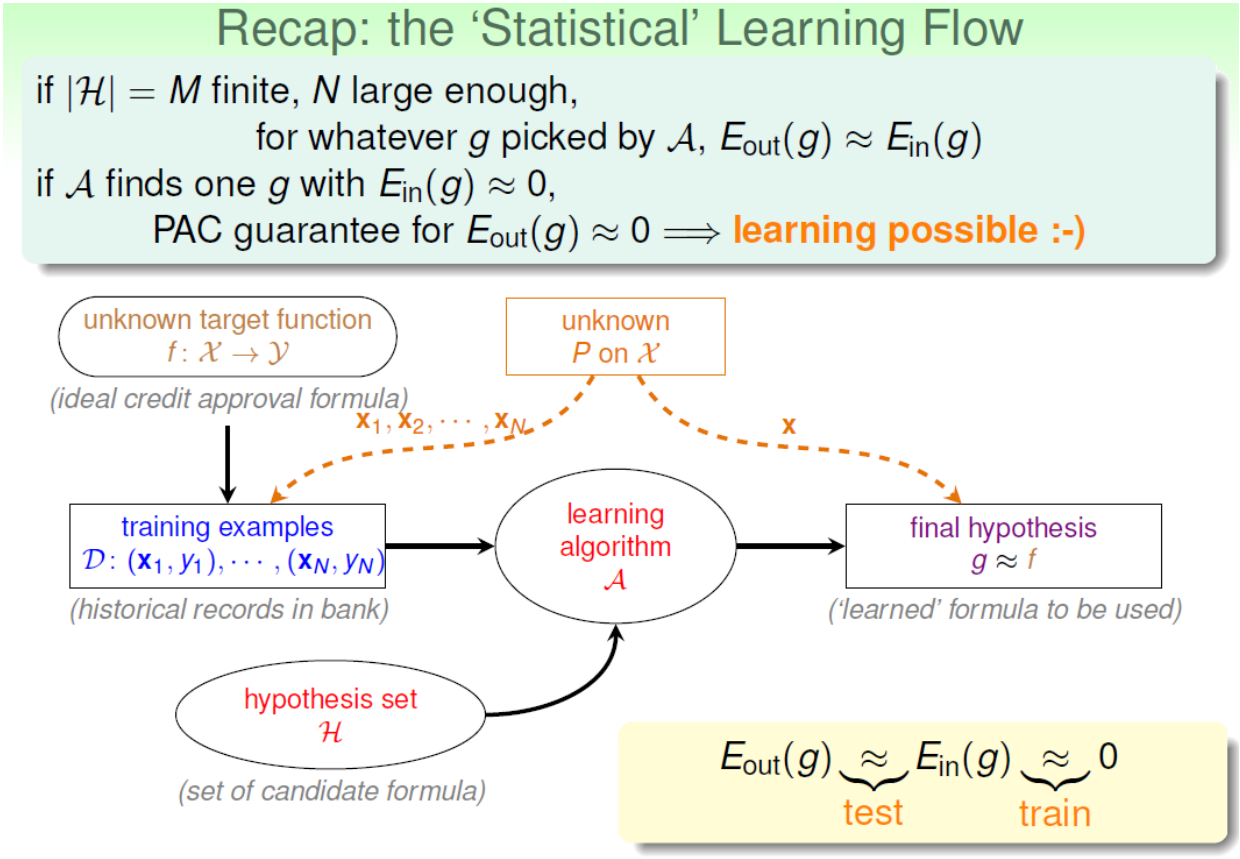
机器学习基石第5、6周笔记

5.1 Recap and Preview



机器学习能够学习的前提：
训练数据$\mathcal{D}$与测试hypothesis的数据都来自同样的一个样本分布

如果**假设集合(hypothesis set)**不太大，是有限的，而且**数据量(N)**够大的话，根据霍夫丁不等式，可以确保对每一个hypothesis而言，$E_{in} \approx E_{out}$,所以干脆选一个E_{in}最小的，如果这个$E_{in} \approx 0$,那么E_{out}也接近于0，这就达到了学习的效果。

Trade-off on M

1. can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. can we make $E_{\text{in}}(g)$ small enough?

small M

1. Yes!, $\mathbb{P}[\textbf{BAD}] \leq 2 \cdot M \cdot \exp(\ldots)$
2. No!, too few choices

large M

1. No!, $\mathbb{P}[\textbf{BAD}] \leq 2 \cdot M \cdot \exp(\ldots)$
2. Yes!, many choices

using the right M (or $\mathcal{H}$) is important
$M = \infty$ **doomed?**

这里将学习拆为两个问题：

- E_{out}与E_{in}是不是接近
- E_{in} 是不是足够小

在讨论这两个问题的时候，涉及到一个参数M，M是hypothesis set的大小，我们现在暂时说hypothesis set是有限大的，那它到底到大，我们用这个M来代表.

- 当M很小的时候，根据公式可知Bad data 发生的概率很小，那么E_{out}与E_{in}就很接近，但是算法可供选择的假设少，不一定能够找到一个很小的E_{in},使$E_{in} \approx 0$
- 当M很大的时候，算法的选择足够多，可以找到一个好的E_{in},使$E_{in} \approx 0$，但是对第一个问题而言，Bad data 发生的概率大大增加，

Known

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \leq 2 \cdot M \cdot \exp\left(-2\epsilon^2 N\right)$$

Todo

- establish **a finite quantity** that replaces M

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \overset{?}{\leq} 2 \cdot m_{\mathcal{H}} \cdot \exp\left(-2\epsilon^2 N\right)$$

- justify the feasibility of learning for infinite M
- study $m_{\mathcal{H}}$ to understand its trade-off for 'right' $\mathcal{H}$, just like M

用$m_{\mathcal{H}}$替换M,用m代表它可能比原来的大M来得小，然后用H代表它跟hypothesisset的一些性质可能是有关系的

Data size: how large do we need?

One way to use the inequality

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \leq \underbrace{2 \cdot M \cdot \exp\left(-2\epsilon^2 N\right)}_{\delta}$$

is to pick a tolerable difference ϵ as well as a tolerable **BAD** probability δ, and then gather data with size (N) large enough to achieve those tolerance criteria. Let $\epsilon = 0.1$, $\delta = 0.05$, and $M = 100$. What is the data size needed?

1. 215
2. 415
3. 615
4. 815

$$\delta = 2M * e^{-2\epsilon^2 N} \Rightarrow N = \frac{1}{2\epsilon^2} ln \frac{2M}{\delta}$$

N=415

5.2 Effective Number of Line

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \leq 2 \cdot M \cdot \exp\left(-2\epsilon^2 N\right)$$

- $\mathcal{B}$AD events $\mathcal{B}_m$: $|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon$
- to give $\mathcal{A}$ freedom of choice: bound $\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \ldots \mathcal{B}_M]$
- worst case: all $\mathcal{B}_m$ non-overlapping

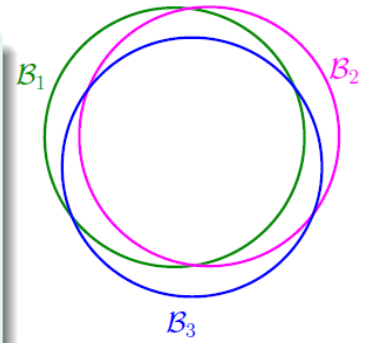
$$\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \ldots \mathcal{B}_M] \underbrace{\leq}_{\text{union bound}} \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \ldots + \mathbb{P}[\mathcal{B}_M]$$

我们用联集的方式把m个不好的事情联集起来:计算每一个不好的事情发生的几率，然后把它加加起来。但是当m无限大的时候，每次这个概率不为0的话，这个union bound 的值会大于1，这样union bound的值就没有了意义。

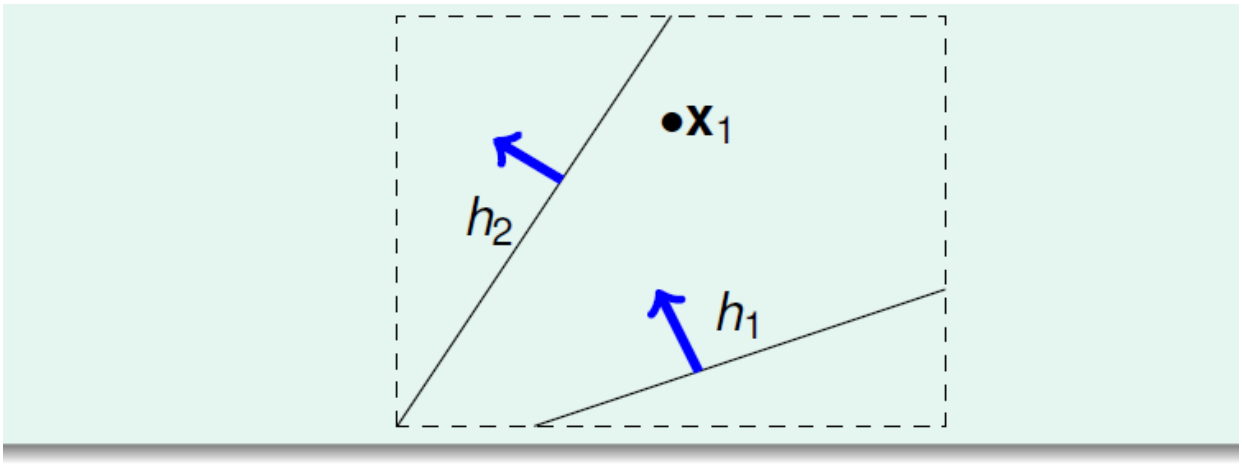
union bound $\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \ldots + \mathbb{P}[\mathcal{B}_M]$

- $\mathcal{B}$AD events $\mathcal{B}_m$: $|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon$

  overlapping for similar hypotheses $h_1 \approx h_2$
- why? (1) $E_{\text{out}}(h_1) \approx E_{\text{out}}(h_2)$
  (2) for most $\mathcal{D}$, $E_{\text{in}}(h_1) = E_{\text{in}}(h_2)$
- union bound **over-estimating**

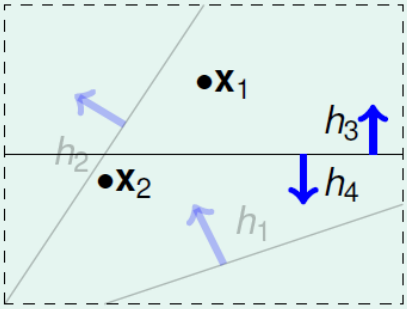


坏事情是重叠的，单纯的加和就会造成过分的估计，所以我们的目的就是将无数个hypothesis分成有限个不同的类，去除掉重叠部分

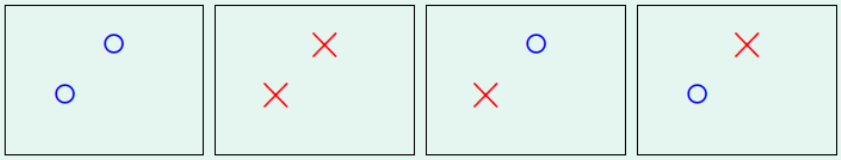


2 kinds: h_1-like($\mathbf{x}_1$) = ○ or h_2-like($\mathbf{x}_1$) = ×

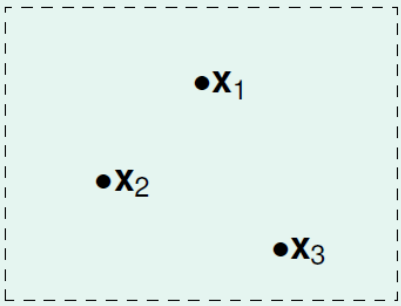
- how many **kinds of** lines if viewed from two inputs $\mathbf{x}_1, \mathbf{x}_2$?



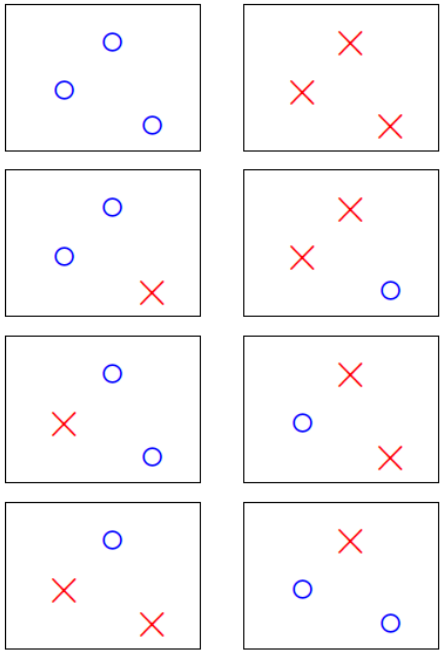
4:



for three inputs $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$

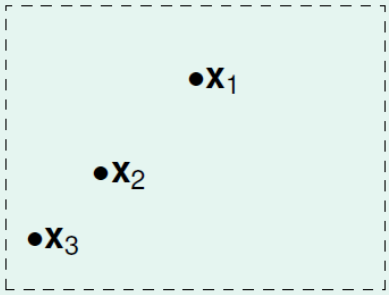


8:



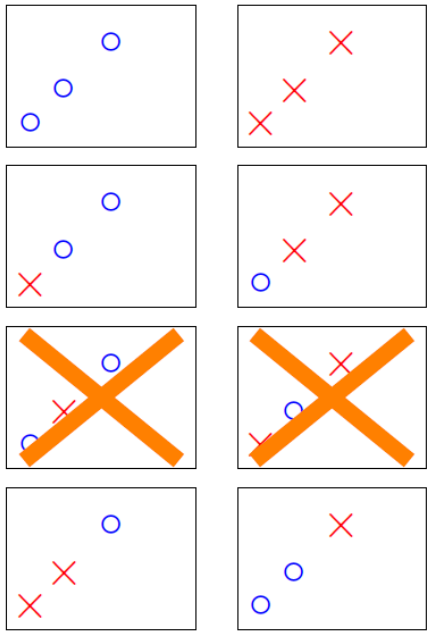
always 8 **for three inputs**?

for another three inputs x_1, x_2, x_3

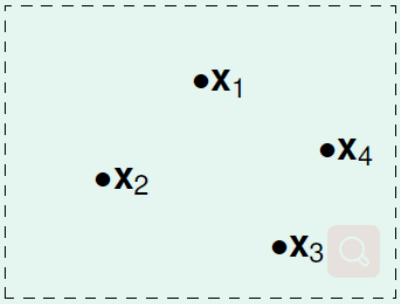


'fewer than 8' when degenerate (e.g. collinear or same inputs)

6:

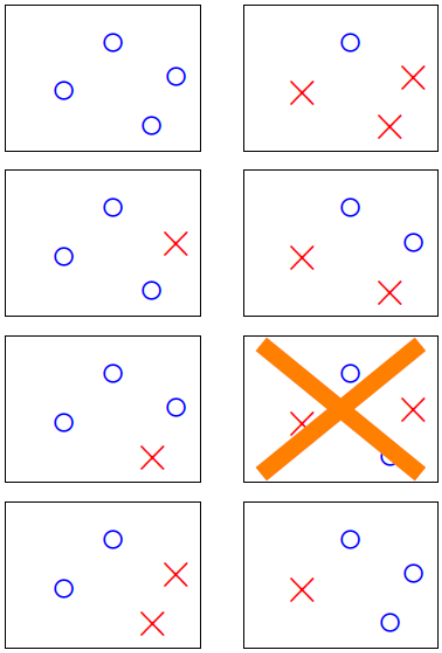


for four inputs x_1, x_2, x_3, x_4



for any four inputs
at most 14

14: 2×



Effective Number of Lines

maximum kinds of lines with respect to N inputs $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$
$\Longleftrightarrow$ **effective number of lines**

- must be $\leq 2^N$ (why?)
- finite 'grouping' of infinitely-many lines $\in \mathcal{H}$
- wish:

$$\mathbb{P}\left[\left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon\right] \leq 2 \cdot \text{effective}(N) \cdot \exp\left(-2\epsilon^2 N\right)$$

lines in 2D

N	effective(N)
1	2
2	4
3	8
4	$14 < 2^N$

if (1) effective(N) can replace M and
(2) effective(N) $\ll 2^N$
learning possible with infinite lines :-)

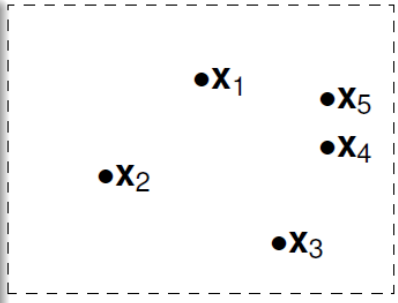
Fun Time

What is the effective number of lines for five inputs $\in \mathbb{R}^2$?

1. 14
2. 16
3. 22
4. 32

Reference Answer: (3)

If you put the inputs roughly around a circle, you can then pick any consecutive inputs to be on one side of the line, and the other inputs to be on the other side. The procedure leads to effectively 22 kinds of lines, which is **much smaller than** $2^5 = 32$. You shall find it difficult to generate more kinds by varying the inputs, and we will give a formal proof in future lectures.



0个点：1种
1个点：5种
2个点：5种

5.3 Effective Number of Hypotheses

Dichotomies: Mini-hypotheses

$$\mathcal{H} = \{\text{hypothesis } h\colon \mathcal{X} \to \{\times, \circ\}\}$$

- call

$$h(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = (h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_N)) \in \{\times, \circ\}^N$$

  a **dichotomy**: hypothesis 'limited' to the eyes of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$
- $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$:
  all dichotomies 'implemented' by $\mathcal{H}$ on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

	hypotheses $\mathcal{H}$	dichotomies $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$
e.g.	all lines in $\mathbb{R}^2$	$\{\circ\circ\circ\circ, \circ\circ\circ\times, \circ\circ\times\times, \ldots\}$
size	possibly infinite	upper bounded by 2^N

$|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)|$: candidate for **replacing** M

dichotomy：二分意思是把我们手上的点 分成两堆，圈圈的一堆，叉叉的一堆
dicotomy只对N个特定的点来做取值，代表着N个点有几种不一样的组合，所以它的上限是2^N
hypothesis是对我们的输入空间X里面所有的点取值，所以用线来区分的话，可能有无限多条线

Growth Function

- $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)|$: depend on inputs $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$
- growth function:
  remove dependence by **taking** max **of all possible** $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)|$$

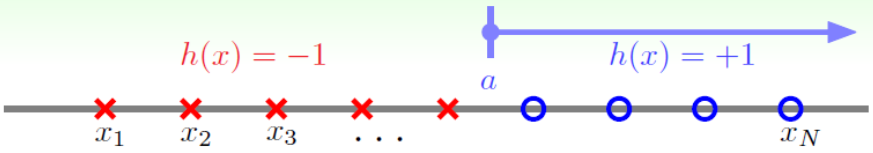
- finite, upper-bounded by 2^N

lines in 2D

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

成长函数(Growth Function)：记为$m_H(N)$,对于由N个点组成的不同集合中，某集合对应的dichotomy最大，那么这个dichotomy值就是$m_H(N)$

Growth Function for Positive Rays



- $\mathcal{X} = \mathbb{R}$ (one dimensional)
- $\mathcal{H}$ contains h, where **each** $h(x) = \text{sign}(x - a)$ **for threshold** a
- 'positive half' of 1D perceptrons

one dichotomy for $a \in$ each spot (x_n, x_{n+1}):

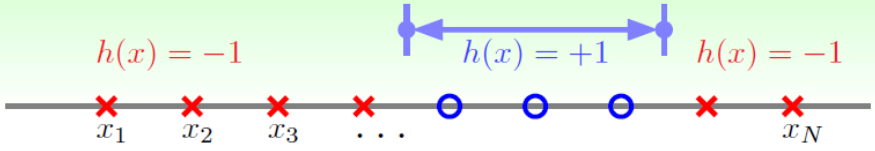
$$m_{\mathcal{H}}(N) = N + 1$$

$(N + 1) \ll 2^N$ when N large!

x_1	x_2	x_3	x_4
o	o	o	o
×	o	o	o
×	×	o	o
×	×	×	o
×	×	×	×

Training versus Testing *Effective Number of Hypotheses*

Growth Function for Positive Intervals



- $\mathcal{X} = \mathbb{R}$ (one dimensional)
- $\mathcal{H}$ contains h, where **each** $h(x) = +1$ **iff** $x \in [\ell, r)$, -1 **otherwise**

one dichotomy for each 'interval kind'

$$m_{\mathcal{H}}(N) = \underbrace{\binom{N+1}{2}}_{\text{interval ends in } N+1 \text{ spots}} + \underbrace{1}_{\text{all } \times}$$

$$= \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

$\left(\frac{1}{2}N^2 + \frac{1}{2}N + 1\right) \ll 2^N$ when N large!

x_1	x_2	x_3	x_4
o	×	×	×
o	o	×	×
o	o	o	×
o	o	o	o
×	o	×	×
×	o	o	×
×	o	o	o
×	×	o	×
×	×	o	o
×	×	×	o
×	×	×	×

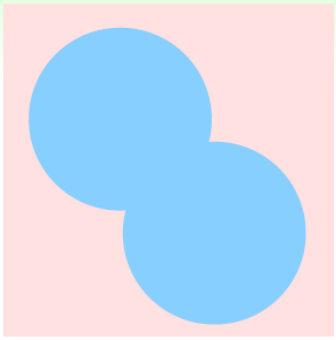
$$C_{N+1}^2 + 1 = \frac{N^2}{2} + \frac{N}{2} + 1$$

C_{N+1}^2 :从N+1个空中取两个空,放置两个端点

1：两个端点放置在同一空中，得到全为叉的情况

Growth Function for Convex Sets (1/2)

convex region in blue



non-convex region

- $\mathcal{X} = \mathbb{R}^2$ (two dimensional)
- $\mathcal{H}$ contains h, where $h(\mathbf{x}) = +1$ **iff x in a convex region,** -1 **otherwise**

what is $m_{\mathcal{H}}(N)$?

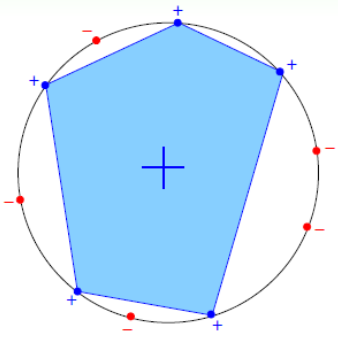
二维集合凸平面

Growth Function for Convex Sets (2/2)

- one possible set of N inputs: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ on a big circle
- **every dichotomy can be implemented** by $\mathcal{H}$ using a convex region slightly extended from contour of positive inputs

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

- call those N inputs '**shattered**' **by** $\mathcal{H}$



$m_{\mathcal{H}}(N) = 2^N \iff$
exists N inputs that can be shattered

shattered:N个点所有可能的分类情况都能够被hypotheses set覆盖

Consider positive **and negative** rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. The hypothesis set is often called **'decision stump'** to describe the shape of its hypotheses. What is the growth function $m_{\mathcal{H}}(N)$?

1. N
2. $N+1$
3. $2N$
4. 2^N

Reference Answer: 3

Two dichotomies when threshold in each of the $N-1$ 'internal' spots; two dichotomies for the all-$\circ$ and all-$\times$ cases.

$$2(N-1)+2=2N$$

N个点中有N-1个空，射线有正有负，所以有$2(N-1)$种，除此之外还有N个点之外的两个空，得到全正全负两种情况，所以一共有$2(N-1)+2=2N$种分类

5.4Break Point

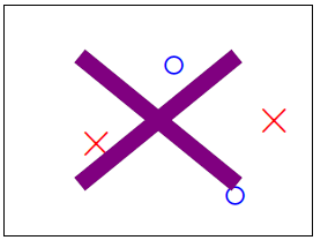
Break Point of $\mathcal{H}$

what do we know about 2D perceptrons now?

three inputs: 'exists' shatter;
four inputs, 'for all' no shatter

if no k inputs can be shattered by $\mathcal{H}$, call k a **break point** for $\mathcal{H}$

- $m_{\mathcal{H}}(k) < 2^k$
- $k+1, k+2, k+3, \ldots$ also break points!
- will study minimum break point k



2D perceptrons: **break point at** 4

2D perceptrons，我们之前分析了3个点，可以做出8种所有的dichotomy，而4个点，就无法做出所有16个点的dichotomy了。所以，我们就把4称为2D perceptrons的break point（5、6、7等都是break point）。令有k个点，如果k大于等于break point时，它的成长函数一定小于2的k次方。

The Four Break Points

- positive rays: $m_{\mathcal{H}}(N) = N + 1 = O(N)$
  break point at 2
- positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 = O(N^2)$
  break point at 3
- convex sets: $m_{\mathcal{H}}(N) = 2^N$
  no break point
- 2D perceptrons: $m_{\mathcal{H}}(N) < 2^N$ **in some cases**
  break point at 4

conjecture:

- no break point: $m_{\mathcal{H}}(N) = 2^N$ (sure!)
- break point k: $m_{\mathcal{H}}(N) = O(N^{k-1})$

excited? wait for next lecture :-)

Fun Time

Consider positive **and negative** rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. As discussed in an earlier quiz question, the growth function $m_{\mathcal{H}}(N) = 2N$. What is the minimum break point for $\mathcal{H}$?

1. 1
2. 2
3. 3
4. 4

Reference Answer: 3

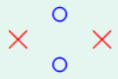
At $k = 3$, $m_{\mathcal{H}}(k) = 6$ while $2^k = 8$.

- positive rays: $m_{\mathcal{H}}(N) = N + 1$

  ○× $m_{\mathcal{H}}(2) = 3 < 2^2$: break point at 2
- positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

  ○×○ $m_{\mathcal{H}}(3) = 7 < 2^3$: break point at 3
- convex sets: $m_{\mathcal{H}}(N) = 2^N$

   $m_{\mathcal{H}}(N) = 2^N$ always: no break point
- 2D perceptrons: $m_{\mathcal{H}}(N) < 2^N$ **in some cases**

   $m_{\mathcal{H}}(4) = 14 < 2^4$: break point at 4

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible =** 3)

如果k=2，那么当N取不同值的时候，计算其成长函数$m_H(N)$是多少。很明显，当N=1时，$m_H(N) = 2$；当N=2时，由break point为2可知，任意两点都不能被shattered（shatter的意思是对N个点，能够分解为种dichotomies）；$m_H(N)$最大值只能是3；当N=3时，简单绘图分析可得其$m_H(N) = 4$，即最多只有4种dichotomies。

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

4 dichotomies , shatter any two points? **yes**

x_1	x_2	x_3
○	○	○
○	○	×
○	×	○
○	×	×

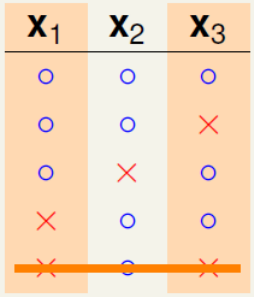
maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

4 dichotomies , shatter any two points? **no**

x_1	x_2	x_3
○	○	○
○	○	×
○	×	○
×	○	○

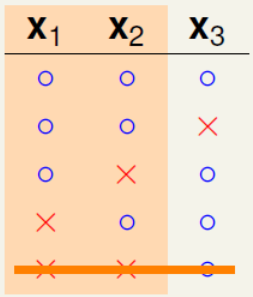
maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

5 dichotomies , shatter any two points? **yes**



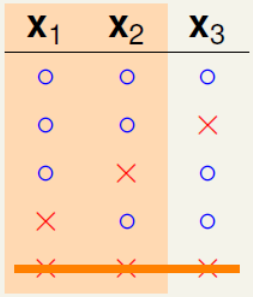
maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

5 dichotomies , shatter any two points? **yes**



maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

5 dichotomies , shatter any two points? **yes**



Restriction of Break Point (2/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible =** 3)
- $N = 3$: **maximum possible =** $4 \ll 2^3$

—break point k **restricts maximum possible** $m_{\mathcal{H}}(N)$ **a lot** for $N > k$

idea: $m_{\mathcal{H}}(N)$
$\leq$ maximum possible $m_{\mathcal{H}}(N)$ given k
$\leq$ $poly(N)$

当N>K时，break point K 限制了$m_H(N)$的大小，所以如果给定了N和K，只要证明$m_H(N)$的最大值的上界是多项式，则根据霍夫丁不等式，就能用$m_H(N)$代替M，得到机器学习是可行的。

When minimum break point $k = 1$, what is the maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$?

1. 1
2. 2
3. 4
4. 8

Reference Answer: 1

Because $k = 1$, the hypothesis set cannot even shatter one point. Thus, every 'column' of the table cannot contain both ○ and ×. Then, after including the first dichotomy, it is not possible to include any other different dichotomy. Thus, the maximum possible $m_{\mathcal{H}}(N)$ is 1.

x_1	x_2	x_3
○	×	○
○	×	×

6.2 Bounding Function: Basic Cases

Bounding Function

bounding function $B(N, k)$:

maximum possible $m_{\mathcal{H}}(N)$ when break point = k

- combinatorial quantity:
  maximum number of length-N vectors with ($\circ$, $\times$)
  while **'no shatter' any length-k** subvectors
- irrelevant of the details of $\mathcal{H}$
  e.g. $B(N, 3)$ bounds both
  - positive intervals ($k = 3$)
  - 1D perceptrons ($k = 3$)

new goal: $B(N, k) \leq poly(N)$**?**

Bounding Function(B(N,k)):当**break point=k**时，成长函数$m_H(N)$最多有多少种Dichotomy的可能.这里我们不去管成长函数到底长什么样子，而只需要关心的是在排列组合上，到底可以做出多少种排列组合。

Table of Bounding Function (1/4)

	$B(N,k)$	k=1	2	3	4	5	6	...
	1							
	2		**3**					
	3		**4**					
N	4							
	5							
	6							
	⋮							

Known

- $B(2, 2) = 3$ (maximum < 4)
- $B(3, 2) = 4$ ('pictorial' proof previously)

Table of Bounding Function (2/4)

	$B(N,k)$	k = 1	2	3	4	5	6	...
	1	**1**						
	2	**1**	3					
	3	**1**	4					
N	4	**1**						
	5	**1**						
	6	**1**						
	⋮	⋮						

Known

- $B(N,1) = 1$ (see previous quiz)

Table of Bounding Function (3/4)

	$B(N,k)$	k = 1	2	3	4	5	6	...
	1	1	**2**	**2**	**2**	**2**	**2**	...
	2	1	3	**4**	**4**	**4**	**4**	...
	3	1	4		**8**	**8**	**8**	...
N	4	1				**16**	**16**	...
	5	1					**32**	...
	6	1						...
	⋮	⋮						

Known

- $B(N,k) = 2^N$ for $N < k$
  —including all dichotomies not violating ‘breaking condition’

Table of Bounding Function (4/4)

	$B(N,k)$	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			**15**	16	16	...
	5	1				**31**	32	...
	6	1					**63**	...
	⋮	⋮						⋱

Known

- $B(N,k) = 2^N - 1$ for $N = k$
  —**removing a single dichotomy** satisfies 'breaking condition'

6.3 Bounding Function: Inductive Cases

'Achieving' Dichotomies of $B(4,3)$

after checking all 2^{2^4} sets of dichotomies, **the winner is ...**

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
01	○	○	○	○
02	×	○	○	○
03	○	×	○	○
04	○	○	×	○
05	○	○	○	×
06	×	×	○	×
07	×	○	×	○
08	×	○	○	×
09	○	×	×	○
10	○	×	○	×
11	○	○	×	×

	$B(N,k)$	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		**11**	15	16	16
	5	1				31	32
	6	1					63

how to reduce $B(4,3)$ to $B(3,?)$ cases?

首先，把B(4,3)所有情况写下来，共有11组。也就是说再加一种dichotomy，任意三点都能被shattered，11是极限。

Reorganized Dichotomies of $B(4,3)$

after checking all 2^{2^4} sets of dichotomies, **the winner is . . .**

	x_1	x_2	x_3	x_4
01	o	o	o	o
02	×	o	o	o
03	o	×	o	o
04	o	o	×	o
05	o	o	o	×
06	×	×	o	×
07	×	o	×	o
08	×	o	o	×
09	o	×	×	o
10	o	×	o	×
11	o	o	×	×

⇒

	x_1	x_2	x_3	x_4
01	o	o	o	o
05	o	o	o	×
02	×	o	o	o
08	×	o	o	×
03	o	×	o	o
10	o	×	o	×
04	o	o	×	o
11	o	o	×	×
06	×	×	o	×
07	×	o	×	o
09	o	×	×	o

orange: pair; **purple**: single

对这11种dichotomy分组，目前分成两组，分别是orange和purple，orange的特点是，x_1, x_2和x_3是一致的，x_4不同并成对，例如1和5，2和8等，purple则是单一的，x_1, x_2, x_3都不同，如6,7,9三组。

将orange去掉x_4后去重得到4个不同vector，记为α,相应的purple记为β,$B(4,3) = 2\alpha + \beta$,因为B(4,3)要求任意三点是不能shatter的，所以，所以由α, β构成的三点组合也不能shatter，即

$$\alpha + \beta \leq B(3,3)$$

Estimating Part of $B(4,3)$ (1/2)

$$B(4,3) = 11 = 2\alpha + \beta$$

	x_1	x_2	x_3
α	o	o	o
	×	o	o
	o	×	o
	o	o	×
β	×	×	o
	×	o	×
	o	×	×

- $\alpha + \beta$: dichotomies on (x_1, x_2, x_3)
- $B(4,3)$ 'no shatter' any 3 inputs $\Longrightarrow$ $\alpha + \beta$ 'no shatter' any 3

	x_1	x_2	x_3	x_4
2α	o	o	o	o
	o	o	o	×
	×	o	o	o
	×	o	o	×
	o	×	o	o
	o	×	o	×
	o	o	×	o
	o	o	×	×
β	×	×	o	×
	×	o	×	o
	o	×	×	o

$$\alpha + \beta \leq B(3,3)$$

另一方面，由于中x_4是成对存在的，且是不能被任意三点shatter的，则能推导出α是不能被任意两点shatter的。这是因为，如果α是不能被任意两点shatter，而x_4又是成对存在的，那么x_1、x_2、x_3、x_4组成的必然能被三个点shatter。这就违背了条件的设定,所以$\alpha + \beta \leq B(3,2)$

Estimating Part of $B(4,3)$ (2/2)

$$B(4,3) = 11 = 2\alpha + \beta$$

	x_1	x_2	x_3
α	o	o	o
	×	o	o
	o	×	o
	o	o	×

- α: dichotomies on (x_1, x_2, x_3) with x_4 **paired**
- $B(4,3)$ 'no shatter' any 3 inputs $\Longrightarrow$ α 'no shatter' any 2

	x_1	x_2	x_3	x_4
2α	o	o	o	o
	o	o	o	×
	×	o	o	o
	×	o	o	×
	o	×	o	o
	o	×	o	×
	o	o	×	o
	o	o	×	×
β	×	×	o	×
	×	o	×	o
	o	×	×	o

$$\alpha \leq B(3,2)$$

由此得出B(4,3)与B(3,x)的关系：

$$
\begin{aligned}
B(4,3) &= 2\alpha+\beta \\
\alpha+\beta &\le B(3,3) \\
\alpha &\le B(3,2) \\
\Rightarrow B(4,3) &\le B(3,3)+B(3,2)
\end{aligned}
$$

归纳为公式如下：

$$
\begin{aligned}
B(N,k) &= 2\alpha+\beta \\
\alpha+\beta &\le B(N-1,k) \\
\alpha &\le B(N-1,k-1) \\
\Rightarrow B(N,k) &\le B(N-1,k)+B(N-1,k-1)
\end{aligned}
$$

根据公式填表：

	$B(N,k)$	1	2	3	4	5	6
					k		
	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	11	15	16	16
	5	1	≤ 6	≤ 16	≤ 26	31	32
	6	1	≤ 7	≤ 22	≤ 42	≤ 57	63

根据递推公式，推导出B(N,K)满足下列不等式：

$$
B(N,k) \le \underbrace{\sum_{i=0}^{k-1}\binom{N}{i}}_{\text{highest term } N^{k-1}}
$$

证明：

(1)当N=0时

$$
B(0,k) \le \sum_{i=0}^{k-1} C_0^i
$$

$$
B(0,k-1) \le \sum_{i=0}^{k-2} C_0^i
$$

(2)假设下列两式成立：

$$
B(N-1,k) = \sum_{i=0}^{k-1} C_{N-1}^i
$$

$$B(N-1,k-1)=\sum_{i=0}^{k-2}C_{N-1}^{i}$$

则:

$$B(N,k)\le B(N-1,k)+B(N-1,k-1)$$
$$B(N-1,k)\le B(N-2,k)+B(N-2,k-1)$$
$$B(N-2,k)\le B(N-3,k)+B(N-3,k-1)$$
$$\vdots$$
$$B(2,k)\le B(1,k)+B(1,k-1)$$
$$B(1,k)\le B(0,k)+B(0,k-1)$$

证明 $B(1,k)$的值:

$$C_n^m=C_{n-1}^m+C_{n-1}^{m-1} \qquad (6\text{-}1)$$

$$B(0,k)=C_0^0+C_0^1+C_0^2+C_0^3\ldots+C_0^{k-2}+C_0^{k-1}$$
$$B(0,k-1)=\quad C_0^0+C_0^1+C_0^2\ldots+C_0^{k-3}+C_0^{k-2}$$

由公式6-1可得:

$$C_0^1+C_0^0=C_1^1$$
$$C_0^2+C_0^1=C_1^2$$
$$\vdots$$
$$C_0^{k-2}+C_0^{k-3}=C_1^{k-2}$$
$$C_0^{k-1}+C_0^{k-2}=C_1^{k-1}$$

所以

$$B(0,k)+B(0,k-1)=C_0^0+C_1^1+C_1^2+\ldots+C_1^{k-2}+C_1^{k-1} \quad (6\text{-}2)$$
$$=C_1^0+C_1^1+C_1^2+\ldots+C_1^{k-2}+C_1^{k-1} \quad (6\text{-}3)$$

所以 $B(1,k)=C_1^0+C_1^1+C_1^2+\ldots+C_1^{k-2}+C_1^{k-1}$

同理: $B(1,k-1)=C_1^0+C_1^1+C_1^2+\ldots+C_1^{k-2}$

根据公式6-1可得: $B(2,k)=C_2^0+C_2^1+C_2^2+\ldots+C_2^{k-2}+C_2^{k-1}$

以此类推 $B(N,k)=C_N^0+C_N^1+C_N^2+C_N^3+\ldots+C_N^{k-2}+C_N^{k-1}$

(3)由(1)、(2)可得

$$B(N,k)=\sum_{i=0}^{k-1}C_N^i$$

成立

上述不等式的右边是最高阶为k-1的N项多项式，也就是说成长函数的上界B(N,K)的上界满足多项式分布poly(N)，

6.4 A Pictorial Proof

want:

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 2 \quad m_{\mathcal{H}}(\ N) \cdot \exp\left(-2 \qquad \epsilon^2 N\right)$$

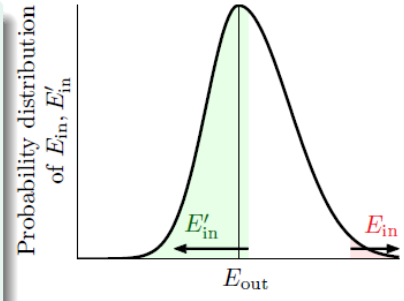
actually, when N large enough,

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 2 \cdot 2 m_{\mathcal{H}}(2N) \cdot \exp\left(-2 \cdot \frac{1}{16} \epsilon^2 N\right)$$

Step 1: Replace E_{out} by E'_{in}

$$\begin{aligned} & \frac{1}{2}\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \\ \leq \quad & \mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \end{aligned}$$

- $E_{\text{in}}(h)$ finitely many, $E_{\text{out}}(h)$ infinitely many —replace the evil E_{out} first
- how? sample verification set $\mathcal{D}'$ of size N to calculate E'_{in}
- BAD h of $E_{\text{in}} - E_{\text{out}}$ $\overset{\text{probably}}{\Longrightarrow}$ BAD h of $E_{\text{in}} - E'_{\text{in}}$



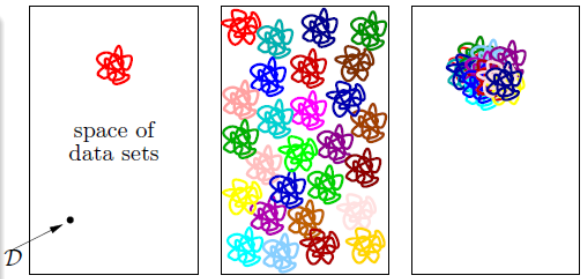
evil E_{out} removed by verification with '**ghost data**'

取另外的N个点拿来做验证，用来估计E_{out}的值 我把这另外N个点叫D'，OK，那这个用D'估计出来的值我叫E'_{in} ,当发生bad data时，E_{out}与E_{in}差很远，如果我有很大的机会抽到E'_{in}，E'_{in}与E_{in}也相差很远，那么便可以换一种坏事情：E'_{in}与E_{in}相差很远。
式子左边的$\frac{1}{2}$对应有很大的几率抽中E'_{in}与E_{in}相差很远

Step 2: Decompose $\mathcal{H}$ by Kind

$$\begin{aligned} \text{BAD} &\leq 2\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \\ &\leq 2m_{\mathcal{H}}(2N)\mathbb{P}\left[\text{fixed } h \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \end{aligned}$$

- E_{in} with $\mathcal{D}$, E'_{in} with $\mathcal{D}'$ —now $m_{\mathcal{H}}$ comes to play
- how? infinite $\mathcal{H}$ becomes $|\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}'_1, \ldots, \mathbf{x}'_N)|$ kinds
- union bound on $m_{\mathcal{H}}(2N)$ kinds



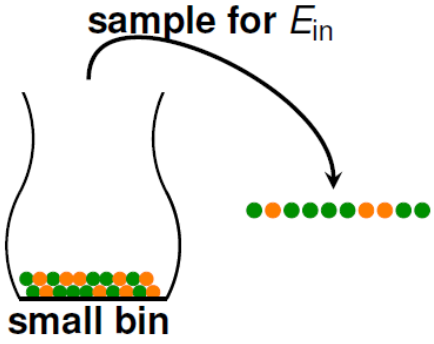
(a) Hoeffding Inequality (b) Union Bound (c) Now

use $m_{\mathcal{H}}(2N)$ to calculate BAD-overlap properly

Step 3: Use Hoeffding without Replacement

$$\begin{aligned} \text{BAD} &\leq 2m_{\mathcal{H}}(2N)\mathbb{P}\left[\text{fixed } h \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \\ &\leq 2m_{\mathcal{H}}(2N) \cdot 2\exp\left(-2\left(\frac{\epsilon}{4}\right)^2 N\right) \end{aligned}$$

- consider bin of $2N$ examples, choose N for E_{in}, leave others for E'_{in} $|E_{\text{in}} - E'_{\text{in}}| > \frac{\epsilon}{2} \Leftrightarrow \left|E_{\text{in}} - \frac{E_{\text{in}}+E'_{\text{in}}}{2}\right| > \frac{\epsilon}{4}$
- so? just 'smaller bin', 'smaller ϵ', and Hoeffding without replacement



use Hoeffding after zooming to fixed h

Vapnik-Chervonenkis (VC) bound:

$$\begin{aligned} & \mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \epsilon\Big] \\ \leq \quad & 4 m_{\mathcal{H}}(2N) \exp\left(-\frac{1}{8}\epsilon^2 N\right) \end{aligned}$$

- replace E_{out} by E'_{in}
- decompose $\mathcal{H}$ by kind
- use Hoeffding without replacement