Machine Learning

Lecture 8: Theory of Generalization (II):
VC dimension

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

本节内容

- Vapnik-Chervonenkis (VC) dimension
- Examples: determining the VC dimension
- VC Inequality
- Structural Risk Minimization
- Sample Complexity

(Left to right) J. Rissanen, V. Vapnik, A. Gammerman, A. Chervonenkis, C. Wallace and R. Solomonoff

We have...

$$\mathbb{P}[\,|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

If we choose tolerance level δ (e.g. $\delta = 0.05$), meaning we want the probability no more than δ

$$2Me^{-2\epsilon^2 N} \leq \delta \implies \epsilon \geq \sqrt{\frac{1}{2N}\ln\left(\frac{2M}{\delta}\right)}$$

We then know with probability at least $1 - \delta$, we have

$$|E_{in}(g) - E_{out}(g)| \leq \sqrt{\frac{1}{2N}\ln\left(\frac{2M}{\delta}\right)}$$

or, equivalently,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N}\ln\left(\frac{2M}{\delta}\right)}$$

What we have so far…

- We define the different behaviors that a hypothesis class has on a specific set of N points $\mathcal{S} = \{x_1, \ldots, x_N \in \mathbb{R}^d\}$ as the ***dichotomies***: $\mathscr{H}(\mathcal{S})$

- The number of different behaviors is then: $|\mathscr{H}(\mathcal{S})|$

- The maximum of $|\mathscr{H}(\mathcal{S})|$ on any N points $\mathcal{S} = \{x_1, \ldots, x_N \in \mathbb{R}^d\}$ is defined as the ***growth function:*** $m_{\mathscr{H}}(N)$

- Clearly, in given $\mathbb{R}^d$ space, $m_{\mathscr{H}}(N)$ is dependent of $\mathscr{H}$ and N

- If a $\mathscr{H}$ satisfying $m_{\mathscr{H}}(N) < 2^N$ for any $N \geq k$ (k is defined as the ***break point***) we define $m_{\mathscr{H}}(N) \leq B(k, N)$ for any such $\mathscr{H}$

- We show that $B(k, N)$ is bounded by a polynomial of N

We are about to answer two questions:

I. What kind of $\mathscr{H}$ has such a break point k?

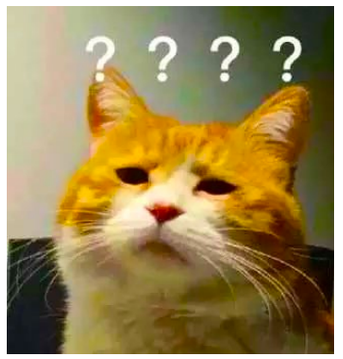
II. If such a k exists, what is our generalization inequality?

Can we simply replace M in the finite case

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N}\ln\left(\frac{2M}{\delta}\right)}$$

with $m_{\mathscr{H}}(N)$ to have

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N}\ln\left(\frac{2m_{\mathscr{H}}(N)}{\delta}\right)}$$



Vapnik-Chervonenkis (VC) dimension

Definition. VC-dimension of $\mathscr{H}$ = maximal number of points N such that $m_{\mathscr{H}}(d_{vc}) = 2^{d_{vc}}$.

- The VC-dimension of a hypothesis class is the maximal number of points for which you can get all possible behaviors

$$d_{vc} = \arg\max_{N} \{N \mid m_{\mathscr{H}}(N) = 2^N\}$$

- In terms of the number of dichotomies, we have

$$d_{vc} = \arg\max_{N} \{N \mid \mathscr{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = 2^N\}$$

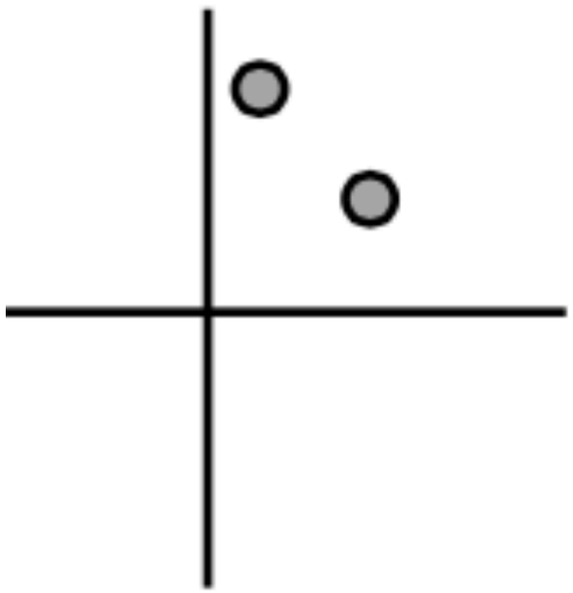
- In terms of break point

$$d_{vc} = k - 1$$

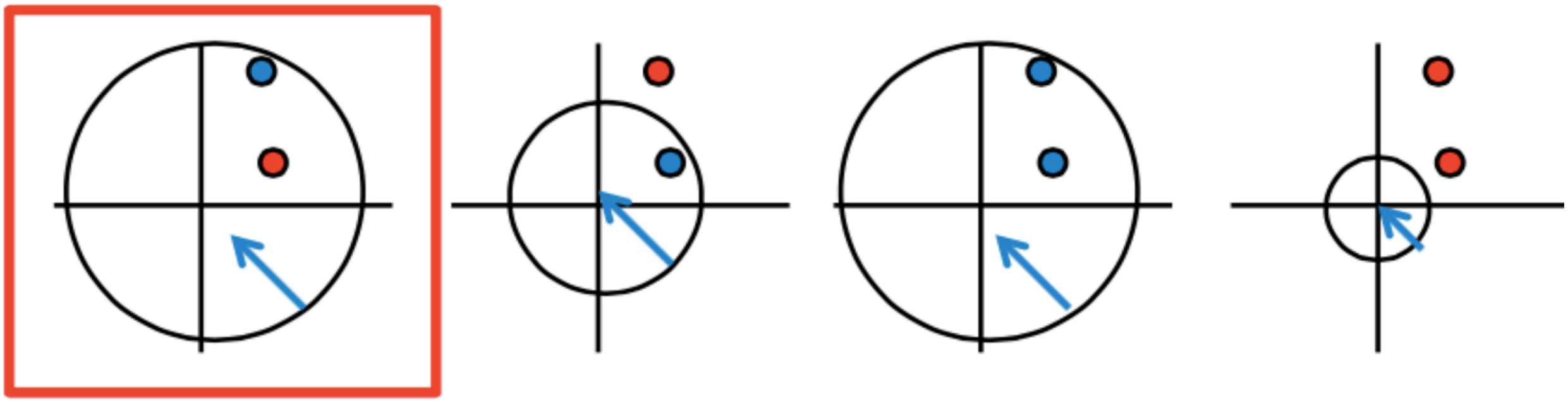
- $\mathscr{H}$ is able to achieve any labeling of some given set of d_{vc} points, meaning it can shatter some given set of d_{vc} points

VC-dimension: example

Can $h_\theta(\mathbf{x}) = \text{sign}(\mathbf{x}^T\mathbf{x} + \theta)$ shatter these points?



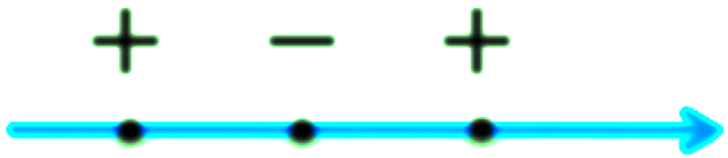
- $d_{vc} = ?$



VC dimension: intervals

- Consider the hypotheses $h = \{x \in [a, b]\}$, which labels x positive if x is in the interval $[a, b]$, negative otherwise
- For the case of two points, we have two dichotomies

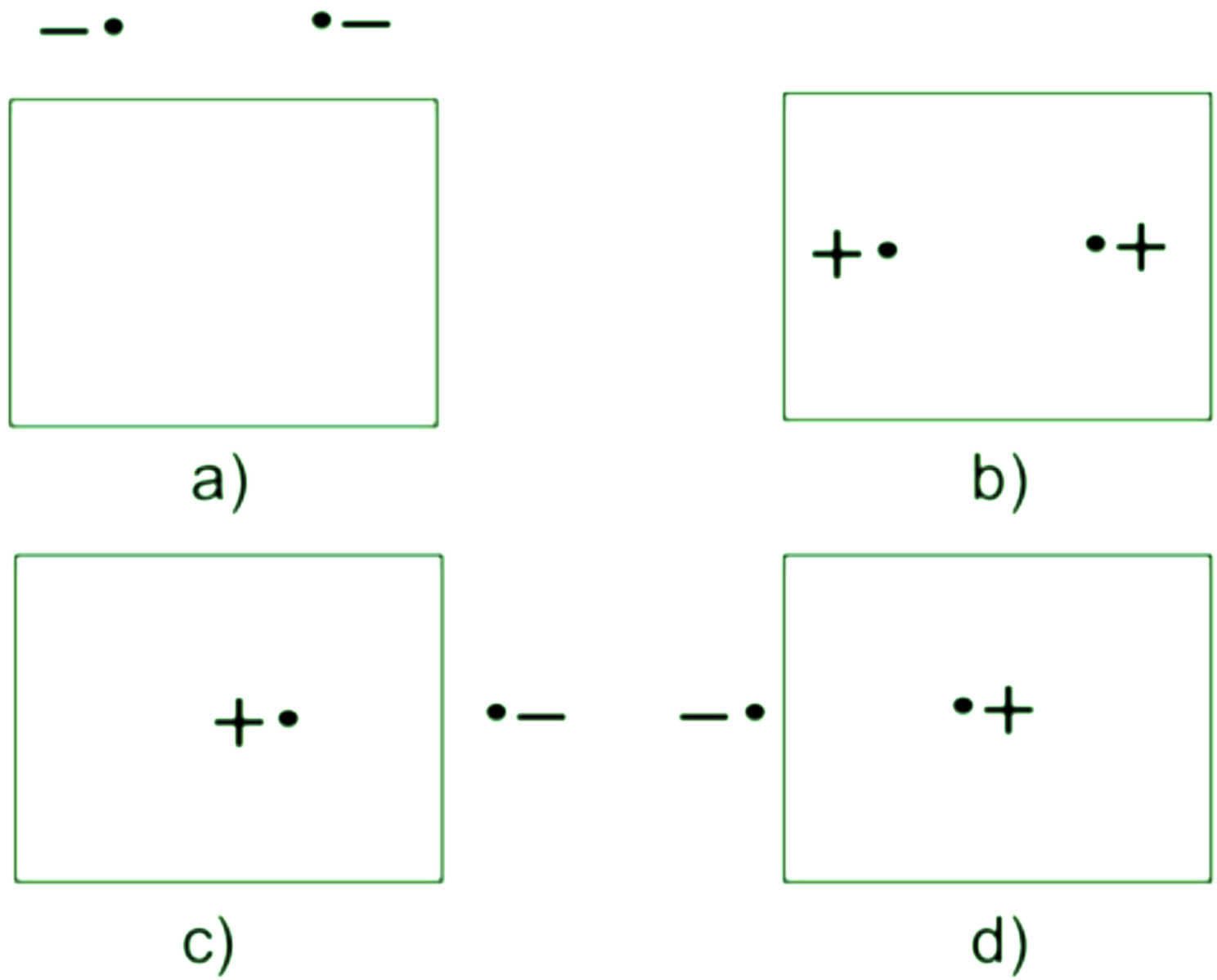
- However, for the case of three points, there exists some dichotomies that cannot be generated, one case is presented below:



- $d_{vc} = 2$

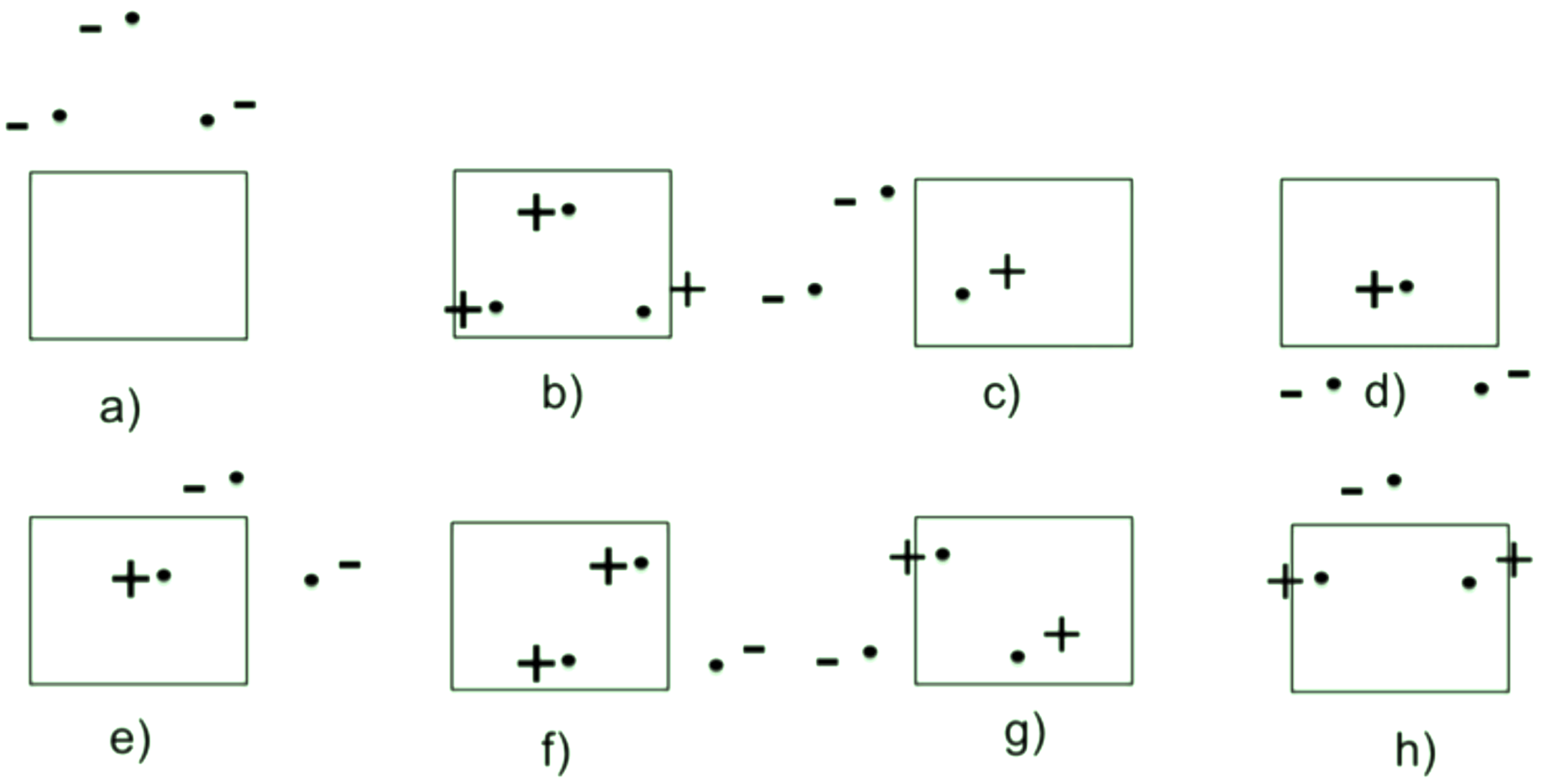
VC dimension: rectangles

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:



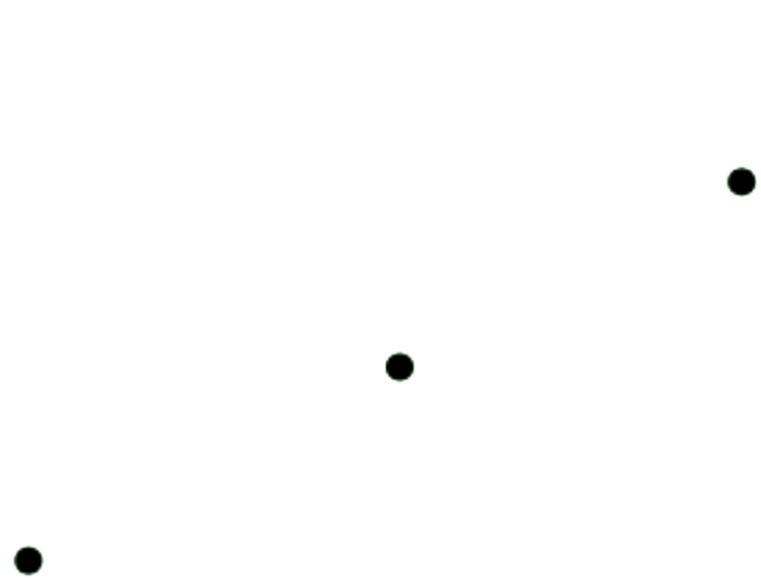
VC dimension: rectangles

For three points:



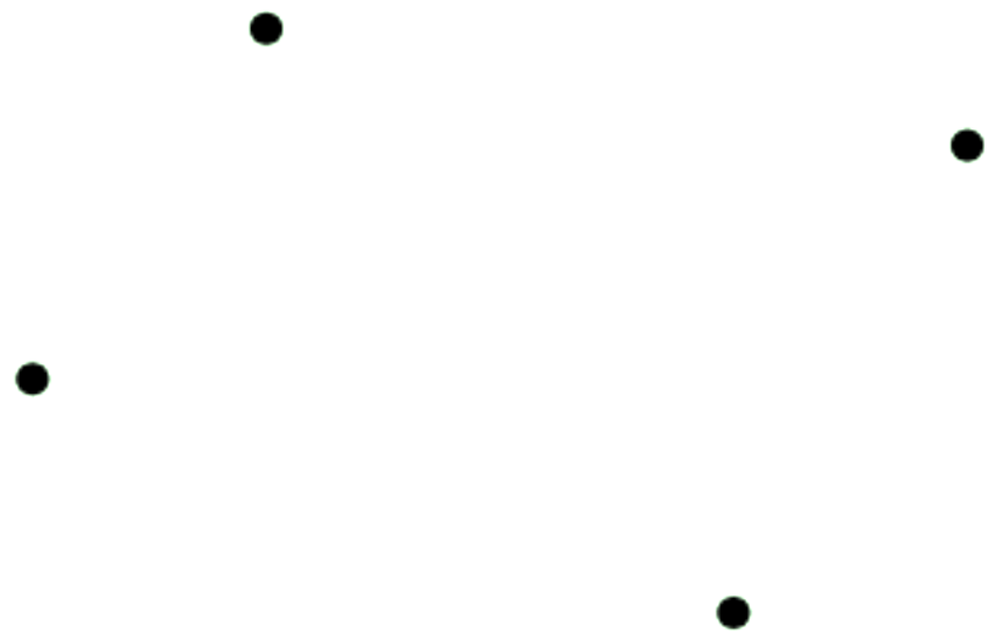
VC dimension: rectangles

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

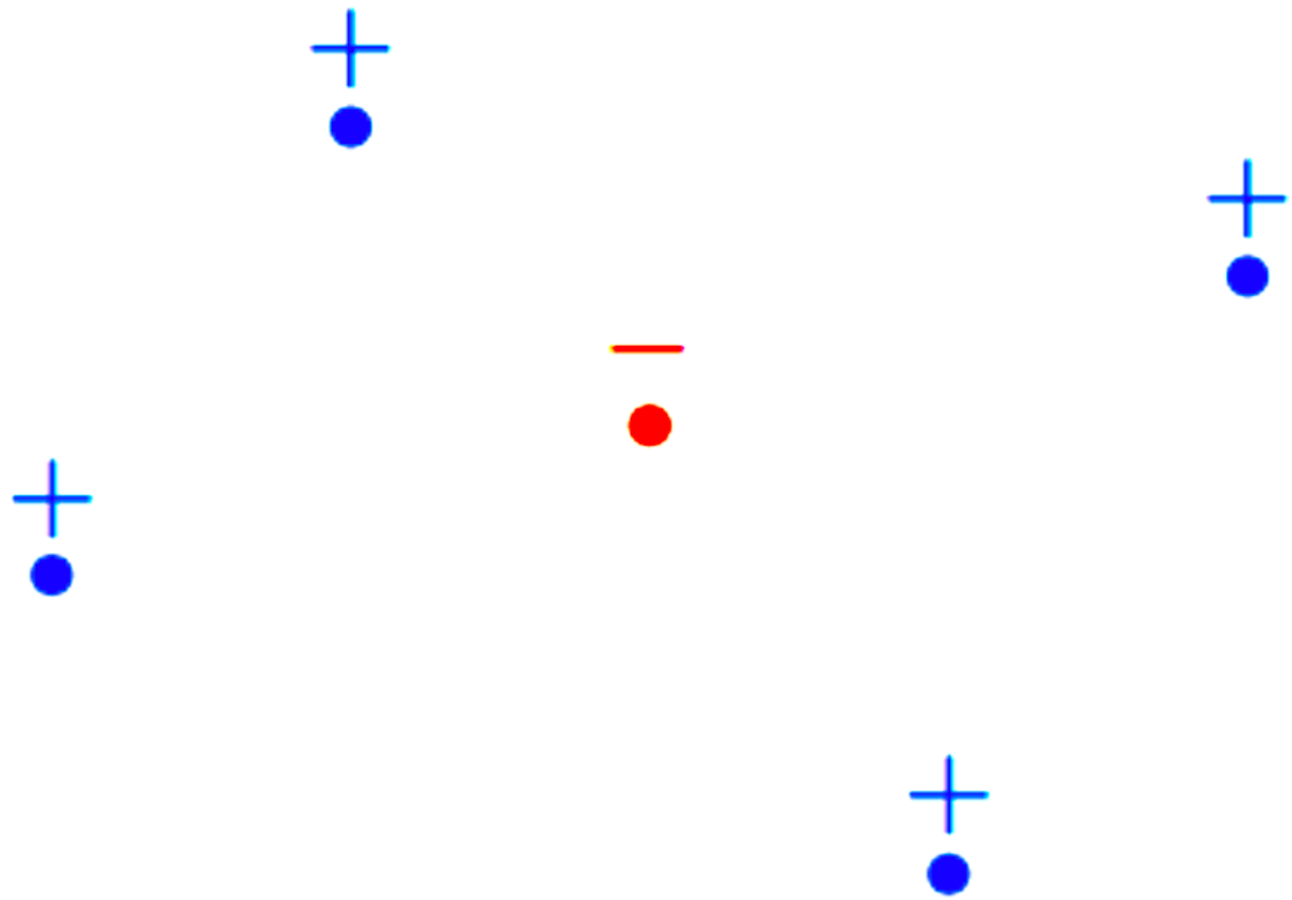
However, this configuration can be shattered!



Therefore, the VC dimension is at least 4

VC dimension: rectangles

But not this one:



VC dimension: rectangles

From the previous examples, we see that:

- For $N = 1, m_{\mathscr{H}}(N) = 2^1 = 2 = 2^N$
- For $N = 2, m_{\mathscr{H}}(N) = 2^2 = 4 = 2^N$
- For $N = 3, m_{\mathscr{H}}(N) = 2^3 = 8 = 2^N$
- For $N = 4, m_{\mathscr{H}}(N) = 2^4 = 16 = 2^N$
- For $N = 5, m_{\mathscr{H}}(N) < 2^5$

Therefore, VC-dimension of $\mathscr{H}$ = 4; break point = 5

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

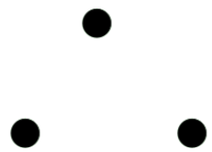
Examples

- $\mathcal{H}$ is positive rays:

$$d_{\mathrm{VC}} = 1$$

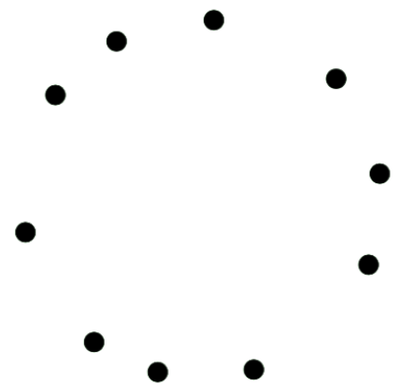
- $\mathcal{H}$ is 2D perceptrons:

$$d_{\mathrm{VC}} = 3$$



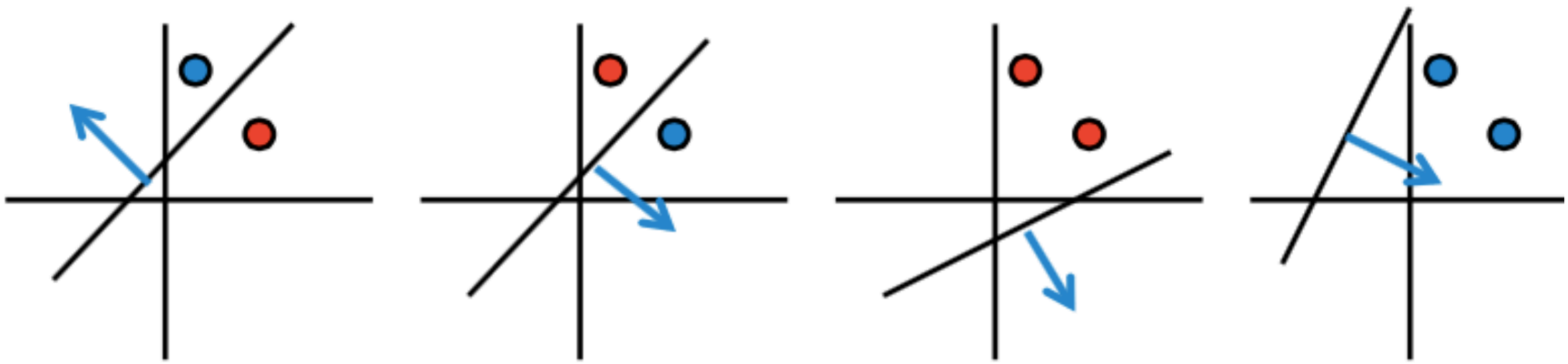
- $\mathcal{H}$ is convex sets:

$$d_{\mathrm{VC}} = \infty$$



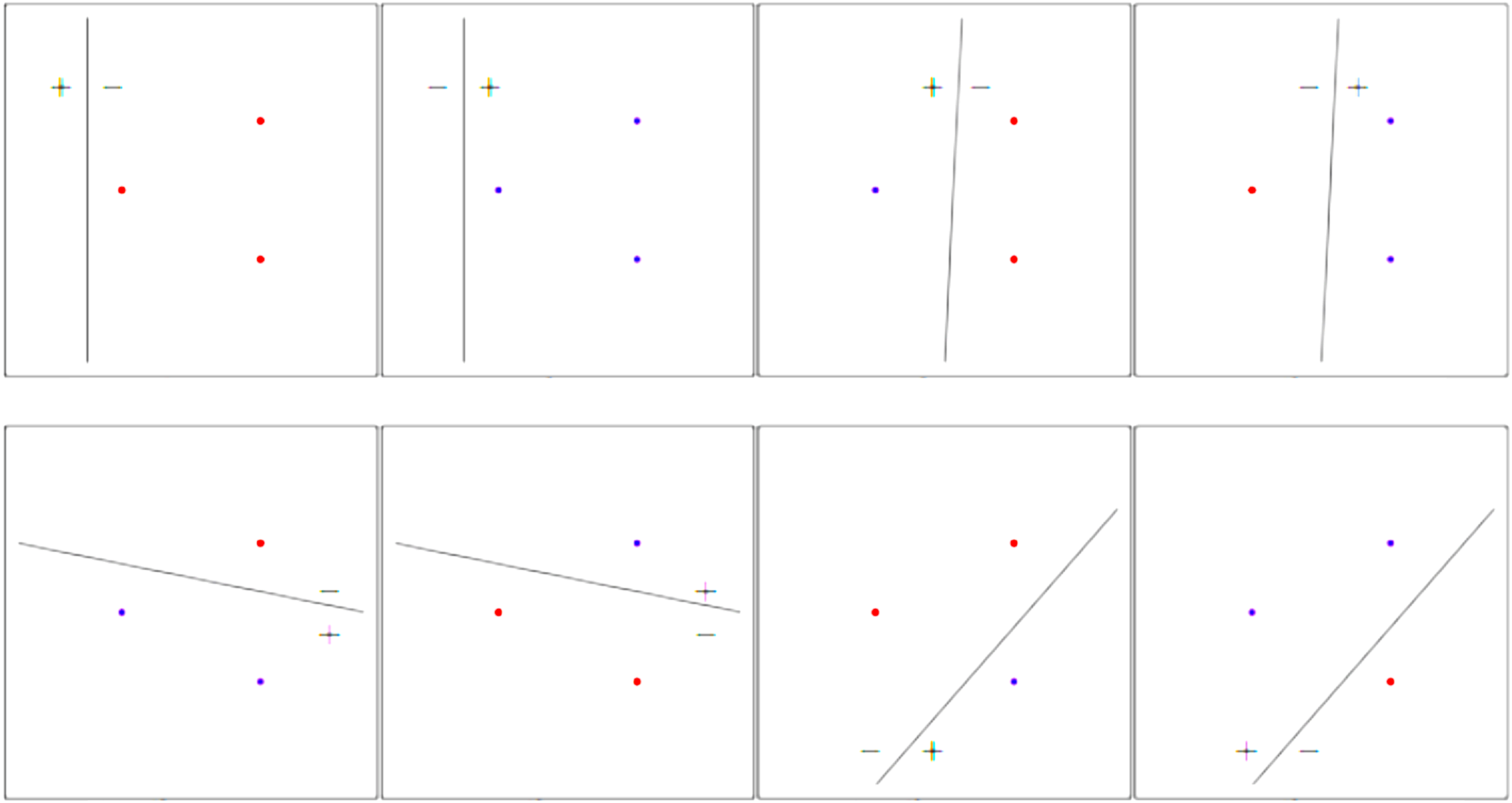
VC dimension: linear separator

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



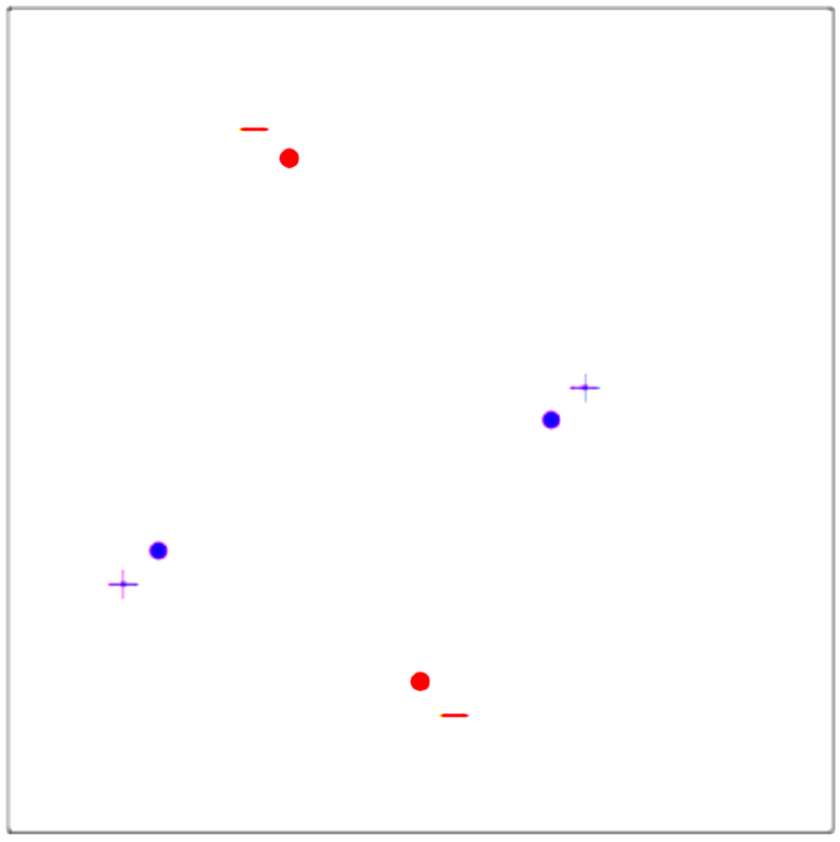
- $d_{vc} = 2?,\ \ d_{vc} \leq 2?,\ \ \ d_{vc} \geq 2?$

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 $d_{vc} = 3$

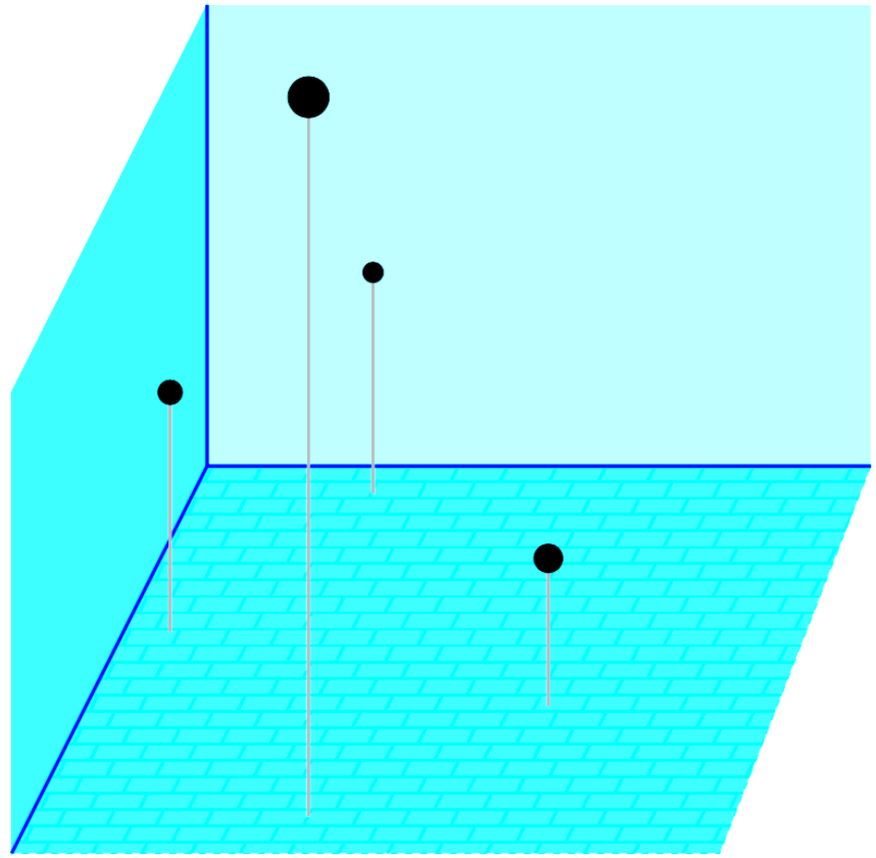
VC dimension

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:

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

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

X is invertible

Can we shatter this data set?

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, 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+1}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 $\mathbb{R}^d$!!

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

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 ($j = 1$)

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

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

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

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

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

y_i $\quad -1 \quad +1 \qquad\qquad +1 \qquad\qquad -1$

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

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i \implies \boldsymbol{w}^T \boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{w}^T \boldsymbol{x}_i$$

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！！导致矛盾！

- 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, $\boldsymbol{y}_i = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_j) = +1$, contradiction!!!

Putting it together…

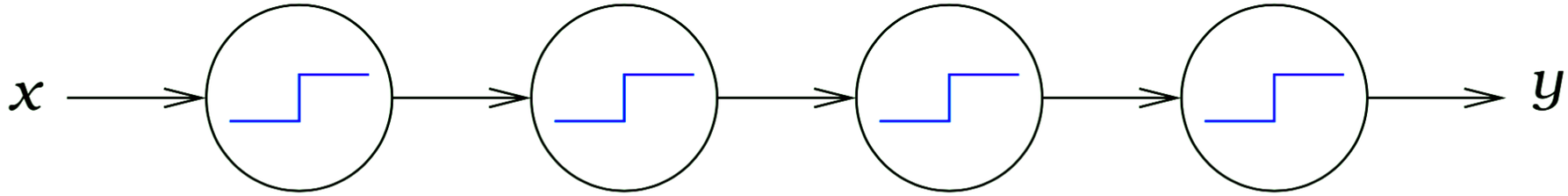
We proved $d_{vc} \leq d + 1$ and $d_{vc} \geq d + 1$

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

Number of Parameters is NOT equivalent to VC dimension

- Parameters create degree of freedom
- # of parameters: analog degrees of freedom
- Parameters may not contribute VC dimension

每一个都是sign($w_1x + w_0$)。但VC维度依然是2。后面的每次复合只是对{+1, − 1}进行变换。



我们依然可以找到1维数据，VC维度 = ∞的例子。

- Few parameters may lead to large VC dimension
- d_{vc} measures the effective number of parameters

The growth function

- The largest value of N for which $m_{\mathscr{H}}(N) = 2^N$
- "The most points $\mathscr{H}$ can shatter"

$$N \leq d_{\text{VC}}(\mathcal{H}) \implies \mathcal{H} \text{ can shatter } N \text{ points}$$

$$k > d_{\text{VC}}(\mathcal{H}) \implies k \text{ is a break point for } \mathcal{H}$$

We have proved:

- In terms of a break point k

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

- In terms of the VC dimension d_{vc}

$$m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{d_{\text{VC}}} \binom{N}{i}}_{\text{maximum power is } N^{d_{\text{VC}}}}$$

最高次幂就是$N^{d_{vc}}$，之后随着N增加，这个幂是定死的，不会增加的。

The growth function and VC dimension

Problem 2.5 Prove by induction that $\sum_{i=0}^{D} \binom{N}{i} \leq N^D + 1$, hence

$$m_{\mathcal{H}}(N) \leq N^{d_{\text{vc}}} + 1.$$

$D = 0$, $\sum_{i=0}^{D} \binom{N}{i} = 1$, $N^D + 1 = 1 + 1 = 2$, the statement is true.

Assume the inequality holds true at $D = k$. For $D = k + 1$, we have

$$\sum_{i=0}^{k+1} \binom{N}{i} = \sum_{i=0}^{k} \binom{N}{i} + \binom{N}{k+1} \leq N^k + 1 + \binom{N}{k+1}$$

On the other hand,

$$\binom{N}{k+1} = \frac{N(N-1)\dots(N-k)}{(k+1)!} \leq N(N-1)\dots(N-k) \leq N^k(N-1)$$

The growth function and VC dimension

Therefore we have:

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

$$= N^{k+1} + 1$$

By induction, we know the statement is true $\sum_{i=0}^{D} \binom{N}{i} \leq N^D + 1$

Therefore, we have shown that:

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

The generalization theory: VC inequality

Theorem A.1 (Vapnik, Chervonenkis, 1971).

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

with probability at least $1-\delta$,

$$E_{\text{out}}(g)\leq E_{\text{in}}(g)+\sqrt{\frac{8}{N}\ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)}$$

If we use the polynomial bound based on d_{VC} instead of $m_{\mathcal{H}}(2N)$, another valid bound on the out-of-sample error,

$$E_{\text{out}}(g)\leq E_{\text{in}}(g)+\sqrt{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{\text{VC}}}+1)}{\delta}\right)}\ .$$

Proof of VC inequality

Theorem A.1 (Vapnik, Chervonenkis, 1971).

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|E_{\text{in}}(h)-E_{\text{out}}(h)\right|>\epsilon\right]\leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

这个sup取到时并不是E_{in}取到sup或者E_{out}取到sup。
这对我们衡量E_{out}造成了困难。左边还有$\mathscr{D}$的随机性。

Key step is to replace $E_{out}(h)$ by $E'_{in}(h)$ on another test set $\mathscr{D}'$ independent of $\mathscr{D}$. We use $E'_{in}(h)$ to estimate $E_{out}(h)$, since $E_{out}(h)$ is generally unknown and difficult to evaluate.

左边概率来自于$\mathscr{D}$是随机的

右边概率来自于$\mathscr{D}$和$\mathscr{D}'$是随机的

Lemma A.2.

$$\left(1-2e^{-\frac{1}{2}\epsilon^2 N}\right)\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|E_{\text{in}}(h)-E_{\text{out}}(h)\right|>\epsilon\right]\leq\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|E_{\text{in}}(h)-E'_{\text{in}}(h)\right|>\frac{\epsilon}{2}\right],$$

where the probability on the RHS is over $\mathcal{D}$ and $\mathcal{D}'$ jointly.

Proof. We can assume that $\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right] > 0$, otherwise there is nothing to prove.

$$
\begin{aligned}
& \mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\right] \\
\geq\ & \mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\ \textbf{and}\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right] \quad \text{(A.1)} \\
=\ & \mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right] \times
\end{aligned}
$$

$$
\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\ \middle|\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right].
$$

左边的概率在$\mathscr{D}$已经选定的条件下只$\mathscr{D}'$是随机的，但现在$\mathscr{D}$也是随机的。式子中依然是2部分都随机。

$$
\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\ \middle|\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]
$$

$$
\int_{\mathscr{D}(\omega)\in\Omega}\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\ \middle|\ \mathscr{D},\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]d\omega
$$

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\,\middle|\,\mathscr{D},\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]$$

$$\geq\ \mathbb{P}\left[|E_{\text{in}}(h^*)-E'_{\text{in}}(h^*)|>\tfrac{\epsilon}{2}\,\middle|\,\mathscr{D},\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right] \tag{A.2}$$

$$\geq\ \mathbb{P}\left[|E'_{\text{in}}(h^*)-E_{\text{out}}(h^*)|\leq\tfrac{\epsilon}{2}\,\middle|\,\mathscr{D},\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right] \tag{A.3}$$

$$\geq\ 1-2e^{-\frac{1}{2}\epsilon^2 N} \tag{A.4}$$

左边成功地去掉了sup

Once $\mathscr{D}$ is selected, the following arguments hold:

1. Inequality (A.2) follows because the event "$|E_{\text{in}}(h^*)-E'_{\text{in}}(h^*)|>\frac{\epsilon}{2}$" implies "$\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\frac{\epsilon}{2}$".

2. Inequality (A.3) follows because the events "$|E'_{\text{in}}(h^*)-E_{\text{out}}(h^*)|\leq\frac{\epsilon}{2}$" and "$|E_{\text{in}}(h^*)-E_{\text{out}}(h^*)|>\epsilon$" (which is given) imply "$|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\frac{\epsilon}{2}$".

3. Inequality (A.4) follows because h^* is fixed with respect to $\mathcal{D}'$ and so we can apply the Hoeffding Inequality to $\mathbb{P}[|E'_{\text{in}}(h^*)-E_{\text{out}}(h^*)|\leq\frac{\epsilon}{2}]$.

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\;\middle|\;\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]$$

$$=\int_{\mathscr{D}(\omega)\in\Omega}\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\;\middle|\;\mathscr{D},\ \sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]d\omega$$

$$\geq\int_{\mathscr{D}(\omega)\in\Omega}1-2e^{-\frac{1}{2}\epsilon^2N}d\omega\quad=\quad 1-2e^{-\frac{1}{2}\epsilon^2N}$$

Back to (A.1), we now have

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\right]$$

$$\geq\ \mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\ \times$$

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\;\middle|\;\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right].$$

$$\geq\ \mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\ \times\left(1-2e^{-\frac{1}{2}\epsilon^2N}\right)$$

Lemma A.2 is proved.

We continue to prove VC inequality

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

Note that we can assume $e^{-\frac{1}{2}\epsilon^2 N}<\frac{1}{4}$, because otherwise the bound in Theorem A.1 is trivially true. In this case, $1-2e^{-\frac{1}{2}\epsilon^2 N}>\frac{1}{2}$, so the lemma implies

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\leq 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\right].$$

现在我们要给这一项估上界。

我们考虑$\mathscr{D}$和$\mathscr{D}'$的选取方式为，先把$S=\mathscr{D}\cup\mathscr{D}'$给选出来，再随机的均分出两个数据集。

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\right]\\
=&\sum_{S}\mathbb{P}[S]\times\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\,\middle|\,S\right]\\
\leq&\sup_{S}\ \mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\,\middle|\,S\right].
\end{aligned}$$

当然也可能是积分

Let $\mathcal{H}(S)$ be the dichotomies that $\mathcal{H}$ can implement on the points in S. By definition of the growth function, $\mathcal{H}(S)$ cannot have more than $m_{\mathcal{H}}(2N)$ dichotomies. Suppose it has $M \leq m_{\mathcal{H}}(2N)$ dichotomies, realized by $h_1, \ldots, h_M$. Thus,

$$\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| = \sup_{h\in\{h_1,\ldots,h_M\}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| .$$

对于这个集合之外的h，他们所产生的结果和这批假设函数有重合，所以不用考虑

Then,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, S\right]$$

$$= \mathbb{P}\left[\sup_{h\in\{h_1,\ldots,h_M\}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, S\right]$$

这一步就又可以用有限元素的$\mathcal{H}$的方法进行讨论

$$\leq \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E'_{\text{in}}(h_m)| > \tfrac{\epsilon}{2} \,\middle|\, S\right] \tag{A.5}$$

$$\leq M \times \sup_{h\in\mathcal{H}} \mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, S\right], \tag{A.6}$$

这里的S是选定的，所以随机性来自于如何均分S

Lemma A.3.

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2}\right]$$

$$\leq m_{\mathcal{H}}(2N) \times \sup_{S} \sup_{h\in\mathcal{H}} \mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, S\right]$$

综合我们已经证过的两个Lemma，现在我们已经有了

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\leq 2m_{\mathcal{H}}(2N)\times\sup_{S}\sup_{h\in\mathcal{H}}\mathbb{P}\left[|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\,\middle|\,S\right]$$

现在我们估右边项的上界：$\sup_{S}\sup_{h\in\mathcal{H}}\mathbb{P}\left[|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\,\middle|\,S\right]$

只需要对两个sup里面的每一个S,h估出一个“一致(uniformly)”的上界即可。

Lemma A.4. For *any* h and *any* S,

$$\mathbb{P}\left[|E_{\text{in}}(h)-E'_{\text{in}}(h)|>\tfrac{\epsilon}{2}\,\middle|\,S\right]\leq 2e^{-\frac{1}{8}\epsilon^2 N},$$

where the probability is over random partitions of S into two sets $\mathcal{D}$ and $\mathcal{D}'$.

给定样本容量为$2N$的S下（样本S可以理解为条件的整体），采用0-1 loss:

$$E_{\text{in}}(h)=\frac{1}{N}\sum_{a_n\in\mathcal{D}}a_n,\ \text{and}\quad E'_{\text{in}}(h)=\frac{1}{N}\sum_{a'_n\in\mathcal{D}'}a'_n$$

这是2次在S里采样的均值

$$\mu=\frac{1}{2N}\sum_{n=1}^{2N}a_n=\frac{E_{\text{in}}(h)+E'_{\text{in}}(h)}{2}$$

这是整体的均值！！

可以验证这两个事件等价：$|E_{\text{in}}-\mu|>t\iff|E_{\text{in}}-E'_{\text{in}}|>2t$

下面就可以利用Hoeffding不等式！！

利用Hoeffding不等式:

$$\mathbb{P}\left[\left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > 2t\right] \;=\; \mathbb{P}\left[\,\left|E_{\text{in}} - \mu\right| > t\right] \leq 2e^{-2t^2 N}$$

再令$t = \epsilon/4$，即可证明Lemma A.4

现在结合我们已经证明的结果和Lemma A.4

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 2m_{\mathcal{H}}(2N) \times \sup_{S} \sup_{h\in\mathcal{H}} \mathbb{P}\left[\left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \tfrac{\epsilon}{2} \;\middle|\; S\right]$$

$$\mathbb{P}\left[\left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \tfrac{\epsilon}{2} \;\middle|\; S\right] \leq 2e^{-\frac{1}{8}\epsilon^2 N}$$

VC-inequality得证:

Theorem A.1 (Vapnik, Chervonenkis, 1971)**.**

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}.$$

Interpretation of VC inequality

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|E_{\text{in}}(h)-E_{\text{out}}(h)|>\epsilon\right]\leq 4((2N)^{d_{\text{vc}}}+1)e^{-\frac{1}{8}\epsilon^2 N}$$

$$E_{\text{out}}(g)\leq E_{\text{in}}(g)+\sqrt{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{\text{vc}}}+1)}{\delta}\right)}$$

with probability at least $1-\delta$

Example: Suppose you use linear separators for a model with 10 attributes and confidence level 95 % . The size of training data is $N = 1000$. What is the bound for $E_{in}(g) - E_{out}(g)$ of your learning?

$$\sqrt{\frac{8}{1000}\ln\left(\frac{4(2\cdot 1000)^{11}+1)}{0.05}\right)} = 0.8390$$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\underbrace{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{\text{vc}}}+1)}{\delta}\right)}_{\Omega(N,\mathcal{H},\delta)}}$$

With probability $\geq 1-\delta$, $\quad |E_{\text{out}} - E_{\text{in}}| \leq \Omega(N,\mathcal{H},\delta)$

泛化能力取决于三部分：训练集容量、假设集合、置信水平

$$E_{\text{out}} \leq E_{\text{in}} + \Omega(N,\mathcal{H},\delta)$$

泛化能力的提高：降低置信水平？增大训练集容量？慎选假设集合？

在给定置信水平$(1-\delta)$、给定训练集容量(N)的前提下，我们只能谨慎选择$\mathscr{H}$来提高泛化能力

Restraining $\mathscr{H}$: structural risk minimization (结构风险极小化) to penalize the model complexity

Out ultimate goal is to minimize the generalization error $E_{out}(g)$

$$E_{\text{out}} \leq E_{\text{in}} + \Omega(N, \mathcal{H}, \delta)$$

Ideal case: minimize this.

Can we minimize its upper bound?

Intuition:

- minimize Ω to guarantee $E_{out}(g)$ are close to $E_{in}(g)$, i.e., we can generalize what we have learned
- minimize $E_{in}(g)$ to guarantee that we have learned "good" information
- These two altogether yields a "good" model
- However, this is always a dilemma; need to strike a balance

  $\Longrightarrow$ this is called regularization (正则化)

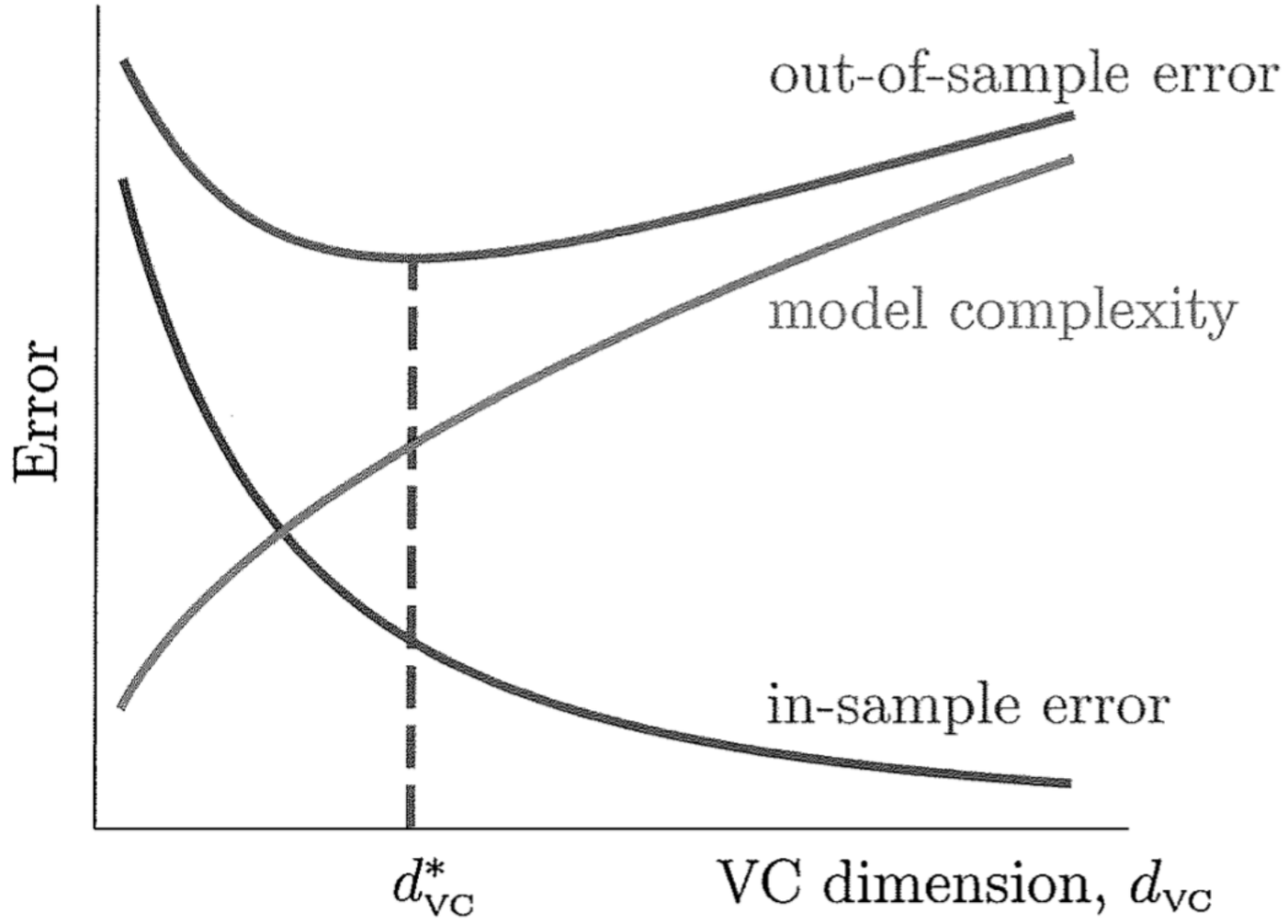
Restraining $\mathscr{H}$: structural risk minimization (结构风险极小化) to penalize the model complexity

This can be achieved by approaches like:

$$\min_{\boldsymbol{w}} \quad E_{in}(\boldsymbol{w}) \quad \text{s.t. } R(\boldsymbol{w}) \leq r$$

$$\min_{\boldsymbol{w}} \quad E_{in}(\boldsymbol{w}) + \lambda R(\boldsymbol{w})$$

$R(\boldsymbol{w})$ represents the model complexity



Sample Complexity

How many training examples N are needed to achieve a certain generalization performance specified by δ and ϵ

$$\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon \quad \Rightarrow \quad \begin{cases} N \geq \frac{8}{\epsilon^2}\ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right) \\ N \geq \frac{8}{\epsilon^2}\ln\left(\frac{4((2N)^{d_{\text{VC}}}+1)}{\delta}\right) \end{cases} \tag{2.13}$$

If we want certain ϵ and δ, how does N depend on d_{VC}?

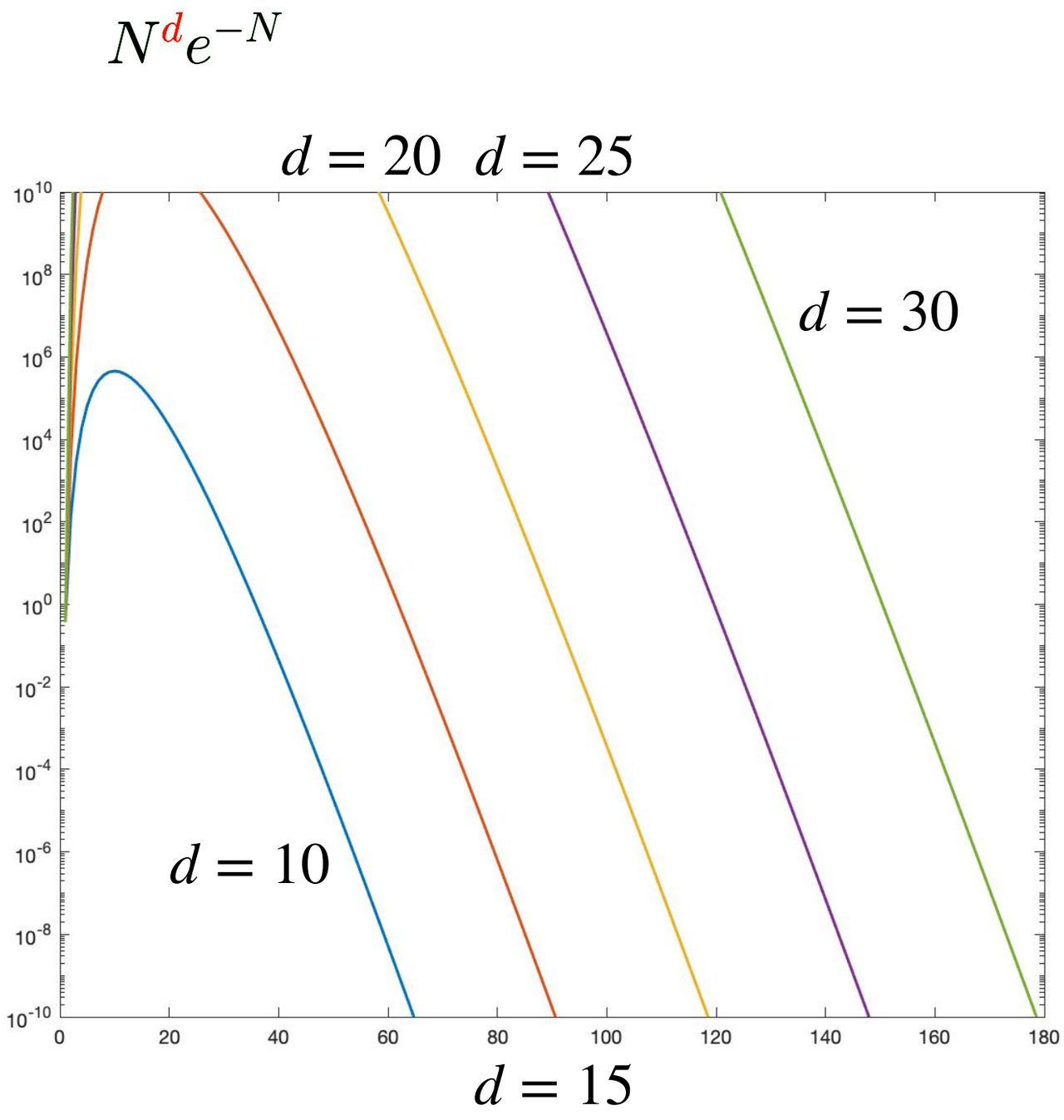
Let us look at $N^d e^{-N}$

Fix $N^{d}e^{-N} =$ small value

How does N change with d?

Rule of thumb:

$$N \geq 10\, d_{\text{VC}}$$



Sample Complexity

Example 2.6. Suppose that we have a learning model with $d_{\text{VC}} = 3$ and would like the generalization error to be at most 0.1 with confidence 90% (so $\epsilon = 0.1$ and $\delta = 0.1$). How big a data set do we need? Using (2.13), we need

$$N \geq \frac{8}{0.1^2} \ln\left(\frac{4(2N)^3 + 4}{0.1}\right).$$

Trying an initial guess of $N = 1,000$ in the RHS, we get

$$N \geq \frac{8}{0.1^2} \ln\left(\frac{4(2 \times 1000)^3 + 4}{0.1}\right) \approx 21,193.$$

We then try the new value $N = 21,193$ in the RHS and continue this iterative process, rapidly converging to an estimate of $N \approx 30,000$. If d_{VC} were 4, a similar calculation will find that $N \approx 40,000$. For $d_{\text{VC}} = 5$, we get $N \approx 50,000$.

You can see that the inequality suggests that the number of examples needed is approximately proportional to the VC dimension, as has been observed in practice. The constant of proportionality it suggests is 10,000, which is a *gross* overestimate; a more practical constant of proportionality is closer to 10. □

Using VC dimension

- Recall how we used validation data, or cross-validation error rates to select a complexity
- Use VC dimension based bound on test error similarly "structural risk minimization" (SRM)
- Other alternatives (statistical tools):
  - Probabilistic models: likelihood under model (rather than classification error)
  - AIC (Aikike information criterion)
  - BIC (Bayesian information criterion)