Lecture 7. Generalisation with Finite VC Dimension

COMP90051 Statistical Machine Learning

Lecturer: Jean Honorio

THE UNIVERSITY OF MELBOURNE

This lecture

- Motivation
- Growth function
  - Considering patterns of labels possible on a data set
  - Gives good generalisation bounds provided possible patterns don't grow too fast in the data set size
- Vapnik-Chervonenkis (VC) dimension
  - Max number of points that can be labelled in all ways
  - Beyond this point, growth function is polynomial in data set size
  - Leads to famous VC generalisation theorem

Motivation

…from last lecture

A Countably Finite Model Class

- Consider we have 2 features and a countably finite set $\mathcal{F}$ of classifiers, containing:

$$f(x) = \operatorname{sgn}(x_1 + x_2) = \begin{cases} +1, \text{if } x_1 + x_2 > 0 \\ -1, \text{if } x_1 + x_2 \leq 0 \end{cases}$$
$$f(x) = \operatorname{sgn}(x_1 - x_2)$$
$$f(x) = \operatorname{sgn}(-x_1 + x_2)$$
$$f(x) = \operatorname{sgn}(-x_1 - x_2)$$
$$f(x) = \operatorname{sgn}(x_1)$$
$$f(x) = \operatorname{sgn}(-x_1)$$
$$f(x) = \operatorname{sgn}(x_2)$$
$$f(x) = \operatorname{sgn}(-x_2)$$

- Here $|\mathcal{F}| = 8$

Empirical Risk Minimisation

- Training data $\boldsymbol{D} = \{\boldsymbol{x}_1, y_1, \dots, \boldsymbol{x}_n, y_n\}$ is a random variable!
  - $(\boldsymbol{x}_i, y_i)$ i.i.d. with distribution P (unknown)
- The empirical risk of a classifier f for loss l is

$$\hat{R}_{\boldsymbol{D}}[f] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i, f(\boldsymbol{x}_i)\big)$$

- ERM: $\hat{f}_{\boldsymbol{D}}$ minimises the empirical risk

$$\hat{f}_{\boldsymbol{D}} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_{\boldsymbol{D}}[f]$$

Go trough all the $|\mathcal{F}| = 8$ classifiers and choose the best for data D

- Given f and n samples in $\boldsymbol{D}$, we can compute $\hat{R}_{\boldsymbol{D}}[f]$

True Risk

- The true risk is the expected value of the loss l
  - Intuitively speaking, the true risk is the empirical risk when using an infinite number of samples
- The true risk of a classifier f for loss l is

$$R[f] = \mathbb{E}\, l\big(Y, f(X)\big) = \int l\big(Y, f(X)\big) P(X, Y)\, dX\, dY$$

aka generalisation error (expected test error) for $l(y, y') = \begin{cases} 1, \text{if } y \neq y' \\ 0, \text{if } y = y' \end{cases}$

- Given f, we cannot compute $R[f]$ because the data distribution P is unknown

Generalisation Theorem

- For a finite model class $\mathcal{F}$, without knowing the data distribution P, with probability $\geq 1 - \delta$ over the choice of the training set D of n i.i.d. samples

$$R[\hat{f}_D] \leq \hat{R}_D[\hat{f}_D] + \sqrt{\frac{\log |\mathcal{F}| + \log(1/\delta)}{2n}}$$

☆ 在uncountable下 $|F|$无限大 → 没意义 vacuous ☆

more certain more pay

$\delta \downarrow \rightarrow \frac{1}{\delta} \uparrow \rightarrow$ bound $\uparrow$

We cannot compute $R[f]$, but we can bound it!

- The proof-sketch required upper bounding

$$\max_{f \in \mathcal{F}} \varphi_D[f] = \underbrace{\max_{f \in \mathcal{F}} \left(R[f] - \hat{R}_D[f]\right)}$$

Non-(Countably Finite) Model Class?

- Finite model class
  - Bounding uniform deviation with union bound and Hoeffding's inequality
- Consider we have 2 features and an uncountable set $\mathcal{F}$ of classifiers, containing for all $w_1 \in \mathbb{R}$, $w_2 \in \mathbb{R}$:

$$f(x) = \operatorname{sgn}(w_1 x_1 + w_2 x_2)$$

- As before, still requires upper bounding

不可数 infinite.

$$\sup_{f \in \mathcal{F}} \left(R[f] - \hat{R}_D[f]\right)$$

max 可数

Mini Summary

- No good for general (countably infinite and uncountable) cases
- Need another fundamentally new idea

Next: Organising analysis around patterns of labels possible on any data set

Growth Function

Focusing on the size of model families on data samples

Example: Decision stumps

- Consider a dataset of 6 samples, each with a single continuous feature (x) and label (y)

x	y
0	+1
4	-1
-2	+1
1	+1
-3	-1
2	-1

- We would like to find a threshold β, and then classify all samples with feature value x above β as $+1$, and feature value x below β as -1 (or viceversa)

Example: Decision stumps

- Lets sort with respect to x

x	y
0	+1
4	-1
-2	+1
1	+1
-3	-1
2	-1

sort

x	y
-3	-1
-2	+1
0	+1
1	+1
2	-1
4	-1

- Lets use the classifier:

$$f(x) = \mathrm{sgn}(x - \beta) = \begin{cases} +1, \text{if } x > \beta \\ -1, \text{if } x \leq \beta \end{cases}$$

- How to find the threshold β? Try all midpoints of x

Example: Decision stumps

- Lets use the classifier:

$$f(x) = \mathrm{sgn}(x - \beta) = \begin{cases} +1, \text{if } x > \beta \\ -1, \text{if } x \leq \beta \end{cases}$$

- Count the number of mistakes for all thresholds β

x	y	best thershold $f(x)$ 设 β是不同数据中点.				
		β=-2.5	β=-1	β=0.5	β=1.5	β=3
-3	-1	-1	-1	-1	-1	-1
-2	+1	+1	-1	-1	-1	-1
0	+1	+1	+1	-1	-1	-1
1	+1	+1	+1	+1	-1	-1
2	-1	+1	+1	+1	+1	-1
4	-1	+1	+1	+1	+1	+1
# mistakes		2	3	4	5	4

Example: Decision stumps

- Lets use the classifier:

$$f(x) = \text{sgn}(\beta - x) = \begin{cases} +1, \text{if } x < \beta \\ -1, \text{if } x \geq \beta \end{cases}$$

- Count the number of mistakes for all thresholds β

x	y	$f(x)$				
		β=-2.5	β=-1	β=0.5	β=1.5	β=3
-3	-1	+1	+1	+1	+1	+1
-2	+1	-1	+1	+1	+1	+1
0	+1	-1	-1	+1	+1	+1
1	+1	-1	-1	-1	+1	+1
2	-1	-1	-1	-1	-1	+1
4	-1	-1	-1	-1	-1	-1
# mistakes		4	3	2	1	2

Example: Decision stumps

- Thus our best decision stump classifier is

$$f(x) = \mathrm{sgn}(1.5 - x) = \begin{cases} +1, \text{if } x < 1.5 \\ -1, \text{if } x \geq 1.5 \end{cases}$$

- We consider all classifiers of the form (for all $\beta \in \mathbb{R}$)

$$f(x) = \mathrm{sgn}(x - \beta) = \begin{cases} +1, \text{if } x > \beta \\ -1, \text{if } x \leq \beta \end{cases}$$

$$f(x) = \mathrm{sgn}(\beta - x) = \begin{cases} +1, \text{if } x < \beta \\ -1, \text{if } x \geq \beta \end{cases}$$

- Although these are simple classifiers, the set of decision stump classifiers $\mathcal{F}$ is uncountable (there are as "many" as real values)

D???

β的值是随便给

Example: Growth function of Decision stumps

- Consider all possible ways we can classify data

① $f(x) = \mathrm{sgn}(x - \beta) = \begin{cases} +1, \text{if } x > \beta \\ -1, \text{if } x \le \beta \end{cases}$ ② $f(x) = \mathrm{sgn}(\beta - x) = \begin{cases} +1, \text{if } x < \beta \\ -1, \text{if } x \ge \beta \end{cases}$

①的 $-\infty$ 和 ②的 $+\infty$ 结果一样去掉；②的 $-\infty$ 和 ①的 $+\infty$ 结果一样，去掉.

x	$f(x)$					
	β=-2.5	β=-1	β=0.5	β=1.5	β=3	β=∞
-3	-1	-1	-1	-1	-1	-1
-2	+1	-1	-1	-1	-1	-1
0	+1	+1	-1	-1	-1	-1
1	+1	+1	+1	-1	-1	-1
2	+1	+1	+1	+1	-1	-1
4	+1	+1	+1	+1	+1	-1

x	$f(x)$					
	β=-2.5	β=-1	β=0.5	β=1.5	β=3	β=∞
-3	+1	+1	+1	+1	+1	+1
-2	-1	+1	+1	+1	+1	+1
0	-1	-1	+1	+1	+1	+1
1	-1	-1	-1	+1	+1	+1
2	-1	-1	-1	-1	+1	+1
4	-1	-1	-1	-1	-1	+1

- A dichotomy (in blue) is one way of classifying the 6 samples
- We have 12 unique dichotomies $2 \times 7 - \underbrace{2}_{\text{重复的}} = 12$ unique.

Dichotomies

- Given dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$ of size $|\mathcal{X}| = n$ and a classifier $f \in \mathcal{F}$, a **dichotomy** is the pattern of labels (n-dimensional vector of labels) produced by f on $\mathcal{X}$

$$\left(f(x_1), \ldots, f(x_n)\right) \in \{-1, +1\}^n.$$

每个数据点进入f里要么+1要么-1　产生不同的组合，一组就是一个 dichotomy.

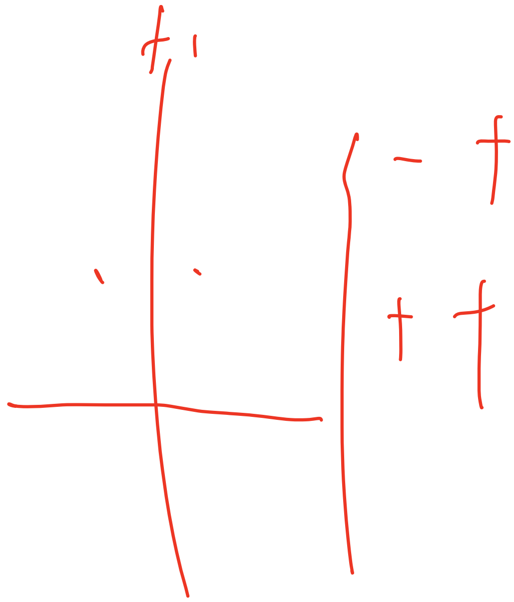
- **Unique dichotomies:** unique patterns of labels possible with all classifiers in the model class $\mathcal{F}$

$$\mathcal{F}(\mathcal{X}) = \left\{\left(f(x_1), \ldots, f(x_n)\right) : f \in \mathcal{F}\right\}$$

模型F中所能产生的所有不同的标签模式
每一种样本的所有结果
这是一个集合所有可能的结果

  - Even when $\mathcal{F}$ infinite, $|\mathcal{F}(\mathcal{X})| \leq 2^n$ (why?) $(1+1)^n$
  - For $\mathcal{F}$ countably finite, $|\mathcal{F}(\mathcal{X})| \leq |\mathcal{F}|$ (why?)

最多：每个dichotomy都不同



Growth Function

- The **growth function**

unique dichotomy 的最大值。

$$S_{\mathcal{F}}(n) = \sup_{|\boldsymbol{X}|=n} |\mathcal{F}(\boldsymbol{X})|$$

对n个样本，F可生成的最多的unique dichotomy数量

(unique dicotomy 的上界；与样本点数量有关的函数)

- is the maximum number of label patterns achievable by classifiers in the model class $\mathcal{F}$ for any set of n samples.
  * Even when $\mathcal{F}$ infinite, $S_{\mathcal{F}}(n) \leq 2^n$ (why?)
  * For $\mathcal{F}$ countably finite, $S_{\mathcal{F}}(n) \leq |\mathcal{F}|$ (why?)

$S_F(n)$ 衡量了F的表达能力（复杂性）：告诉我们F内的分类器可以有多少不同方式对样本n进行分类

Example: Growth function of Decision stumps

- In general, the set of decision stump classifiers lead to $2n$ unique dichotomies for n samples
  - We classify the n samples as -1's followed by +1's (先 个+1 - 个-1) (n+1)
  - We also classify the n samples as +1's followed by -1's (再反过来) (n+1)

有所重复的

- Thus, $S_{\mathcal{F}}(n) = 2n$ $2\times(n+1)-2=2n$
- More complex classifiers would lead to more than $2n$ unique dichotomies for n samples

对 decision stump 来说 $S_F(n)=2n$，对更复杂的分类器会更多

$S_F($ stump

Growth-Function Generalisation Theorem

- For a model class $\mathcal{F}$ with growth function $S_{\mathcal{F}}(n)$, without knowing the data distribution P, with probability $\geq 1-\delta$ over the choice of the training set D of n i.i.d. samples

$$R[\hat{f}_{\boldsymbol{D}}] \leq \hat{R}_{\boldsymbol{D}}[\hat{f}_{\boldsymbol{D}}] + \sqrt{8\frac{\log S_{\mathcal{F}}(2n) + \log(4/\delta)}{n}}$$

(Proof outside scope of COMP90051)

* $|\mathcal{F}|$ becomes $S_{\mathcal{F}}(2n)$, and few negligible extra constants
* If $S_{\mathcal{F}}(n)$ grows exponentially in n, e.g., $S_{\mathcal{F}}(n) = 2^n$ then $\frac{\log S_{\mathcal{F}}(2n)}{n} = 2$, the bound does not decay with more samples n

越多data n个 → 希望bound更紧 才学习能力强

Mini Summary

- Better to organise families by possible patterns of labels on a data set: the dichotomies of the model class
- Counting possible dichotomies gives the growth function
- Generalisation bound with growth function potentially tackles general (countably infinite and uncountable) families provided growth function is sub-exponential in data size

Next: VC dimension for a computable bound on growth functions, with the polynomial behaviour we need! Gives our final VC generalisation bound

The VC dimension

Computable, bounds growth function

Vapnik-Chervonenkis dimension

- The **VC dimension** $\mathrm{VC}(\mathcal{F})$ of a model class $\mathcal{F}$ is the largest n such that $S_{\mathcal{F}}(n) = 2^n$. (使得增长函数等于2^n的最大的n)

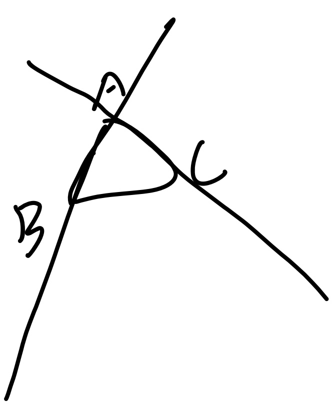
  可以等于二分法数量.

- Set of samples $\mathcal{X} = \{x_1, \ldots, x_n\}$ are shattered by $\mathcal{F}$ if $|\mathcal{F}(\mathcal{X})| = 2^n$, that is, if $\mathcal{X}$ can be classified in all possible ways

  VC维能拟合2^n种可能的组合

  F中的分类器能把X的{-1,+1}可能性都表达出来 => shatter 粉碎

- $\mathrm{VC}(\mathcal{F})$ is the size of the largest set of samples shattered by $\mathcal{F}$



Shatter: (能够分类的最大样本数量)

VC：衡量模型复杂度的指标 VC↑ => model 越复杂

→能够拟合更多的模式

Example: VC Dimension of Decision Stumps

VC dimension: the largest number n such that $S_{\mathcal{F}}(n)=2^n$

- Recall that for decision stump classifiers $S_{\mathcal{F}}(n) = 2n$
- Find the maximum n for which $2n = 2^n$
- The VC dimension is $VC(\mathcal{F}) = 2$

n	$2n$	2^n
1	2	2
2	4	4
3	6	8

- For more intuition, see the $2n$ ways of classifying n samples

n=1

+1	-1

n=2

+1	+1	-1	-1
+1	-1	+1	-1

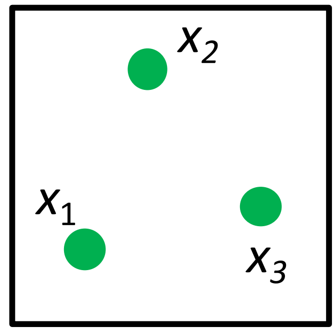
n=3

+1	+1	+1	+1	-1	-1	-1	-1
+1	+1	-1	-1	+1	+1	-1	-1
+1	-1	+1	-1	+1	-1	+1	-1

2 ways (2^3-2*3 = 2) of classifying (in red) are not -1's followed by +1's, neither +1's followed by -1's

Example 2: Growth function for linear classifiers in 2D

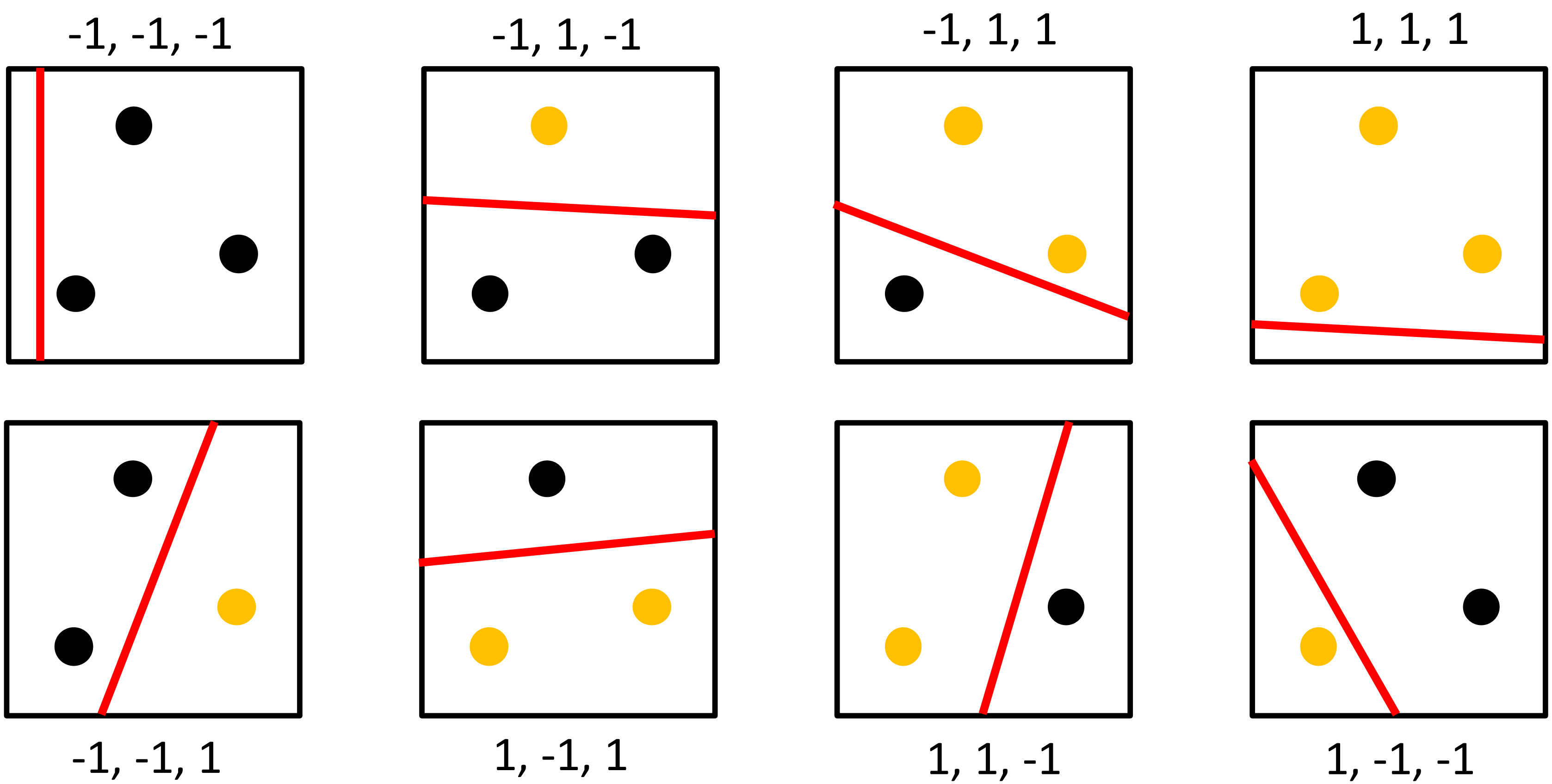
- Black means f(x)=-1
- Yellow means f(x)=1



po

$$S_{\mathcal{F}}(3) = 8$$

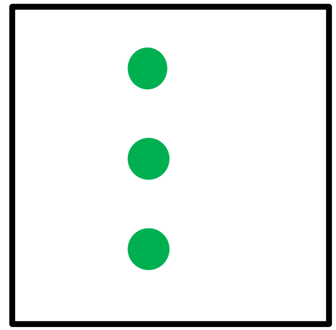
最大的 unique dicotomy 数量.



Example 2: Growth function for linear classifiers in 2D

$$\sup_{|X|=n} |F(X)| = 8$$

这个最大值



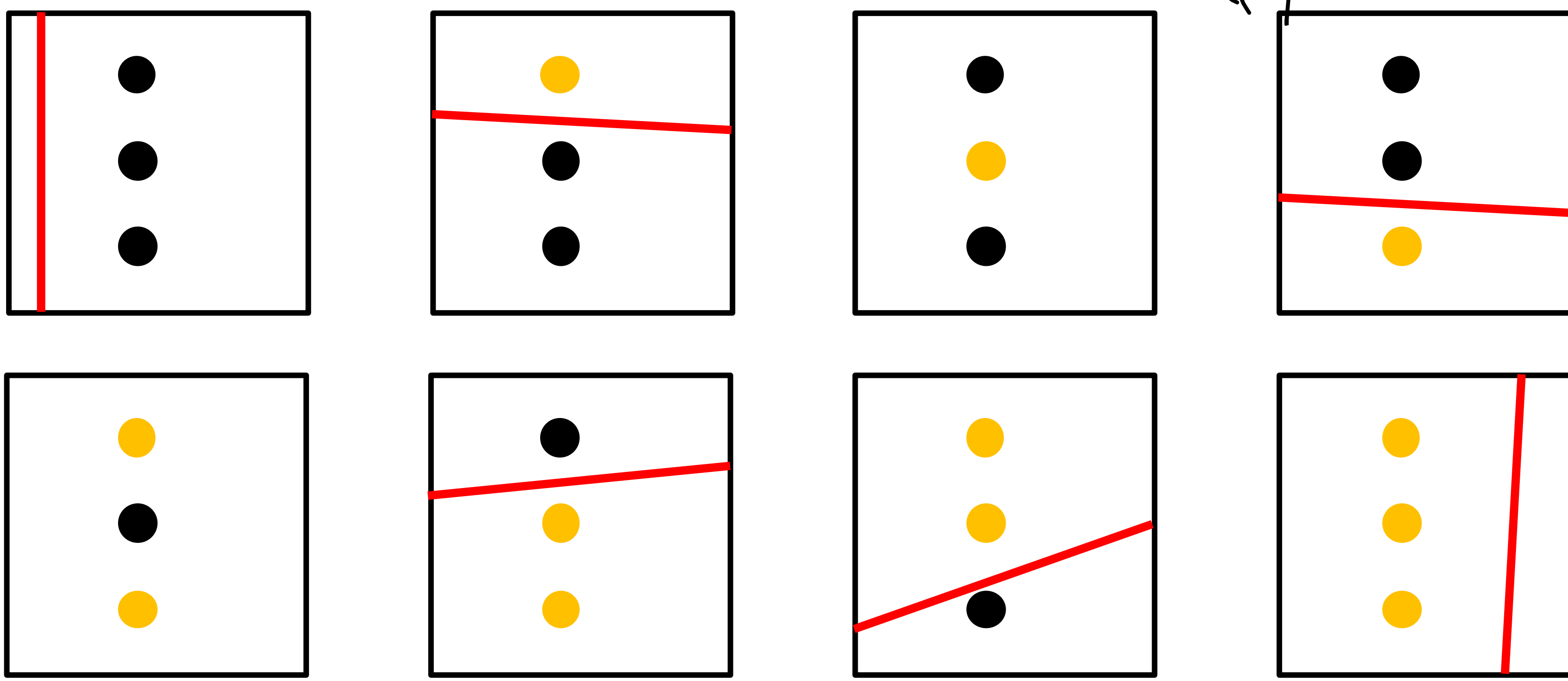
$|\mathcal{F}(\boldsymbol{X})| = 6$

but still have

$S_{\mathcal{F}}(3) = 8$

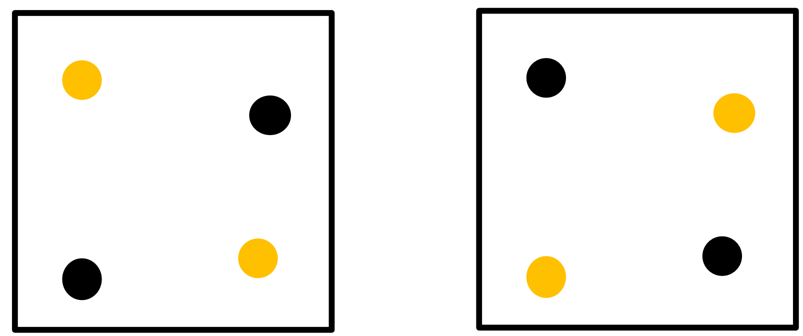
growth function is always superemum

The possible patterns should be



Example 2: Growth function for linear classifiers in 2D

- What about $n = 4$ points?
- Can never produce the criss-cross (XOR) dichotomy

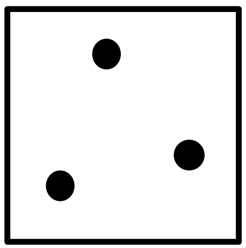


- In fact $\underbrace{S_{\mathcal{F}}(4) = 14 < 2^4}_{\text{grow}}$

小于 2^4 就不出 all possible ways 了.

Example 2: VC dimension for linear classifiers in 2D

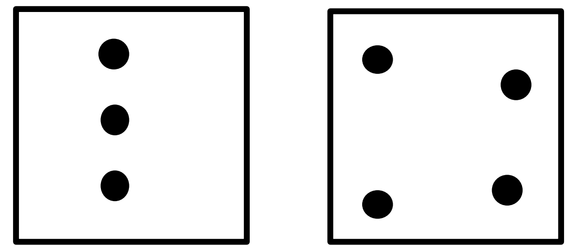
- Example: linear classifiers in $\mathbb{R}^2$, $\mathrm{VC}(\mathcal{F}) = 3$



Shattered

可以shattered的样本的最多的数量

Not shattered

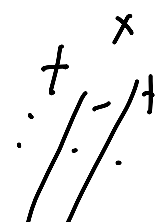


- Guess: VC dimension of linear classifiers in $\mathbb{R}^d$?

增长函数：取最大的可以 粉碎的 样本数量 n

R^d 维下的 LR的VC值 : $\Rightarrow d+1$

$d+1 \le VC \le d+2$ for hyperplane



4点

Example 3: VC dimension from dichotomies on whole domain?

① 每行都应为 unique

② 若VC=4 且 |X|=4 ⇒ 有 2^4 c VC dimension is the points that I can classify in all possible ways

③ 但现在不足16行 ⇒ VC≠4

x_1	x_2	x_3	x_4
0	0	0	0
0	1	1	0
1	0	0	1
1	1	0	1
0	1	0	0
1	0	1	0
1	1	1	1
0	0	1	1
0	1	0	1
1	1	1	0

Note we're using labels {0,1} instead of {-1,+1}. Why OK?

- Columns are *all* points in domain
- Each row is a dichotomy on entire input domain
- Obtain dichotomies on a subset of samples $\mathcal{X}' \subseteq \{x_1, \dots, x_4\}$ by: drop columns, drop dupe rows
- $\mathcal{F}$ shatters $\mathcal{X}'$ if number of rows is $2^{|\mathcal{X}'|}$

x_1	x_2	x_4
0	0	0
0	1	0
1	0	1
1	1	1
~~0~~	~~1~~	~~0~~
1	0	0
~~1~~	~~1~~	~~1~~
0	0	1
0	1	1
1	1	0

This example:

- ① Dropping column 3 leaves 8 rows behind: $\mathcal{F}$ shatters $\{x_1, x_2, x_4\}$
- Original table has $< 2^4$ rows: $\mathcal{F}$ doesn't shatter more than 3
- $\mathrm{VC}(\mathcal{F}) = 3$

$2^3 = 8$

Sauer-Shelah Lemma

- Consider any model class $\mathcal{F}$ with finite $\mathrm{VC}(\mathcal{F})$, and any sample size n. Then

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^{\mathrm{VC}(\mathcal{F})} \binom{n}{i}$$

(Proof outside scope of COMP90051)

降维 bound 以 growth function 不随 n 的增加而 decay.

- Since $\sum_{i=0}^{k} \binom{n}{i} \leq (n+1)^k$, the above implies

growth function ≤ VC

$$\log S_{\mathcal{F}}(n) \leq \mathrm{VC}(\mathcal{F}) \log(n+1)$$

VC Generalisation Theorem

- For a model class $\mathcal{F}$ with VC dimension $\mathrm{VC}(\mathcal{F})$, without knowing the data distribution P, with probability $\geq 1-\delta$ over the choice of the training set D of n i.i.d. samples

有限数值

$$R\left[\hat{f}_{\boldsymbol{D}}\right] \leq \hat{R}_{\boldsymbol{D}}\left[\hat{f}_{\boldsymbol{D}}\right] + \sqrt{8\frac{\mathrm{VC}(\mathcal{F})\log(2n+1)+\log(4/\delta)}{n}}$$

  * Proof-sketch: From the growth-function generalization theorem and since

$$\log S_{\mathcal{F}}(2n) \leq \mathrm{VC}(\mathcal{F})\log(2n+1)$$

Structural Risk Minimisation

- Choose the model class $\mathcal{F}$ with best guarantee of generalisation:

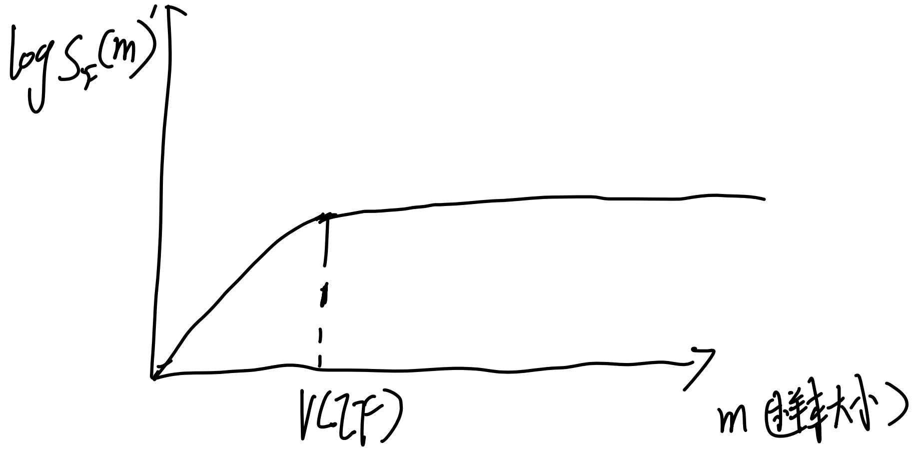
$$\underbrace{\hat{R}_{\boldsymbol{D}}\left[\hat{f}_{\boldsymbol{D}}\right]}_{\substack{\text{Large for simple classifiers,}\\ \text{small for complex classifiers}}} + \underbrace{\sqrt{8\frac{\mathrm{VC}(\mathcal{F})\log(2n+1)+\log(4/\delta)}{n}}}_{\substack{\text{Small for simple classifiers (small } \mathrm{VC}(\mathcal{F})\text{),}\\ \text{large for complex classifiers (large } \mathrm{VC}(\mathcal{F})\text{)}\\[4pt] \text{Large for small } n \text{ (few samples),}\\ \text{small for large } n \text{ (many samples)}}}$$

Mini Summary

- VC dimension is the size of the largest set of samples shattered by a model class
  * It is $d+1$ for linear classifiers in $\mathbb{R}^d$
- Sauer-Shelah: The growth function grows only polynomially in the set size beyond the VC dimension
- As a result, VC generalisation bounds true risk and empirical risk deviation by $O\left(\sqrt{(\mathrm{VC}(\mathcal{F})\log n)/n}\right)$

Much more…

- Finite VC dimension equivalent to Provably approximately correct (PAC) learning
- VC dimension is not the only tool in learning theory
  - Some problems might have infinite VC dimension
  - Other problems beyond classification
- The generalization of some methods require different complexity measures or analysis frameworks, such as:
  - Fat shattering dimension
  - Provably approximately correct (PAC) Bayes bounds
  - Rademacher complexity



随着数据↑ => 模型表示拟合能力↑ → 后期稳了

VC↑　　　　VC —

4个样本点

若unique table有9行，则 $2^4=16<9$ => 当前model class不能粉碎4 => 达不到16种 => 行只能粉碎3以内的

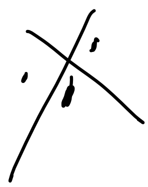
Vc dimension:

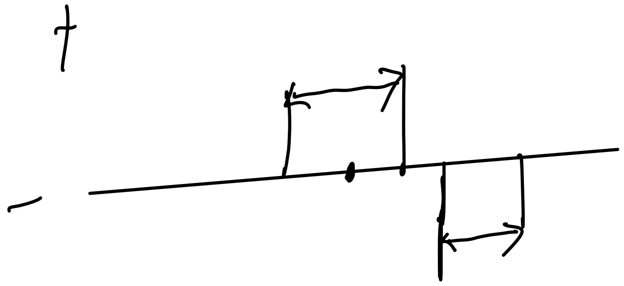
① a set of d can be shatterred. by H ✓ ① d is lower bound

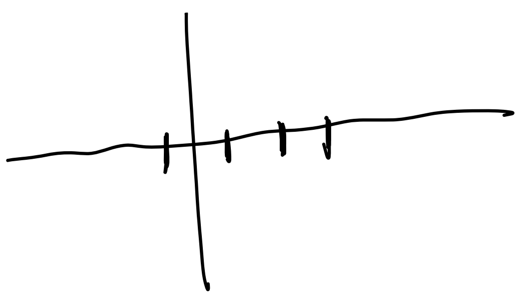
② d+1 is upper bound

H = { interval on the real line }

$X = \mathbb{R}$







H = line

$X = \mathbb{R}^2$