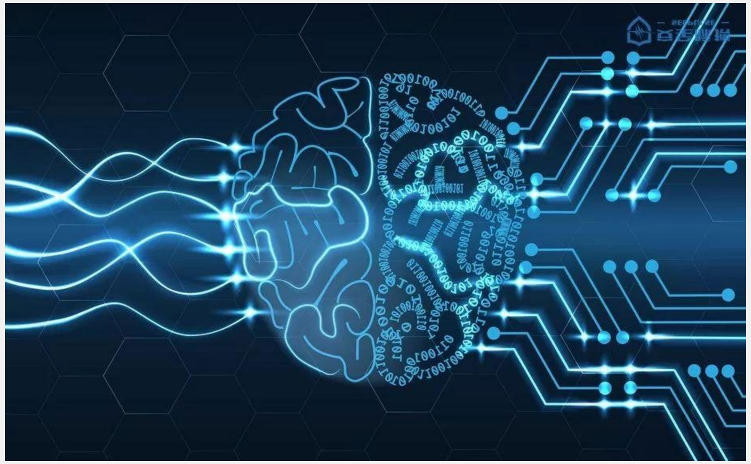


机器学习

Machine Learning

授课老师：刘雪明
电　　话：87543563
办 公 室：南一楼西303
邮　　箱：xm_liu@hust.edu.cn

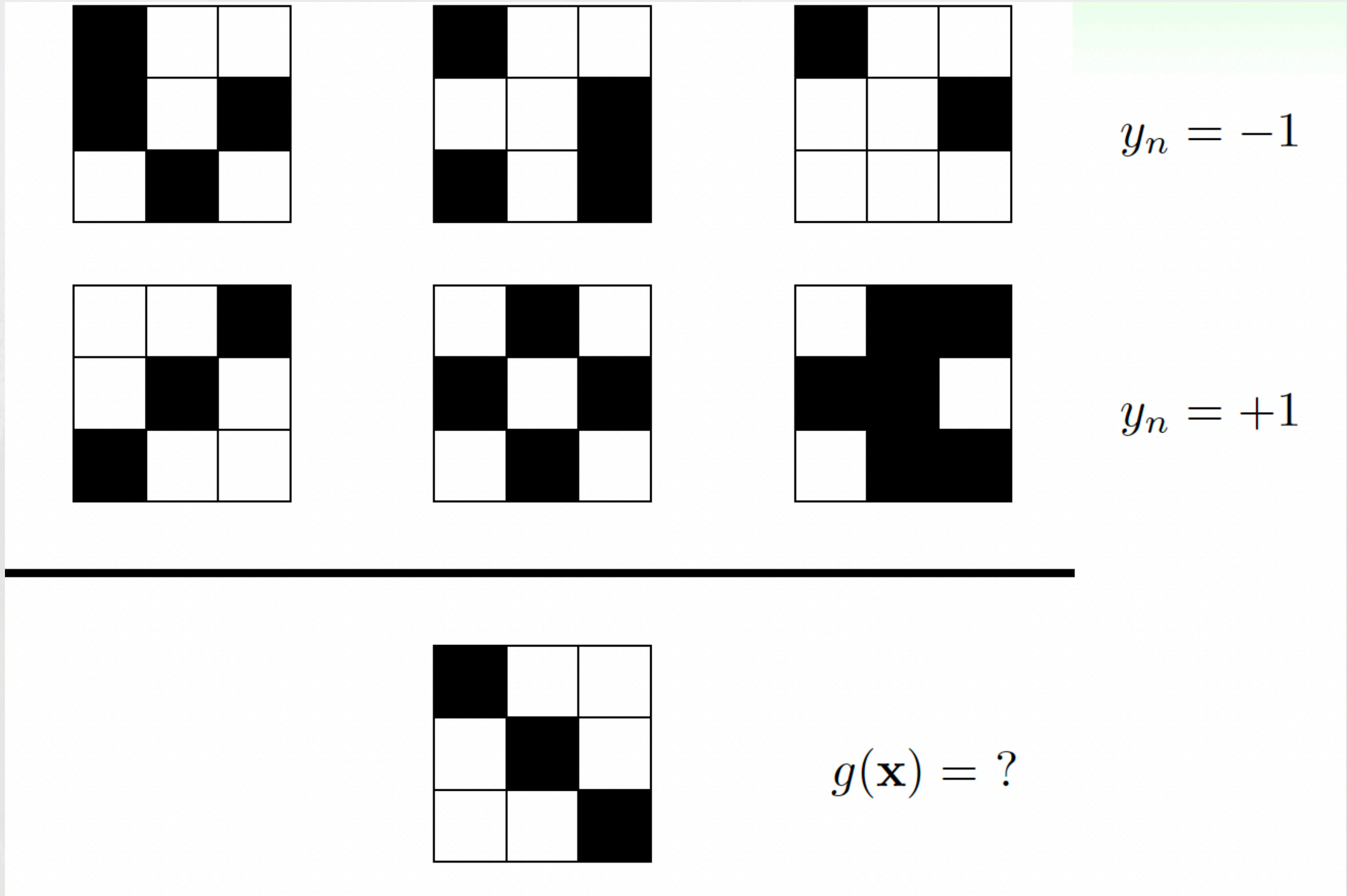
第二章、机器学习理论

目录
CONTENTS

本章学习的目的：**理解机器为什么可以学习**

1、机器可学习性分析

➢ 通过学习前6幅图的规律推测$g(x)$的值



✓ $g(x) = 1$

可能的理由1：对称

可能的理由2：左上角格子白色

……

✓ $g(x) = -1$

可能的理由1：右上角白色且中间列最多1个黑色格子

可能的理由2：左上角格子黑色

……

1、机器可学习性分析

➢ 一个简单的二分类问题，给定如下已知条件，可否学到一个函数g接近于目标函数f？

$\mathbf{x}_n$	$y_n = f(\mathbf{x}_n)$
0 0 0	○
0 0 1	×
0 1 0	×
0 1 1	○
1 0 0	×

目标函数f未知，可能的假设空间$\mathcal{H}$有2^8种可能，从中选择前5个结果与给定条件一致的2^3个假设

1、机器可学习性分析

	$\mathbf{x}$	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\mathcal{D}$	0 0 0	○	○	○	○	○	○	○	○	○	○
$\mathcal{D}$	0 0 1	×	×	×	×	×	×	×	×	×	×
$\mathcal{D}$	0 1 0	×	×	×	×	×	×	×	×	×	×
$\mathcal{D}$	0 1 1	○	○	○	○	○	○	○	○	○	○
$\mathcal{D}$	1 0 0	×	×	×	×	×	×	×	×	×	×
	1 0 1		?	○	○	○	○	×	×	×	×
	1 1 0		?	○	○	×	×	○	○	×	×
	1 1 1		?	○	×	○	×	○	×	○	×

- ✓ 这8个假设均能保证数据集$\mathcal{D}$上$g \approx f$
- ✓ 在数据集$\mathcal{D}$外没法确定哪个假设更准确 （没有免费午餐定理）

1、机器可学习性分析

➢ 没有免费午餐定理(No Free Lunch Theorem)

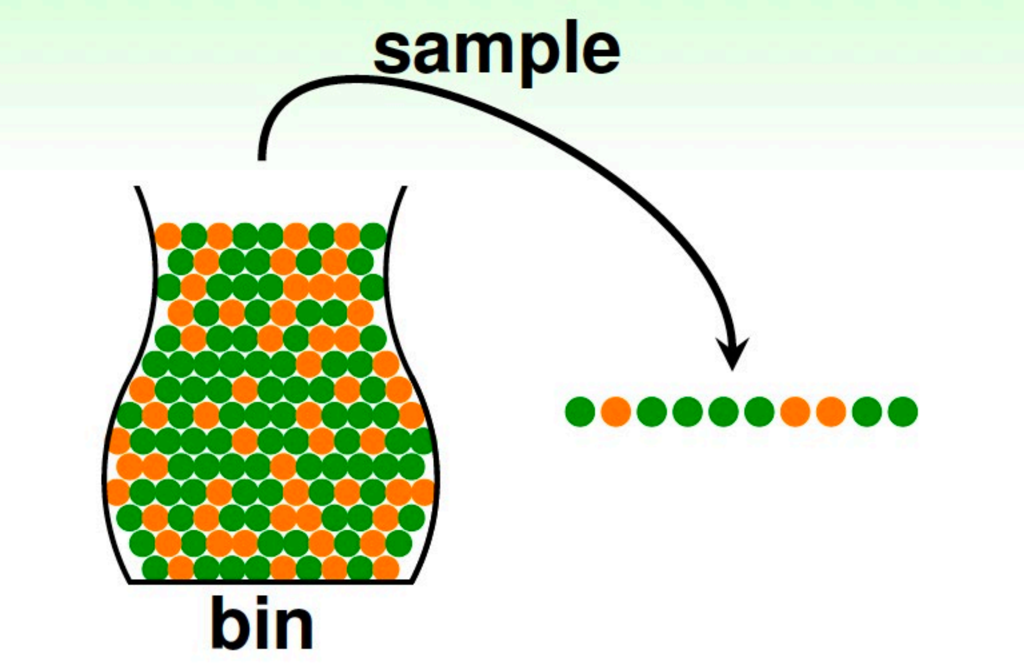
任何一个预测函数，如果在一些训练样本上表现好必然在另一些训练样本上表现不好，如果不对数据在特征空间的先验分布有一定假设，那么表现好与表现不好的情况一样多。

沃尔珀特(Wolpert)

✓ 如何确定在数据集$\mathcal{D}$外$g \approx f$?

1、机器可学习性分析

- ➢ 统计学上如何估计未知量：如何估计罐子里橙色球的比例?



- ✓ 假设罐子里橙色球比例为u
- ✓ 采样样本中橙色球比例为v

- ➢ 是否可以根据v 来推测u?

- ✓ 可能罐子里大部分是橙色球，而抽样全是绿球
- ✓ 可能两者近似相等$u \approx v$

1、机器可学习性分析

- ✓ 未知量：罐子里橙色球比例为u
- ✓ 已知量：采样样本中橙色球比例为v

Wassily Hoeffding

➢ **霍夫丁不等式**(Hoeffding’s Inequality)

当抽样样本(N)足够大时，u与v可能近似相等（在ε范围内）

$$P(|u-v|>\varepsilon)\le 2\exp(-2\varepsilon^2 N)$$

这时称“$u\approx v$”是可能近似正确的（probably approximately correct (PAC))

1、机器可学习性分析

➢ **霍夫丁不等式**(Hoeffding’s Inequality)

当抽样样本(N)足够大时， u与v可能近似相等（在ε 范围内）

$$P(|u-v|>\varepsilon)\le 2\exp(-2\varepsilon^2 N)$$

测验：u=0.4，采样10个样本得到的比例v<0.1，那u和v近似相等可能性的上限是多少?

A、0.67 B、0.4 C、0.33 D、0.05

✔

1、机器可学习性分析

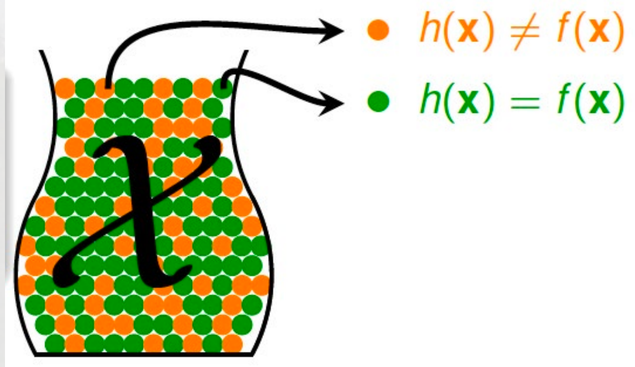
➤ 抽样模型及霍夫丁不等式与学习的联系

罐子抽样模型

- 需确定的量：橙色球比例u
- 球的空间：罐子
- 橙色球
- 绿色球
- 从罐子里抽N个样本

学习模型

- 需确定的量：给定假设上$\mathrm{h}(\vec{x}) = f(\vec{x})$是否成立?
- 输入空间：$\vec{x} \in \mathcal{X}$
- h正确：$\mathrm{h}(\vec{x}) = f(\vec{x})$
- h错误：$\mathrm{h}(\vec{x}) \neq f(\vec{x})$
- 数据集$\mathcal{D} = \{(x_n, y_n)\}$上确定$h$是否正确



数据集规模足够大时，可通过数据集上$\mathrm{h}(\vec{x}) \neq f(\vec{x})$的概率来推测输入空间上$\mathrm{h}(\vec{x}) \neq f(\vec{x})$的概率

1、机器可学习性分析

➤ 抽样模型及霍夫丁不等式与学习的联系

罐子抽样模型	学习模型
▪ 需确定的量：橙色球比例u	▪ 需确定的量：给定假设上$\mathrm{h}(\vec{x}) = f(\vec{x})$是否成立？
▪ 球的空间：罐子	▪ 输入空间：$\vec{x} \in \mathcal{X}$
▪ 橙色球	▪ h正确：$\mathrm{h}(\vec{x}) = f(\vec{x})$
▪ 绿色球	▪ h错误：$\mathrm{h}(\vec{x}) \neq f(\vec{x})$
▪ 从罐子里抽N个样本	▪ 数据集$\mathcal{D} = \{(x_n, y_n)\}$上确定$h$是否正确

备注：统计分析要求样本是均匀采样得到的

机器学习要求样本满足**独立同分布** (independent and identically distributed, 简写**i.i.d.**)

1、机器可学习性分析

➢ 抽样模型及霍夫丁不等式与学习的联系

定义数据集上的错误率及$h(\vec{x}) \neq f(\vec{x})$的概率为$E_{in}(h)$，样本空间上的错误率为$E_{out}(h)$，当数据集规模很大时，两者在ε误差范围内近似相等：

$$\boldsymbol{P(|E_{in}(h) - E_{out}(h)| > \varepsilon) \leq 2\exp(-2\varepsilon^2 N)}$$

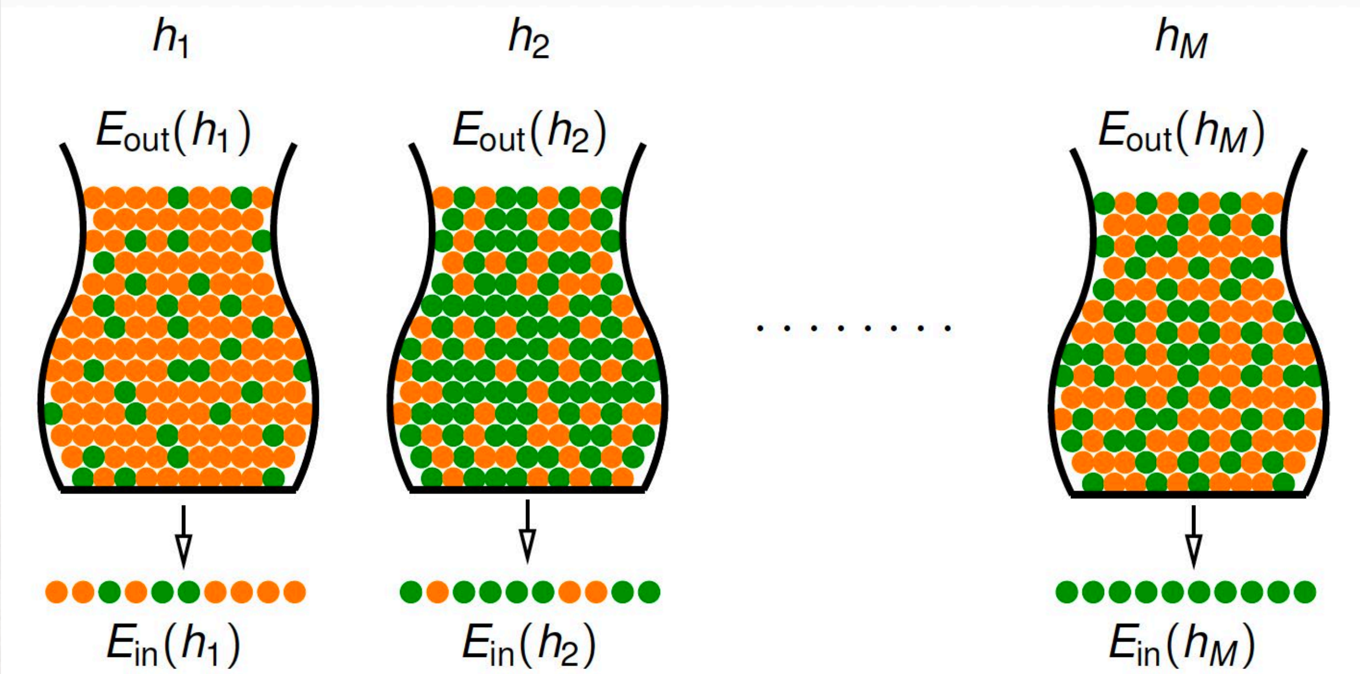
✓ 若$E_{in}(h) \approx E_{out}(h)$ 且$E_{in}(h)$很小，
则样本空间上的错误率$E_{out}(h)$小，即$h \approx f$

✓ 若选择假设h作为学习函数g，且$E_{in}(h)$小，则" $g \approx f$" PAC

✓ 反之，则" $g \neq f$" PAC

1、机器可学习性分析

➢ 存在多个假设的情况



❖ 在假设空间$\mathcal{H} = \{h_1, h_2, \dots h_M\}$上选择样本错误率$E_{in}(h)$低的假设$h_k$作为学习函数$g$，能说明在所有样本空间上的错误率$E_{out}(h)$也低吗?

1、机器可学习性分析

- ✓ 依据墨菲定律， 虽然某事件发生几率极低，但如果重复次数达到大规模的时候，则小概率的事件必然发生。即$E_{in}(h)$与$E_{out}(h)$相差甚远的情况（称为BAD样本）可能发生

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	...	Hoeffding
h	**BAD**					**BAD**		$\mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h] \le \ldots$

- ✓ 依据霍夫丁不等式， BAD样本产生的概率低

$$\mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D}] = \sum_{\text{all possible}\mathcal{D}} \mathbb{P}(\mathcal{D}) \cdot [\![\textbf{BAD } \mathcal{D}]\!]$$

1、机器可学习性分析

- 存在多个假设时

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	Hoeffding
h_1	**BAD**					**BAD**	$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_1] $\leq \ldots$
h_2		**BAD**					$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_2] $\leq \ldots$
h_3	**BAD**	**BAD**				**BAD**	$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_3] $\leq \ldots$
...							
h_M	**BAD**					**BAD**	$\mathbb{P}_{\mathcal{D}}$ [**BAD** $\mathcal{D}$ for h_M] $\leq \ldots$
all	**BAD**	**BAD**				**BAD**	**?**

- 算法在假设空间 $\mathcal{H} = \{h_1, h_2, \ldots h_M\}$ 中自由选择时选到BAD样本的概率为

$$
\begin{aligned}
& \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D}] \\
= \; & \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_1 \textbf{ or } \textbf{BAD } \mathcal{D} \text{ for } h_2 \textbf{ or } \ldots \textbf{ or } \textbf{BAD } \mathcal{D} \text{ for } h_M] \\
\leq \; & \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_1] + \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_2] + \ldots + \mathbb{P}_{\mathcal{D}}[\textbf{BAD } \mathcal{D} \text{ for } h_M] \\
& \text{(union bound)} \\
\leq \; & 2\exp\left(-2\epsilon^2 N\right) + 2\exp\left(-2\epsilon^2 N\right) + \ldots + 2\exp\left(-2\epsilon^2 N\right) \\
= \; & 2M\exp\left(-2\epsilon^2 N\right)
\end{aligned}
$$

1、机器可学习性分析

➢ 存在多个假设时，$P_{\mathcal{D}}[BAD\ \mathcal{D}] \leq 2M exp(-2\varepsilon^2 N)$

✓ 若假设空间规模$|\mathcal{H}| = M$有限，而样本量N足够大时，任选假设h_k作为学习函数g，都有$E_{in}(g) \approx E_{out}(g)$（**predict**）

✓ 若选到的学习函数$E_{in}(g) \approx 0$（**train**），则对应的$E_{out}(g) \approx 0$

➡ 数据集$\mathcal{D}$外$g \approx f$ ➡ 机器可学习

❖ 问题：当M无限时，怎么论证机器可学习?

❖ 参考资料：**Learning from data**, 2012, Abu-Mostafa et al.

1、机器可学习性分析

➢ 测验：根据霍夫定不等式

$$P(|E_{in}(h) - E_{out}(h)| > \varepsilon) \leq 2M\exp(-2\varepsilon^2 N) = \delta$$

可由给定的容忍误差ε和对坏样本的容忍误差δ来确定需要收集多大的数据集N来满足该要求。给定$\boldsymbol{\varepsilon = 0.1, \delta = 0.05, M = 100}$,那么需要多大的数据集。

A、215　　B、415 ✓　　C、615　　D、815

$$N = \frac{1}{2\epsilon^2} \ln \frac{2M}{\delta}$$

2、Hoeffding不等式与模式二分性

❖ 问题：当M无限时，怎么论证机器可学习？

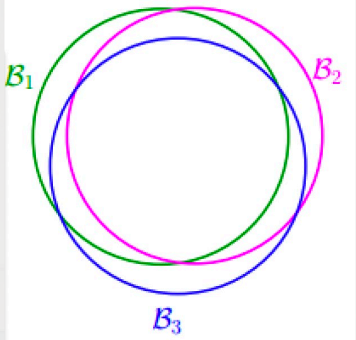
➢ 存在多个假设时，Hoeffding不等式$P_D[BAD\ \mathcal{D}] \le 2M exp(-2\varepsilon^2 N)$中$M$的由来：

$$
\begin{aligned}
& \mathbb{P}_{\mathcal{D}}[\textbf{BAD}\ \mathcal{D}] \\
= \ & \mathbb{P}_{\mathcal{D}}[\textbf{BAD}\ \mathcal{D} \text{ for } h_1 \textbf{ or } \textbf{BAD}\ \mathcal{D} \text{ for } h_2 \textbf{ or } \ldots \textbf{ or } \textbf{BAD}\ \mathcal{D} \text{ for } h_M] \\
\le \ & \mathbb{P}_{\mathcal{D}}[\textbf{BAD}\ \mathcal{D} \text{ for } h_1] + \mathbb{P}_{\mathcal{D}}[\textbf{BAD}\ \mathcal{D} \text{ for } h_2] + \ldots + \mathbb{P}_{\mathcal{D}}[\textbf{BAD}\ \mathcal{D} \text{ for } h_M] \\
& \text{(union bound)} \\
\le \ & 2\exp\left(-2\epsilon^2 N\right) + 2\exp\left(-2\epsilon^2 N\right) + \ldots + 2\exp\left(-2\epsilon^2 N\right) \\
= \ & 2M\exp\left(-2\epsilon^2 N\right)
\end{aligned}
$$

这里的union bound计算假定不同假设间没有交集

2、Hoeffding不等式与模式二分性

- 通常不同假设对应的坏事件间存在交集，上述式子中union bound被过高估计了



- 问题：如何寻找这些坏事件间重叠的部分？使得霍夫丁不等式右边无限大的M可被某个有限值$m_{\mathcal{H}}$所限定，即

$$P(|E_{in}(h) - E_{out}(h)| > \varepsilon) \leq 2m_{\mathcal{H}}\exp\left(-2\varepsilon^2 N\right)$$

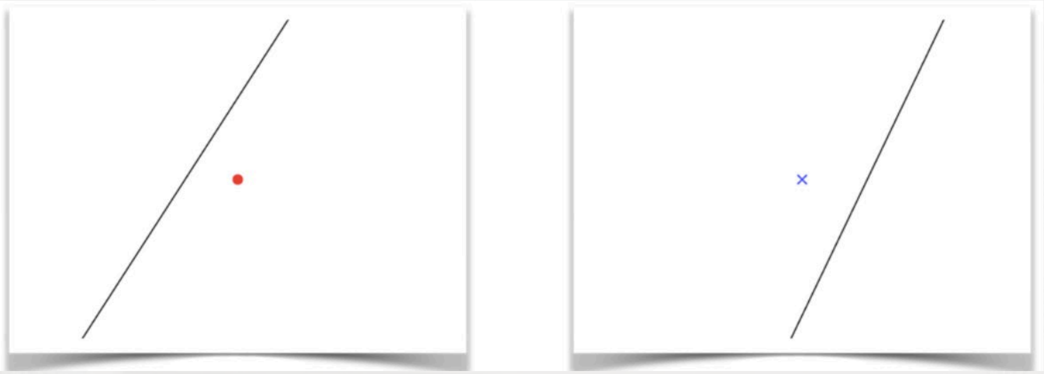
下面以感知器学习模型PLA (Perceptron Learning Algorithm) 为例来说明

2、Hoeffding不等式与模式二分性

- 感知器学习模型PLA（二分类模型且线性可分）中可能的假设空间$\mathcal{H}$是无限大的（存在无数可能的分界线$|\mathcal{H}| = M = \infty$），但有效直线的种类$effective(N) = |\mathcal{H}(x_1, x_2, \ldots x_N)|$却是有限的

2、Hoeffding不等式与模式二分性

- ✓ 考虑2维平面上1个样本的二分类模型（$\mathcal{H}$是平面上所有的线)



该模型存在的无限个假设可被归类为上述2种（$effective(N) = 2$）：

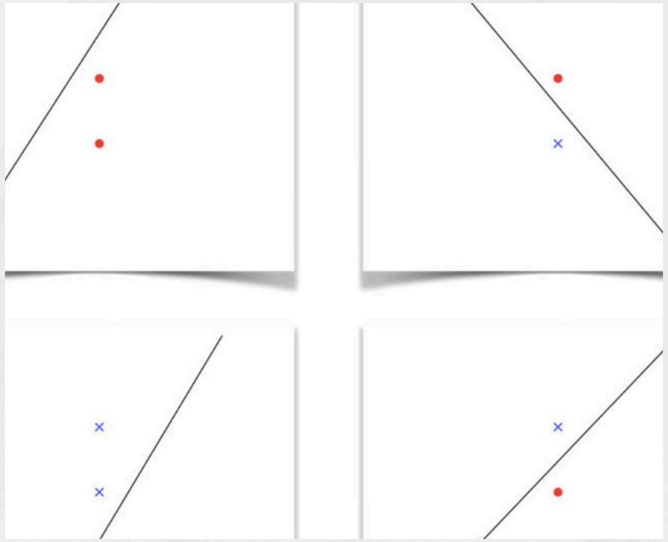
  - ✓ 线在样本左边，样本分类为+1（红圈）；
  - ✓ 线在样本右边，样本分类为-1（蓝叉）

- ➢ 样本要么被判定为+1，要么被判定为-1，其中任意一种输出叫做一种dichotomy（二分、对分）

2、Hoeffding不等式与模式二分性

✓ 下面考虑2个样本的二分类模型

该模型无限多的假设可被归类为右边4种，输出分别为(+1,+1)，(-1,+1)，(-1,-1)，(-1,-1)，称有效直线数量为$effective(N) = 4$

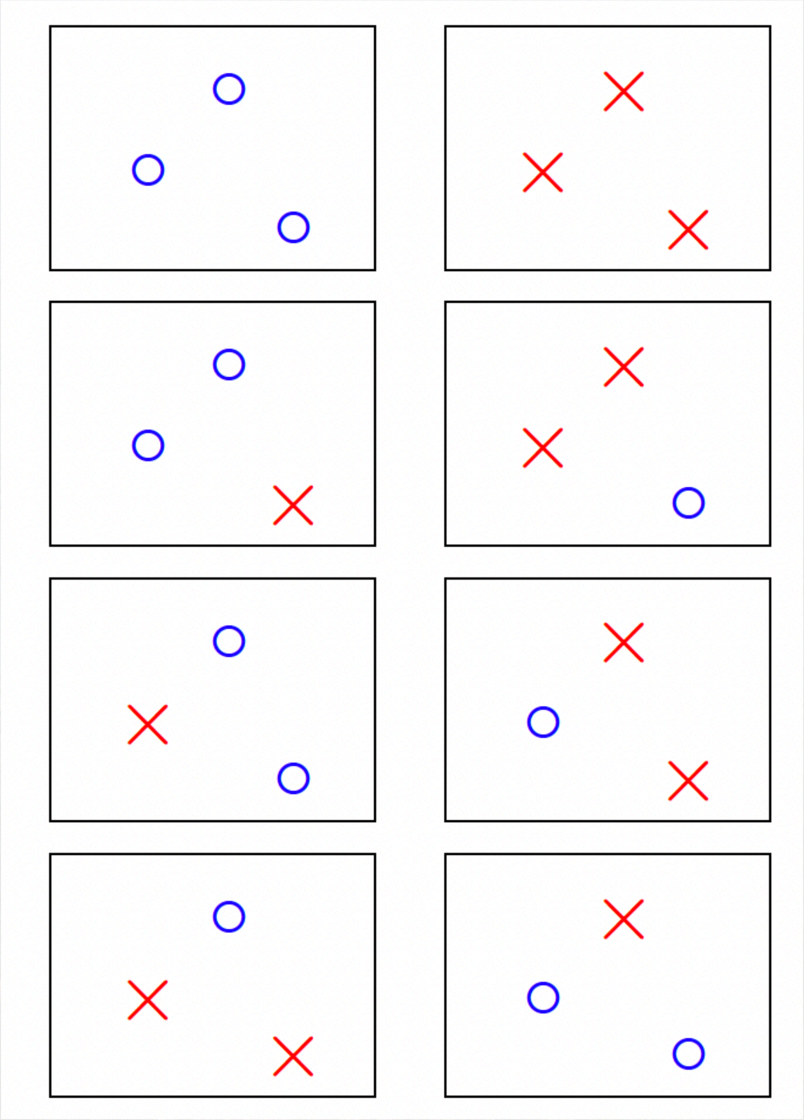


- $N = 2$的感知器分类模型的假设空间有4种dichotomies
- 若两个假设h_1, h_2对训练数据的输出$\mathcal{D}$是相同的，即$\left(h_1(x_1), h_1(x_2), \ldots h_1(x_N)\right) = \left(h_2(x_1), h_2(x_2), \ldots h_2(x_N)\right)$，则称$h_1$，$h_2$对这N个样本是“等效”的，其输出称为$\mathcal{D}$的一种dichotomy
- N个样本的感知器分类模型最多存在2^N种dichotomies

2、Hoeffding不等式与模式二分性

✓ 下面考虑3个样本的二分类模型

该模型无限多的假设可被归类为右边8种，即有效直线数量为$effective(N) = 8$



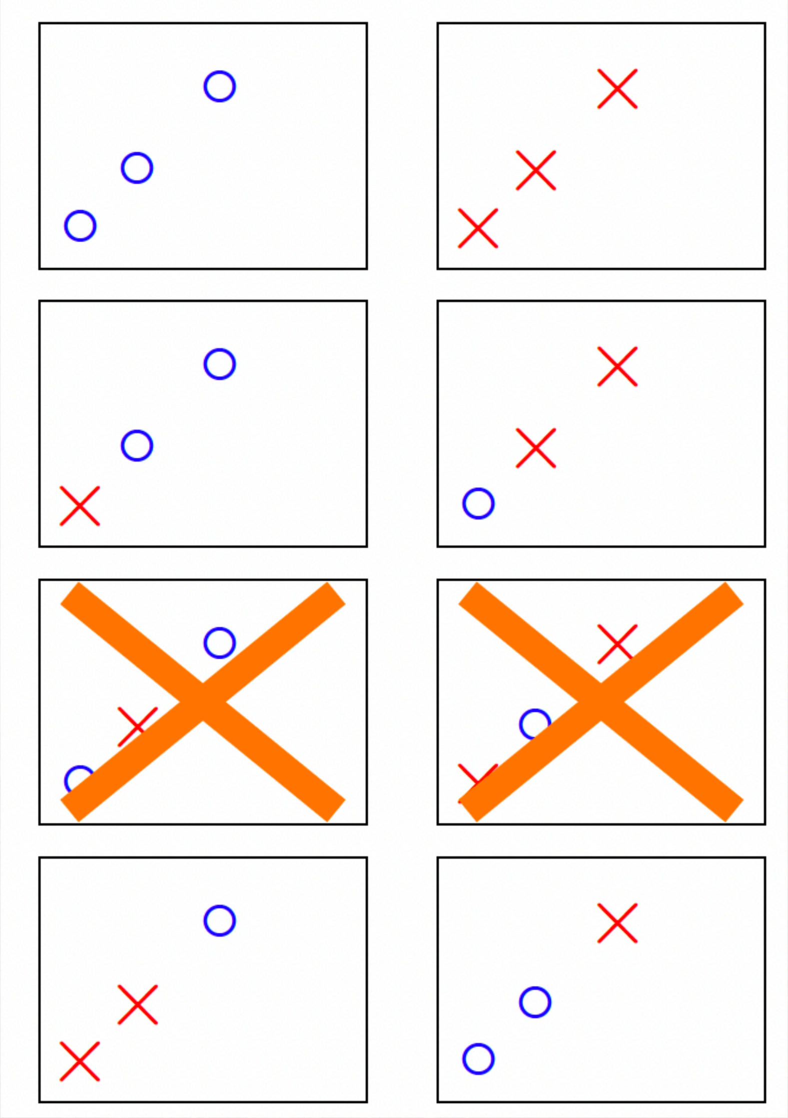
➢ 3个样本的感知器分类模型存在2^3种dichotomies

2、Hoeffding不等式与模式二分性

✓ 下面考虑直线排列的3个样本的二分类模型

该模型无限多的假设可被归类为右边6种，即有效直线数量为$effective(N) = 6$

➢ 直线排列的3个样本的感知器分类模型存在$6 < 2^3$种dichotomies

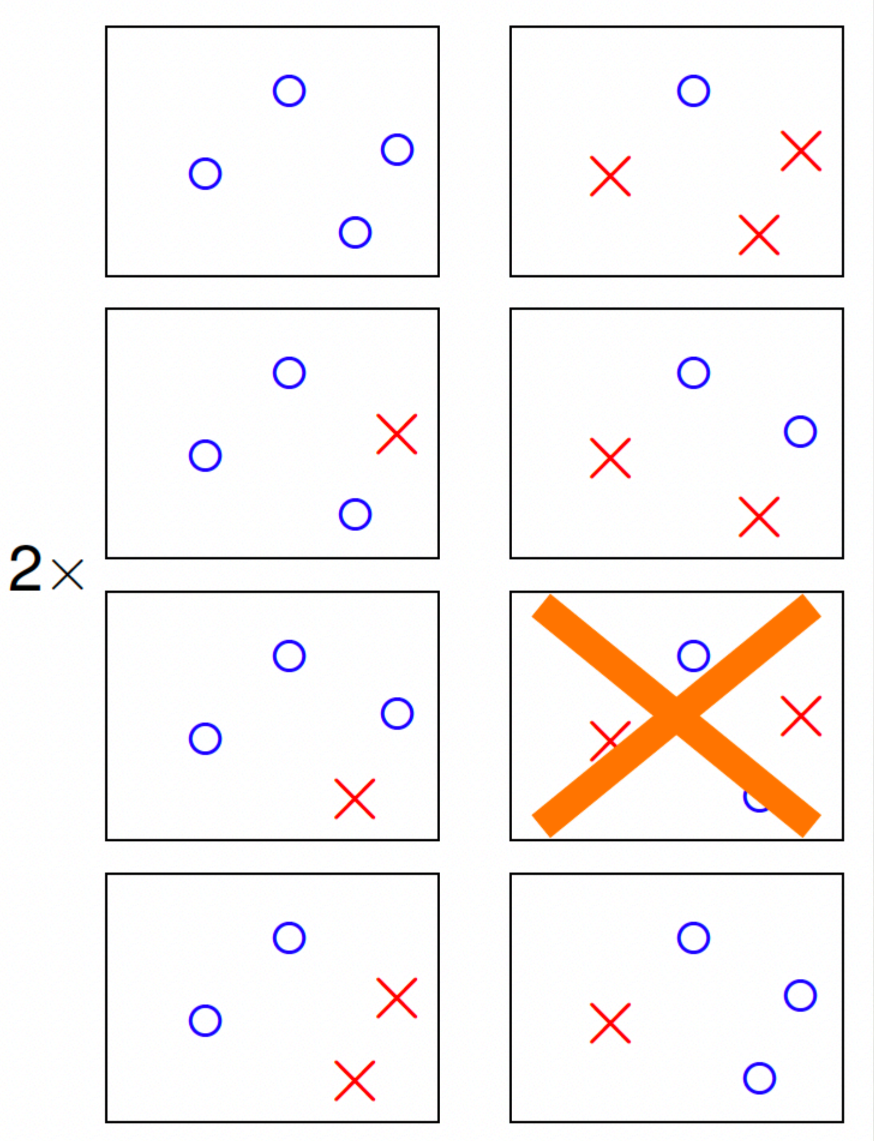


2、Hoeffding不等式与模式二分性

✓ 下面考虑4个样本的二分类模型

该模型无限多的假设可被归类为右边14种

即有效直线数量为$effective(N) = 14$



➢ 4个样本的感知器分类模型存在

$14 < 2^3$种dichotomies

2、Hoeffding不等式与模式二分性

- 测验：2维平面上直线排列的4样本二分类模型存在多少dichotomies?

2、Hoeffding不等式与模式二分性

- 用有效直线的数量$effective(N)$代替M，霍夫定不等式可写成

$$\boldsymbol{P(|E_{in}(h)-E_{out}(h)|>\varepsilon)\le 2}effective(N)\mathbf{exp}\left(-2\varepsilon^2N\right)$$

- 已知$effective(N)\le 2^N$，若能保证$effective(N)\ll 2^N$，则上式右边接近于0，$E_{in}(h)\approx E_{out}(h)$，说明机器学习是可能的
- 2维平面上的感知器模型中，$effective(N)=|\mathcal{H}(x_1,x_2,\dots x_N)|$与样本数量$N$及输入本身性质（如：排列方式）有关
- 定义成长函数$m_{\mathcal{H}}(N)=max_{x_1,x_2,\dots x_N\in\mathcal{X}}|\mathcal{H}(x_1,x_2,\dots x_N)|$消除对输入的依赖，其值与假设空间$\mathcal{H}$与样本数量$N$有关

2、Hoeffding不等式与模式二分性

- 用成长函数$m_{\mathcal{H}}(N)$代替M，霍夫定不等式可写成

$$P(|E_{in}(h)-E_{out}(h)|>\varepsilon)\le 2m_{\mathcal{H}}(N)\exp\left(-2\varepsilon^2N\right)$$

  - ✓ 若$m_{\mathcal{H}}(N)$为多项式，则 $E_{\text{in}}(h)\approx E_{\text{out}}(h)$成立
  - ✓ 若$m_{\mathcal{H}}(N)$为指数函数，则 $E_{\text{in}}(h)\approx E_{\text{out}}(h)$不成立

- 2维平面上感知器学习模型的成长函数$m_{\mathcal{H}}(N)$为：

N	$m_{\mathcal{H}}(N)$
1	2
2	4
3	max(…,6,8)=8
4	max(…,8,14)=14<2^4
…	…
N	$<2^N$

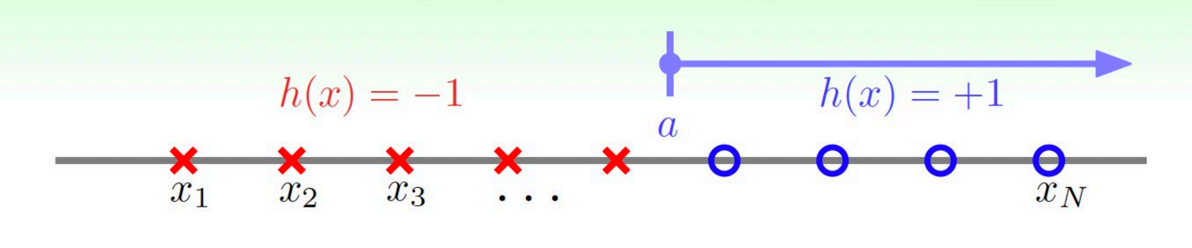
2、Hoeffding不等式与模式二分性

➢ 1维平面上感知器学习模型的成长函数$m_{\mathcal{H}}(N)$

❖ Positive Rays

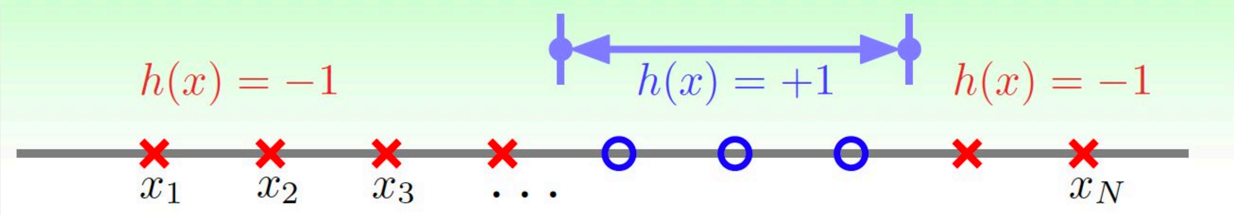
$$h(x) = \text{sign}(x - a)$$

$$m_{\mathcal{H}}(N) = N + 1$$



❖ Positive Intervals

$$h(x) = \begin{cases} +1, \text{if } x \in [l, r) \\ -1, \quad \text{其他} \end{cases}$$



$$m_{\mathcal{H}}(N) = C_{N+1}^{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

2、Hoeffding不等式与模式二分性

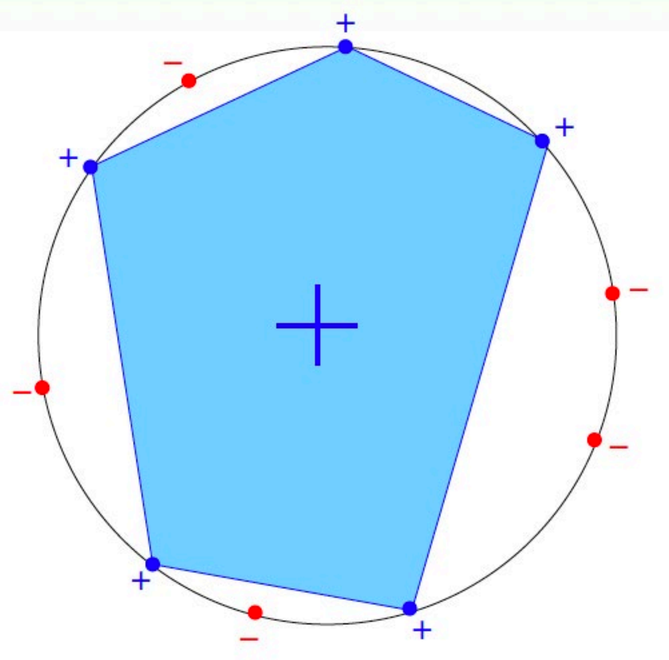
- Convex set的成长函数$m_{\mathcal{H}}(N)$

✓ 考虑一种输入：所有样本点在分布在一个圆上

✓ 假设空间$\mathcal{H}$为任意凸多边形，该多边形所覆盖到的点为+1，其他点为-1

✓ 成长函数为：

$$m_{\mathcal{H}}(N) = 2^N$$



➡ 这N个样本可被$\mathcal{H}$所“打散”（shattered）

2、Hoeffding不等式与模式二分性

- 不同假设空间对应的成长函数$m_{\mathcal{H}}(N)$
  - 一维空间上positive rays
  - 一维空间上positive intervals
  - Convex set
  - 二维平面上的感知器模型

$m_{\mathcal{H}}(N) = N + 1$

$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

$m_{\mathcal{H}}(N) = 2^N$

$m_{\mathcal{H}}(N) \le 2^N$

- 若k个样本的任意输入都不能被$\mathcal{H}$所打散，即 $m_{\mathcal{H}}(k) < 2^k$ 则称k为$\mathcal{H}$的break point（突破点），且$k+1, k+2, \ldots$也是break point。我们关注最小的break point

2、Hoeffding不等式与模式二分性

- 不同假设空间对应的**最小break point**

❖ positive rays	$m_{\mathcal{H}}(N) = N + 1$	2
❖ positive intervals	$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$	3
❖ Convex set	$m_{\mathcal{H}}(N) = 2^N$	\
❖ 二维平面上的感知器模型	$m_{\mathcal{H}}(N) \leq 2^N$	4

- 定理：若存在break point $k \geq 3$，且$N \geq 2$，则$m_{\mathcal{H}}(N) \leq N^{k-1}$

  证明可参考： **Learning from data**, 2012, Abu-Mostafa et al.

➡ 若对于$\mathcal{H}$存在break point k，则$E_{\text{in}}(g) \approx E_{\text{out}}(g)$

2、Hoeffding不等式与模式二分性

➢ 测验：直线排列的二分类模型中考虑假设空间$\mathcal{H}$为positive and negative rays，即$h(x) = \pm\text{sign}(x - a)$，其成长函数$m_{\mathcal{H}}(N) = ?$ ，

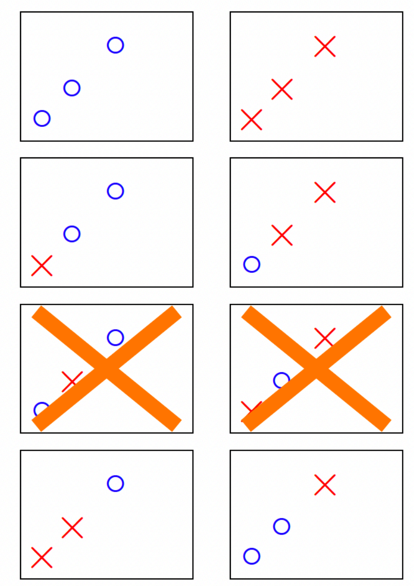
A、N　　B、$N + 1$　　C、$2N$　　D、2^N

✓

$\mathcal{H}$最小的break point是多少?

A、1　　B、2　　C、3　　D、4

✓



如3样本情形

3、VC维理论与模型复杂性

- 前面的论述用$m_{\mathcal{H}}(N)$代替M，从$\mathcal{H}$中选取学习函数g，有

$$P(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 2m_{\mathcal{H}}(N)\exp(-2\varepsilon^2 N)$$

但直接替换是有问题的，Vapnik 和 Chervonenkis 提出并完整地证明准确的式子如下（具体证明参考论文[Vapnik et al. 1971]）

$$P(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 2 \cdot 2m_{\mathcal{H}}(2N)\exp\left(-2 \cdot \frac{1}{16}\varepsilon^2 N\right)$$

- 上式右边是**VC Bound**，若存在break point $k \geq 3$，则

$$P(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 4(2N)^{k-1}\exp\left(-\frac{1}{8}\varepsilon^2 N\right)$$

3、VC维理论与模型复杂性

- VC维（VC Dimension）的定义

假设空间$\mathcal{H}$的VC维$d_{\mathrm{VC}}(\mathcal{H})$表示$\mathcal{H}$所能“打散”（Shatter）的最大样本数$N_s$，即（$d_{\mathrm{VC}} = N_s$）

N_s是使得$m_{\mathcal{H}}(N_s) = 2^{N_s}$成立的最大值

- $d_{\mathrm{VC}}(\mathcal{H})$=最小的break point $k-1$

- 若$N_s \geq 2$，即$d_{\mathrm{VC}} \geq 2$，则$m_{\mathcal{H}}(N) \leq N^{d_{\mathrm{VC}}}$

3、VC维理论与模型复杂性

- 不同假设空间对应的**VC维**

❖ positive rays	$m_{\mathcal{H}}(N) = N + 1$	$d_{\mathrm{VC}} = 1$
❖ positive intervals	$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$	$d_{\mathrm{VC}} = 2$
❖ Convex set	$m_{\mathcal{H}}(N) = 2^N$	$d_{\mathrm{VC}} = \infty$
❖ 二维平面上的感知器模型	$m_{\mathcal{H}}(N) \leq N^3$,对于 N≥ 2	$d_{\mathrm{VC}} = 3$

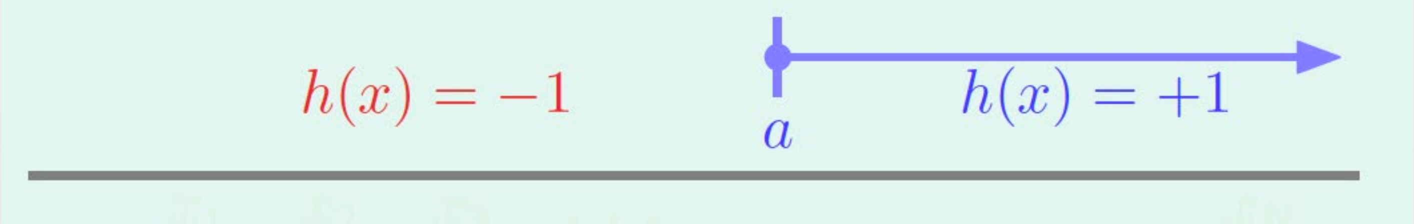
$$P(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 4(2N)^{d_{\mathrm{VC}}} \exp\left(-\frac{1}{8}\varepsilon^2 N\right)$$

- 若VC维d_{VC}有限，则在样本数N大的情况下：$E_{\mathrm{in}}(g) \approx E_{\mathrm{out}}(g)$

3、VC维理论与模型复杂性

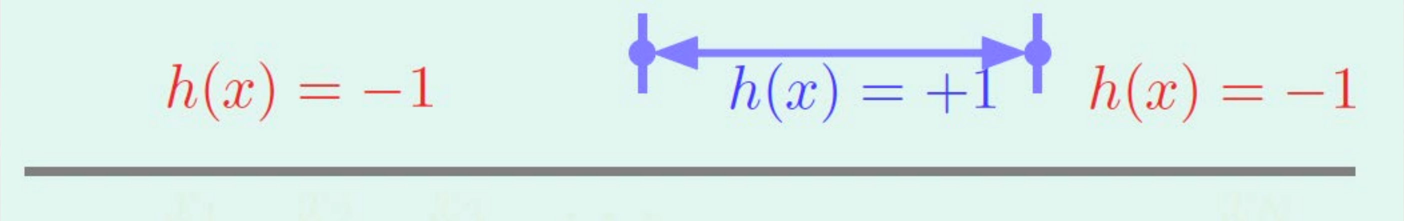
➢ VC维可能的物理意义

❖ positive rays



$\boldsymbol{d_{\mathrm{VC}} = 1}$　　$h(x) = \mathrm{sign}(x - a)$　　自由参数：a

❖ positive intervals



$\boldsymbol{d_{\mathrm{VC}} = 2}$　　$h(x) = \begin{cases} +1, \text{if } x \in [l, r) \\ -1, \quad \text{其他} \end{cases}$　　自由参数：l, r

VC维$d_{\mathrm{VC}} \approx$ 自由参数的数量（*并不总是成立*）

3、VC维理论与模型复杂性

- M与VC维d_{VC}对机器可学习性的影响，机器可学习说明

  1) 预测误差接近于训练误差$E_{out}(h) \approx E_{in}(h)$；

  2) $\mathcal{H}$存在可选的某假设g使得训练误差足够小$E_{in}(g) \approx 0$

M小	M大
1) 满足，$P(BAD) \le 2Mexp(\dots)$ 2) 不一定满足，假设空间选择太少	1) 不满足，$P(BAD) \le 2Mexp(\dots)$ 2) 满足，假设空间选择很多

d_{VC}小	d_{VC}大
1) 满足，$P(BAD) \le 4(2N)^{d_{VC}} exp(\dots)$ 2) 不一定满足，假设空间有效模式少	1) 不满足，$P(BAD) \le 4(2N)^{d_{VC}} exp(\dots)$ 2) 满足，假设空间有效模式多

3、VC维理论与模型复杂性

- 重写VC Bound

$$P(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 4(2N)^{d_{vc}} \exp\left(-\frac{1}{8}\varepsilon^2 N\right)$$

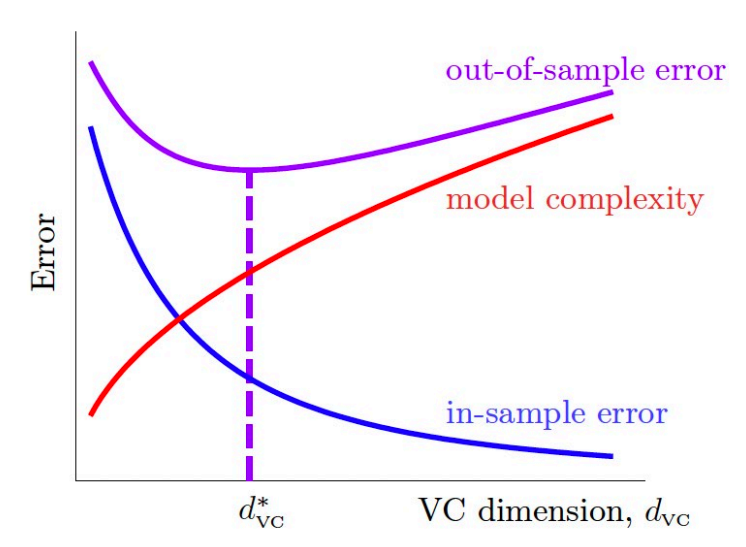
- 令 $4(2N)^{d_{vc}} \exp\left(-\frac{1}{8}\varepsilon^2 N\right) = \delta$ ➡ $\varepsilon = \sqrt{\frac{8}{N}\ln(\frac{4(2N)^{d_{vc}}}{\delta})}$

- 霍夫丁不等式表示坏事件发生的概率， $|E_{in}(h) - E_{out}(h)| \leq \varepsilon$ 则表示好事发生，即样本预测误差与训练误差相当，有

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

3、VC维理论与模型复杂性

- 式子 $\Omega(N,\mathcal{H},\delta)=\sqrt{\frac{8}{N}\ln(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta})}$ 表示**模型复杂度**
- 预测误差$E_{out}(g)\le E_{in}(g)+\sqrt{\frac{8}{N}\ln(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta})}$
- 误差$E_{out}(g)$和$E_{in}(g)$，模型复杂度 Ω 与VC维的关系如下：



- $d_{\mathrm{VC}}\uparrow$:　$\Omega\uparrow$,　$E_{in}\downarrow$
- $d_{\mathrm{VC}}\downarrow$:　$\Omega\downarrow$,　$E_{in}\uparrow$
- 最好的d^*_{VC}是中间某个值

3、VC维理论与模型复杂性

- 测验：若$d_{VC} \leq d+1$，下面哪个说法是正确的？

  A、存在某个或某些$d+1$数量的输入能被“打散”

  B、任何数量为$d+1$的输入都能被“打散”

  C、存在某个或某些 $d+2$ 数量的输入不能被“打散”

  ✓ D、任何数量为$d+2$的输入都不能被“打散”

本章小结

- 机器可学习性分析：$E_{\text{out}}(g) \approx E_{\text{in}}(g) \approx 0$
  - ✓ 当假设空间规模M有限时，由霍夫定不等式$P(|E_{in}(h) - E_{out}(h)| > \varepsilon) \le 2M\exp(-2\varepsilon^2 N)$可知存在$E_{in}(g) \approx E_{out}(g)$
  - ✓ 当M无限时，若假设空间$\mathcal{H}$存在break point，即其VC维有限，则其成长函数$m_{\mathcal{H}}(N)$为多项式函数，机器可学习
  - ✓ VC维d_{VC}对模型复杂度及模型误差均有影响，存在使得预测误差最小的中间值d_{VC}^*