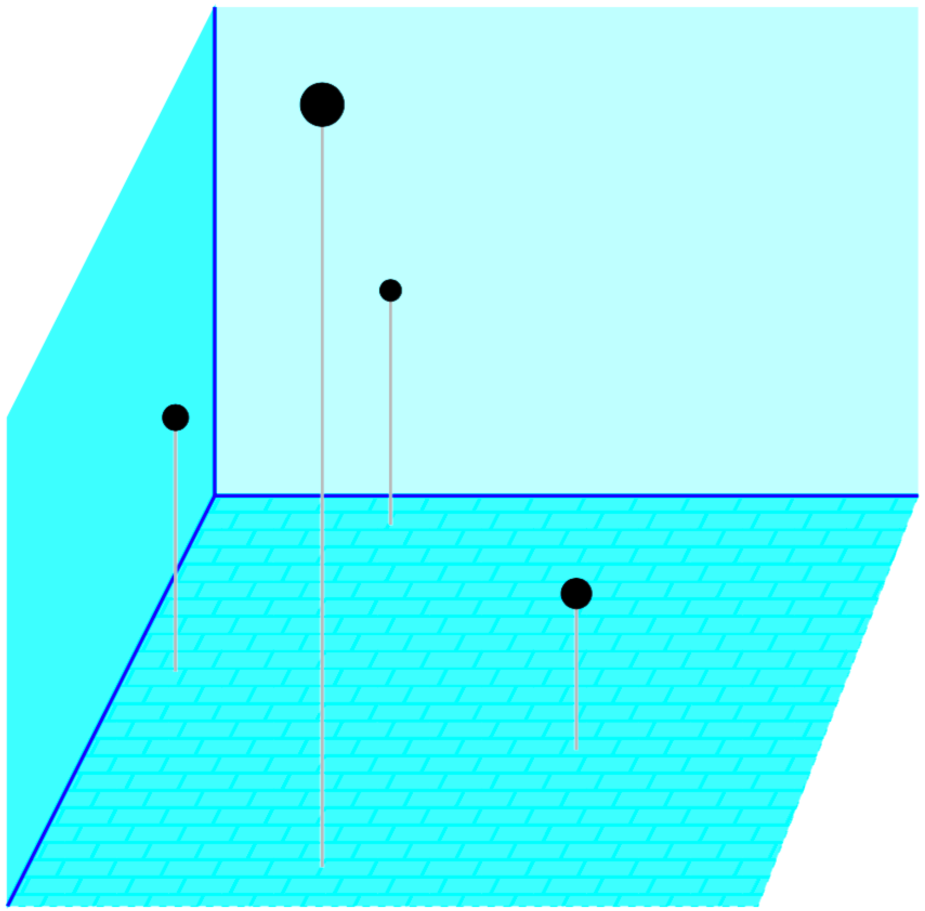
VC dimension

In general, linear classifier (perceptron) in d dimensions with a constant term

$$d_{vc} = d + 1$$

For $d = 2$, $d_{\text{VC}} = 3$

In general, $\boxed{d_{\text{VC}} = d + 1}$



We will prove two directions:

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

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

数据是在d维空间里的，但是分离平面的参数要加上常数项，一共是d+1个参数。sign$(w_0 \cdot 1 + w_1 x_1 + \ldots + w_d x_d)$

Here is one direction

A set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the perceptron:

$$\boldsymbol{x} = [1, x_1, x_2, \ldots, x_d]^T$$

$$\mathrm{X} = \begin{bmatrix} — \mathbf{x}_1^\intercal — \\ — \mathbf{x}_2^\intercal — \\ — \mathbf{x}_3^\intercal — \\ \vdots \\ — \mathbf{x}_{d+1}^\intercal — \end{bmatrix} = \underbrace{\left.\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & 0 \\ 1 & 0 & \ldots & 0 & 1 \end{bmatrix}\right\} d+1}_{d+1}$$

X is invertible

Can we shatter this data set?

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm1 \\ \pm1 \\ \vdots \\ \pm1 \end{bmatrix}$, can we find a vector $\mathbf{w}$ satisfying

$$\text{sign}(\mathrm{X}\mathbf{w}) = \mathbf{y}$$

Easy! Just make $\mathrm{X}\mathbf{w} = \mathbf{y}$

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

我们对于一个特定的包含$d+1$个数据点的数据集，可以产生所有的2^d个dichotomies。这意味着我们可以“粉碎”某个$d+1$样本容量的数据集。所以“断点”肯定不是$d+1$。

We can shatter these $d+1$ points

This implies what?

[a] $d_{\text{VC}} = d+1$

[b] $d_{\text{VC}} \geq d+1$ ✓

[c] $d_{\text{VC}} \leq d+1$

[d] No conclusion

Now, to show that $d_{vc} \leq d + 1$

We need to show that:

[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 ✓

Prove for "ANY"!!!

Here is the other direction

Take any $d+2$ points in $\mathbb{R}^d$!!

$X_i = [1, x_1, x_2, \ldots, x_d]^T$

For any $d+2$ points in $\mathbb{R}^d$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{d+1}, \boldsymbol{x}_{d+2}$ $(d+2) \times (d+1)$

More points than dimensions $\Longrightarrow$ we must have

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i$$

where not all the a_is are zeros

Our purpose is then to design a dichotomy that any linear separator cannot generate on $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{d+1}, \boldsymbol{x}_{d+2}$!!

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i$$

- Consider the following dichotomy:

$$y_i = \text{sign}(a_i) \qquad \text{for } \boldsymbol{x}_i\text{'s with non-zero } a_i$$
$$y_j = -1 \qquad \text{for } \boldsymbol{x}_j$$

- No perceptron can implement such dichotomy!

- The dichotomy we construct ($j = 1$)

$$\boldsymbol{x}_j \qquad \boldsymbol{x}_1 = a_2\boldsymbol{x}_2 + \ldots + 0 \cdot \boldsymbol{x}_{d+1} + a_{d+2}\boldsymbol{x}_{d+2}$$

$$y_j \qquad -1 \qquad \text{sign}(a_2) \qquad \text{whatever} \qquad \text{sign}(a_{d+2})$$

- Show for any $\boldsymbol{w} \in \mathbb{R}^{d+1}$, this dichotomy cannot appear!

- The dichotomy we construct ($j = 1$)

$$\boldsymbol{x}_j \quad \boldsymbol{x}_1 = 1 \cdot \boldsymbol{x}_2 + \ldots + 0 \cdot \boldsymbol{x}_{d+1} - 1 \cdot \boldsymbol{x}_{d+2}$$

-1

$$y_j \quad -1 \quad +1 \quad +1 \quad -1$$

$y_i = \text{sign}(w^T x_i)$

- Show for any $\boldsymbol{w} \in \mathbb{R}^{d+1}$, this dichotomy cannot appear!
- Notice that $y_i = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_i)$

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i \implies \underbrace{\boldsymbol{w}^T \boldsymbol{x}_j}_{y_j} = \underbrace{\sum_{i \neq j} a_i \boxed{\boldsymbol{w}^T \boldsymbol{x}_i}}_{>0}$$

这项的sign就是y_i，假设我们可以选$\boldsymbol{w}$使得这项的sign和y_i匹配。则由于我们在设定dichotomy的时候，把y_i选的和a_i的sign选的一样！！到处矛盾！

- Since $\text{sign}(\boldsymbol{w}^T \boldsymbol{x}_i) = y_i = \text{sign}(a_i)$, then $a_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$
- This forces $\boldsymbol{w}^T \boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$
- Therefore, $\boldsymbol{y}_j = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_j) = +1$, contradiction!!!

Putting it together...

We proved $d_{vc} \leq d + 1$ and $d_{vc} \geq d + 1$

$$\boxed{d_{vc} = d + 1}$$