Lecture 7. VC Theory

COMP90051 Statistical Machine Learning

Semester 2, 2020
Lecturer: Ben Rubinstein

THE UNIVERSITY OF MELBOURNE

This lecture

- PAC learning bounds:
  - Countably infinite case works as we've done so far
  - General infinite case? Needs new ideas!
- Growth functions for the general PAC case
  - Considering patterns of labels possible on a data set
  - Gives good PAC 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, general PAC bound from VC theory
- Optional proofs at end (just for fun)

Countably infinite $\mathcal{F}$?


Josh Staiger (CCA2.0)

- Hoeffding gave us for a single $f \in \mathcal{F}$

$$\Pr\left(R[f] - \hat{R}[f] \geq \sqrt{\frac{\log\left(\frac{1}{\delta(f)}\right)}{2m}}\right) \leq \delta(f)$$

…where we're free to choose (varying) $\delta(f)$ in $[0,1]$.

- Union bound "works" (sort of) for this case

$$\Pr\left(\exists f \in \mathcal{F}, R[f] - \hat{R}[f] \geq \sqrt{\frac{\log\left(\frac{1}{\delta(f)}\right)}{2m}}\right) \leq \sum_{f \in \mathcal{F}} \delta(f)$$

Try this for finite $\mathcal{F}$ with uniform $p(f)$

- Choose confidences to sum to constant δ, then this works
  - E.g. $\delta(f) = \delta \cdot p(f)$ where $1 = \sum_{f\in\mathcal{F}} p(f)$

- By inversion: w.h.p $1-\delta$, for all f, $R[f] \leq \hat{R}[f] + \sqrt{\frac{\log\left(\frac{1}{p(f)}\right) + \log\left(\frac{1}{\delta}\right)}{2m}}$

Ok fine, but general case?

- Much of ML has continuous parameters
  - Countably infinite covers only discrete parameters ☹
- Our argument fails! ☹ ☹
  - $p(f)$ becomes a density
  - Its zero for all f. No divide by zero!
  - Need a new argument!

$$\sqrt{\frac{\log\left(\frac{1}{p(f)}\right) + \log\left(\frac{1}{\delta}\right)}{2m}}$$

- Idea introduced by VC theory: intuition
  - Don't focus on whole class $\mathcal{F}$ as if each f is different
  - Focus on differences over sample $Z_1, \ldots, Z_m$

Mini Summary

- Can eek out PAC bounds on countably infinite families using Hoeffding bound + union bound
- No good for general (uncountably infinite) cases
- Need another fundamentally new idea

Next: Organising analysis around patterns of labels possible on a data set, to avoid wort-case bad events

Growth Function

Focusing on the size of model families on data samples

Bad events: Unreasonably worst case?

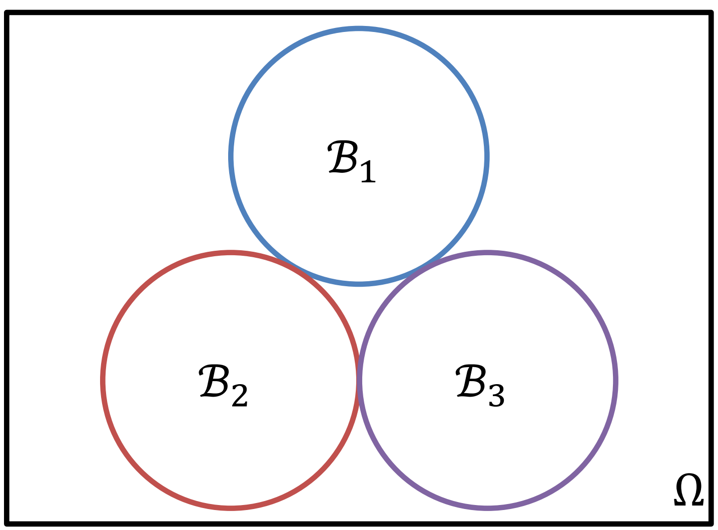
- Bad event $\mathcal{B}_i$ for model f_i

  $R[f_i] - \hat{R}[f_i] \geq \varepsilon$ with probability $\leq 2\exp(-2m\varepsilon^2)$

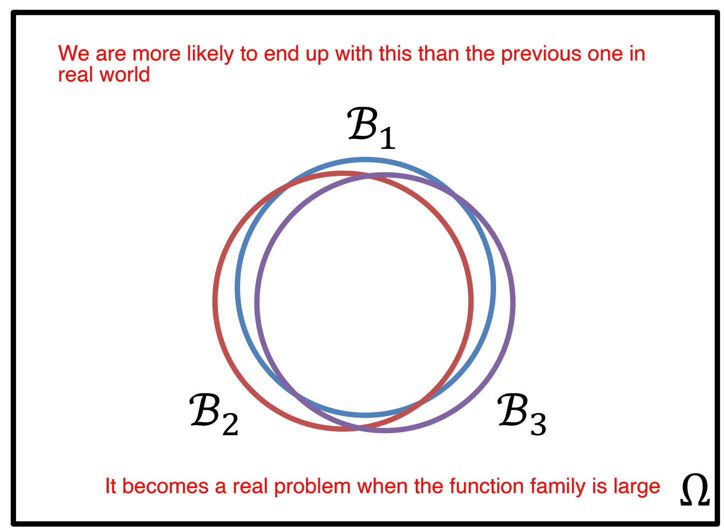
- Union bound: bad events don't overlap!?

$$\Pr\left(\mathcal{B}_1 \text{ or} \ldots \text{or } \mathcal{B}_{|\mathcal{F}|}\right) \leq \Pr(\mathcal{B}_1) + \cdots + \Pr\left(\mathcal{B}_{|\mathcal{F}|}\right) \leq 2|\mathcal{F}|\exp(-2m\varepsilon^2)$$

When they disjoint, the union bound is too conservative.



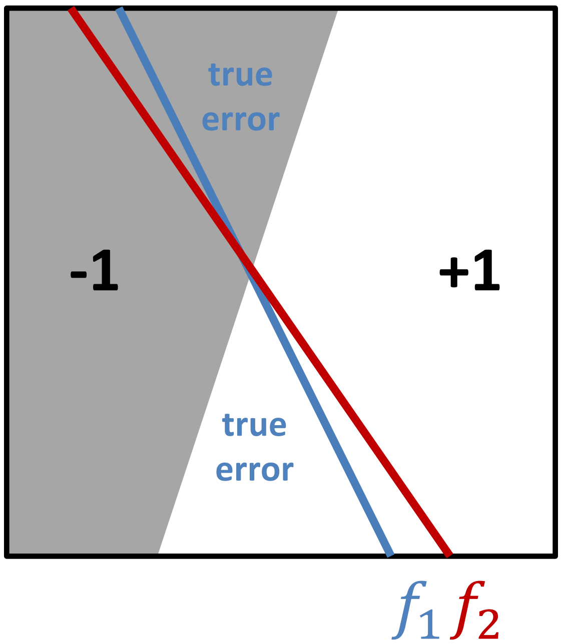
Tight bound: No overlaps



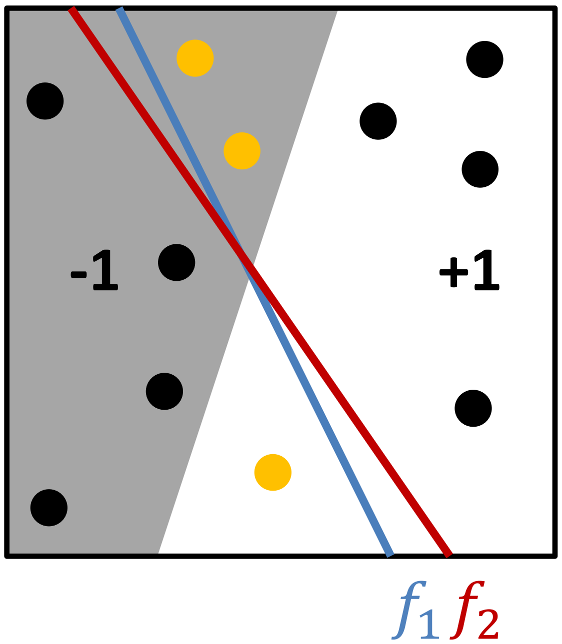
Loose bound: Overlaps

How do overlaps arise?

Instead of focusing on the whole population of models. We are going to focus on the data. We trying to examine the patterns of label that the model can make. In this case, f1 and f2 can be classified as the same pattern. In this case, we would like to have separate bad event per pattern that can be produced by our function family.



Whole of population

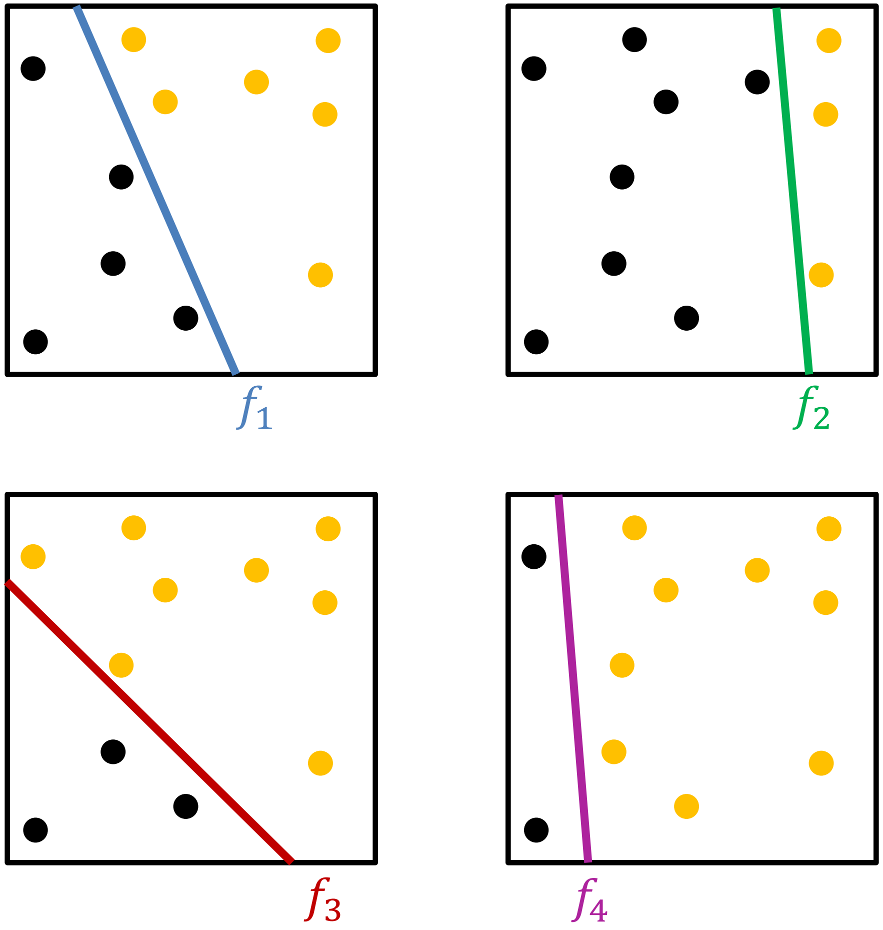


On a sample

Significantly overlapping events $\mathcal{B}_1$ and $\mathcal{B}_2$

How do overlaps arise?

VC theory focuses on the pattern of labels any $f \in \mathcal{F}$ could make



Dichotomies and Growth Function

- Definition: Given sample $x_1, \dots, x_m$ and family $\mathcal{F}$, a **dichotomy** is a $\left(f(x_1), \dots, f(x_m)\right) \in \{-1, +1\}^m$ for some $f \in \mathcal{F}$.

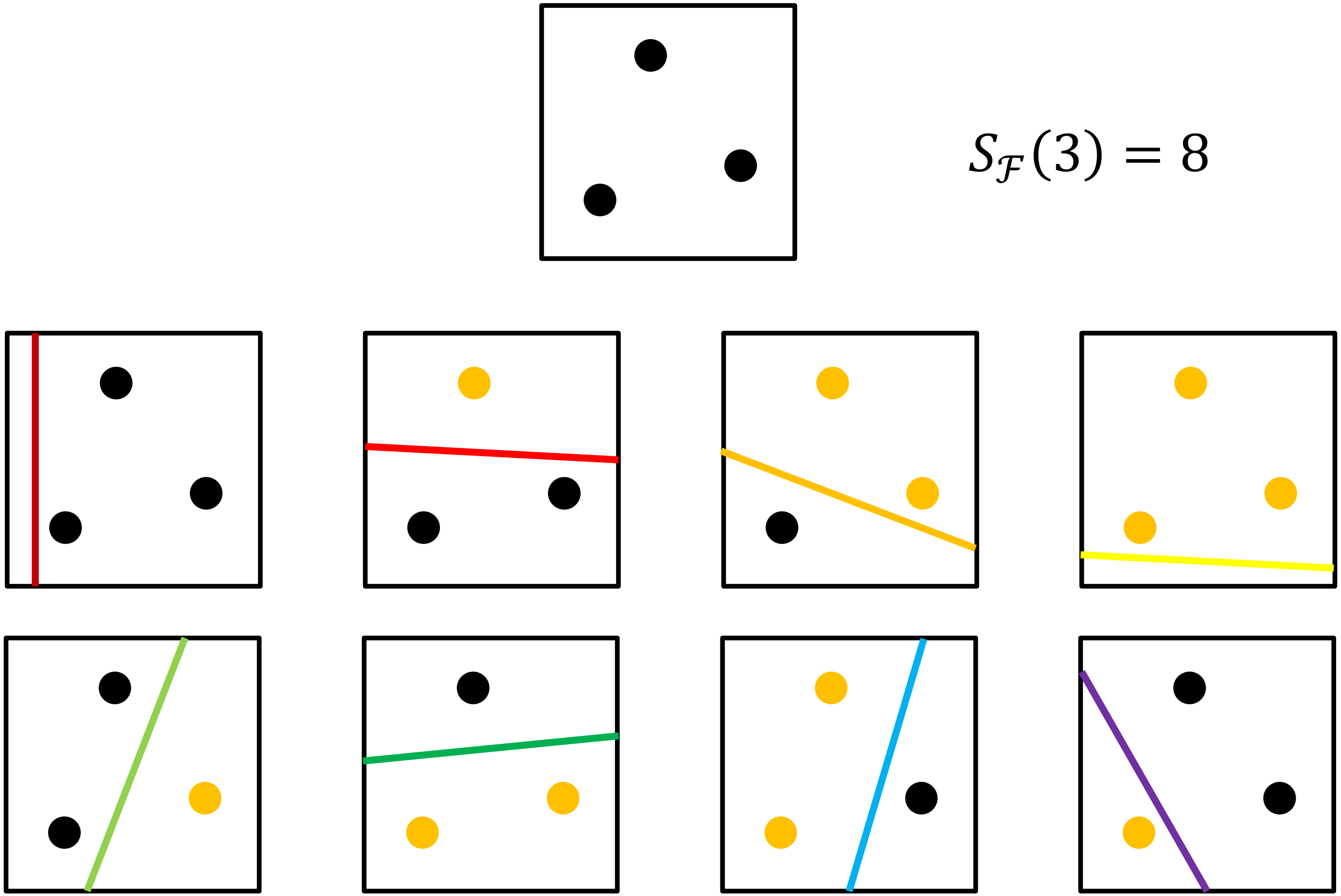
All possible labellings

- Unique dichotomies $\mathcal{F}(\mathbf{x}) = \left\{\left(f(x_1), \dots, f(x_m)\right) : f \in \mathcal{F}\right\}$, patterns of labels possible with the family
- Even when $\mathcal{F}$ infinite, $|\mathcal{F}(\mathbf{x})| \leq 2^m$ (why?)
- And also (relevant for $\mathcal{F}$ finite, tiny), $|\mathcal{F}(\mathbf{x})| \leq |\mathcal{F}|$ (why?)
- *Intuition: $|\mathcal{F}(\boldsymbol{x})|$ might replace $|\mathcal{F}|$ in union bound? How remove $\mathbf{x}$?*

In this binary classification context, always 2^(m)

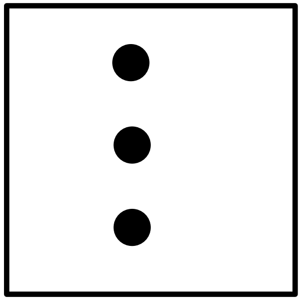
- Definition: The **growth function** $S_{\mathcal{F}}(m) = \sup_{\mathbf{x} \in \mathcal{D}^m} |\mathcal{F}(\mathbf{x})|$ is the max number of label patterns achievable by $\mathcal{F}$ for any m sample.

$S_{\mathcal{F}}(3)$ for $\mathcal{F}$ linear classifiers in 2D?

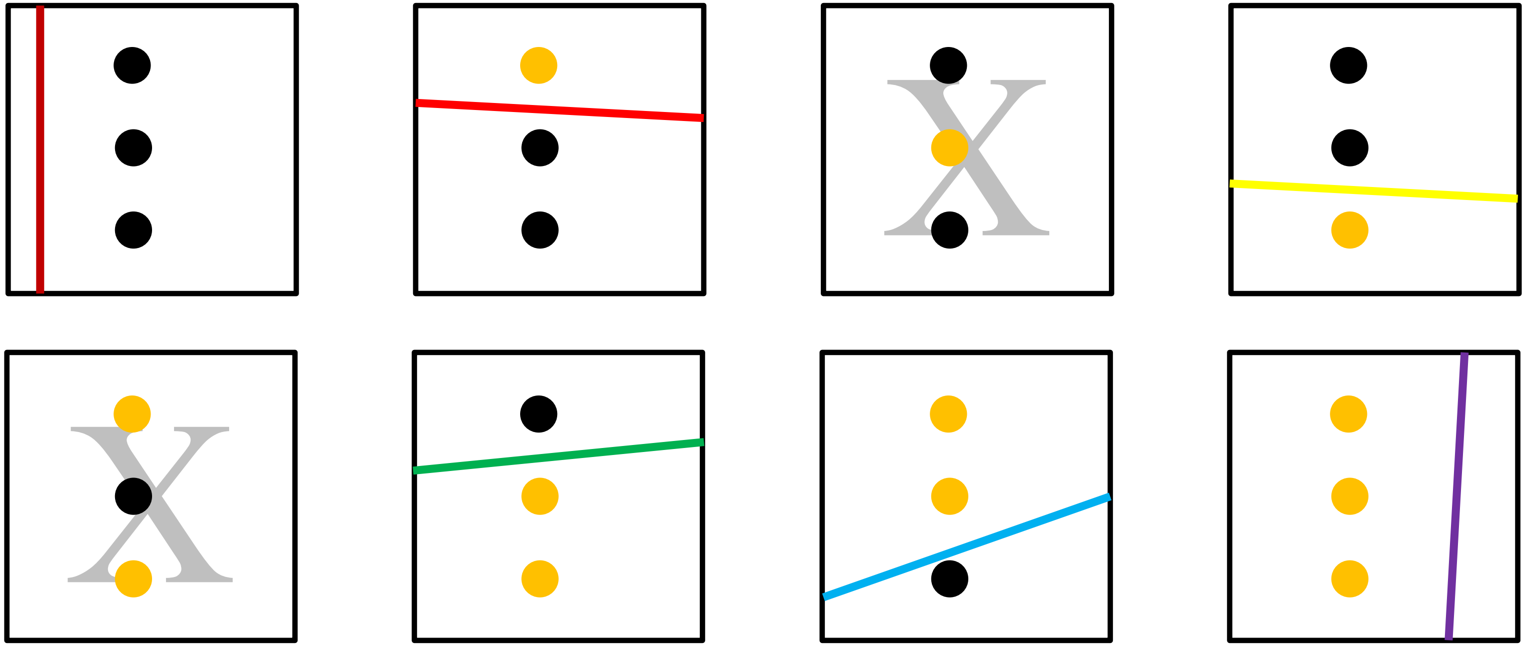


$S_{\mathcal{F}}(3) = 8$

$S_{\mathcal{F}}(3)$ for $\mathcal{F}$ linear classifiers in 2D?

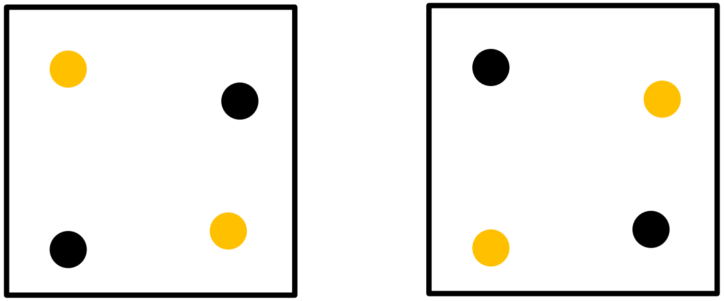


$|\mathcal{F}(\mathbf{x})| = 6$
but still have
$S_{\mathcal{F}}(3) = 8$



$S_{\mathcal{F}}(4)$ for $\mathcal{F}$ linear classifiers in 2D?

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



- In fact $S_{\mathcal{F}}(4) = 14 < 2^4$
- Guess/exercise: What about general m and dimension?

PAC Bound with Growth Function

- Theorem: Consider any $\delta > 0$ and any class $\mathcal{F}$. Then w.h.p. at least $1-\delta$: For all $f \in \mathcal{F}$

$$R[f] \leq \hat{R}[f] + 2\sqrt{2\frac{\log S_{\mathcal{F}}(2m) + \log(4/\delta)}{m}}$$

- Proof: out of scope ("only" 2-3pgs), optional reading.

- Compare to PAC bounds so far
  - A few negligible extra constants (the 2s, the 4)
  - $\mathcal{F}$ has become $\log S_{\mathcal{F}}(2m)$
  - $S_{\mathcal{F}}(m) \leq |\mathcal{F}|$, not "worse" than union bound for finite $\mathcal{F}$
  - $S_{\mathcal{F}}(m) \leq 2^m$, very bad for big family with exponential growth function gets $R[f] \leq \hat{R}[f] + Big\ Constant$. Even $R[f] \leq \hat{R}[f] + 1$ meaningless!!

Mini Summary

- The previous PAC bound approach that organises bad events by model and applies uniform bound is only tight if bad events are disjoint
- In reality some models generate overlapping bad events
- Better to organise families by possible patterns of labels on a data set: the dichotomies of the family
- Counting possible dichotomies gives the growth function
- PAC bound with growth function potentially tackles general (uncountably infinite) 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, general, PAC bound

The VC dimension

Computable, bounds growth function

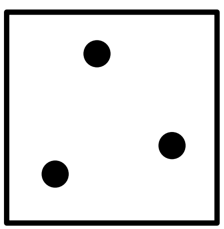
Vapnik-Chervonenkis dimension

All possible labellings

- Unique dichotomies $\mathcal{F}(\mathbf{x}) = \{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}$, patterns of labels possible with the family

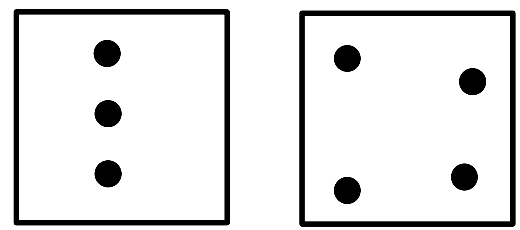
- <u>Definition</u>: The VC dimension $\mathrm{VC}(\mathcal{F})$ of a family $\mathcal{F}$ is the largest m such that $S_{\mathcal{F}}(m) = 2^m$.
  * Points $\mathbf{x} = (x_1, \ldots, x_m)$ are shattered by $\mathcal{F}$ if $|\mathcal{F}(\mathbf{x})| = 2^m$
  * So $\mathrm{VC}(\mathcal{F})$ is the size of the largest set shattered by $\mathcal{F}$

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



Shattered

Not shattered



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

Example: VC($\mathcal{F}$) from $\mathcal{F}(\mathbf{x})$ on whole domain?

10 rows

But 2^4 = 16

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 points $\mathbf{x}' \subseteq \{x_1, \dots, x_4\}$ by: drop columns, drop dupe rows
- $\mathcal{F}$ shatters $\mathbf{x}'$ if number of rows is $2^{|\mathbf{x}'|}$

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:

- 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
- $\mathrm{VC}(\mathcal{F}) = 3$

Sauer-Shelah Lemma

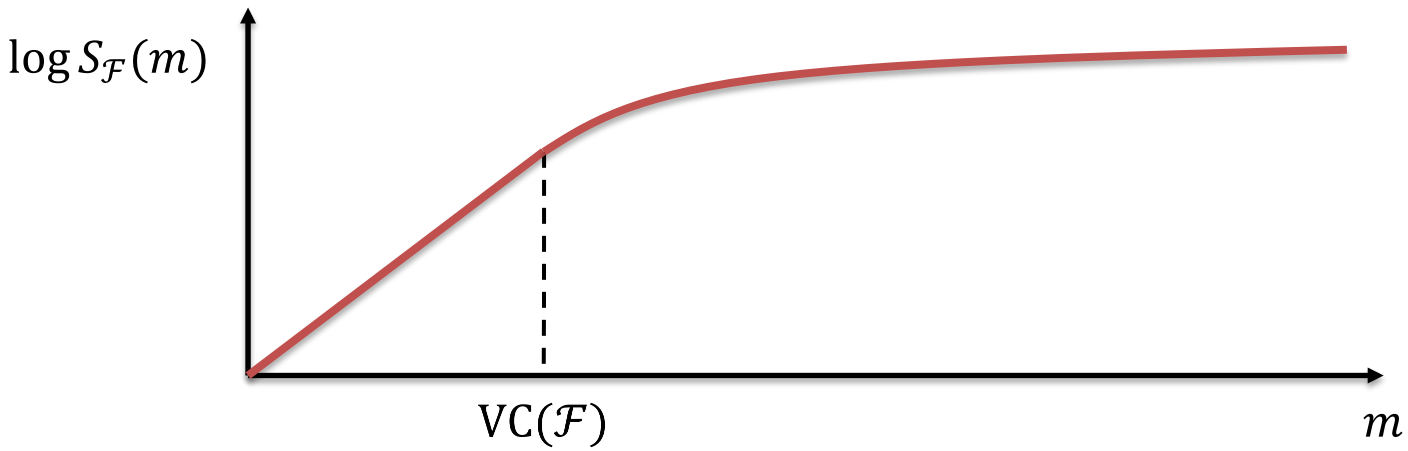
- Lemma (Sauer-Shelah): Consider any $\mathcal{F}$ with finite $\mathrm{VC}(\mathcal{F}) = k$, any sample size m. Then $S_{\mathcal{F}}(m) \leq \sum_{i=0}^{k} \binom{m}{i}$.

- From basic facts of Binomial coefficients
  - Bound is $O(m^k)$: finite VC $\Rightarrow$ eventually polynomial growth!
  - For $m \geq k$, it is bounded by $\left(\frac{em}{k}\right)^k$

- Theorem (VC bound): Consider any $\delta > 0$ and any VC-k class $\mathcal{F}$. Then w.h.p. at least $1 - \delta$: For all $f \in \mathcal{F}$

$$R[f] \leq \hat{R}[f] + 2\sqrt{2\frac{k \log \frac{2em}{k} + \log \frac{4}{\delta}}{m}}$$

VC bound big picture



- (Uniform) difference between $R[f], \hat{R}[f]$ is $O\left(\sqrt{\frac{k \log m}{m}}\right)$ down from ∞
- Limiting complexity of $\mathcal{F}$ leads to better generalisation
- VC dim, growth function measure "effective" size of $\mathcal{F}$
- VC dim doesn't count functions, but uses geometry of family: projections of family members onto possible samples
- Example: linear "gap-tolerant" classifiers (like SVMs) with "margin" Δ have $\mathrm{VC} = O(1/\Delta^2)$. Maximising "margin" reduces VC-dimension.

Mini Summary

- VC-dim is the largest set size shattered by a family
  - It is $d+1$ for linear classifiers in $\mathbb{R}^d$
  - Can calculate it on entire-domain dichotomies of a family by dropping columns and counting unique rows
- Sauer-Shelah: The growth function grows only polynomially in the set size beyond the VC-dim
- As a result, VC PAC bounds uniform risk and empirical risk deviation by $\mathrm{O}\left(\sqrt{(\mathrm{VC}(\mathcal{F})\log m)/m}\right)$

Next: Two selected proofs. Optional but beautiful.

Two Selected Proofs

Green slides: Not examinable.

Food for thought. Soul food.

Linear classifiers in d-dim: $\mathrm{VC}(\mathcal{F}) \geq d + 1$

- Goal: construct $m = d + 1$ specific points in $\mathbb{R}^d$ that are shattered by the linear classifier family
- Data in rows of $\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$ is invertible!
- Any dichotomy $y \in \{-1,1\}^{d+1}$, need $\mathbf{w}$ with $\mathrm{sign}(\mathbf{Xw}) = \mathbf{y}$
- $\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$ works*!!* $\mathrm{sign}(\mathbf{Xw}) = \mathrm{sign}(\mathbf{X}\mathbf{X}^{-1}\mathbf{y}) = \mathrm{sign}(\mathbf{y}) = \mathbf{y}$
- We've shown that $\mathcal{F}$ can shatter $d + 1$ points: $\mathrm{VC}(\mathcal{F}) \geq d + 1$

Linear classifiers in d-dim: $\mathrm{VC}(\mathcal{F}) \leq d+1$

- Goal: cannot shatter any set of $d+2$ points
- Any $\mathbf{x}_1, \dots, \mathbf{x}_{d+2}$, have more pts than dims: linear dependent

  $\mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{x}_i$, for some j, where not all a_i's are zero

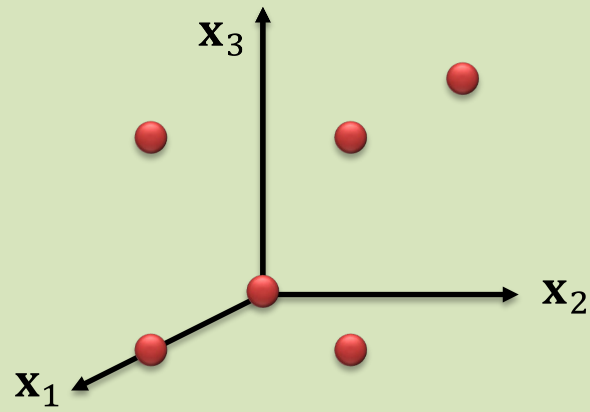
- Possible dichotomy $\mathbf{y}$? $y_i = \begin{cases} \mathrm{sign}(a_i), & \text{if } i \neq j \\ -1, & \text{if } i = j \end{cases}$
  * Suppose $\mathbf{w}$ generated $i \neq j$: $\mathrm{sign}(a_i) = \mathrm{sign}(\mathbf{w}'\mathbf{x}_i)$ so $a_i \mathbf{w}'\mathbf{x}_i > 0$
  * Can $\mathbf{w}$ generate $i = j$??
  * $\mathbf{w}'\mathbf{x}_j = \mathbf{w}' \sum_{i \neq j} a_i \mathbf{x}_i = \sum_{i \neq j} a_i \mathbf{w}'\mathbf{x}_i > 0$ so $\mathrm{sign}\left(\mathbf{w}'\mathbf{x}_j\right) \neq y_i$
- We've shown $\mathrm{VC}(\mathcal{F}) < d+2$, in other words $\mathrm{VC}(\mathcal{F}) = d+1$

Proof of Sauer-Shelah Lemma (by Haussler '95)

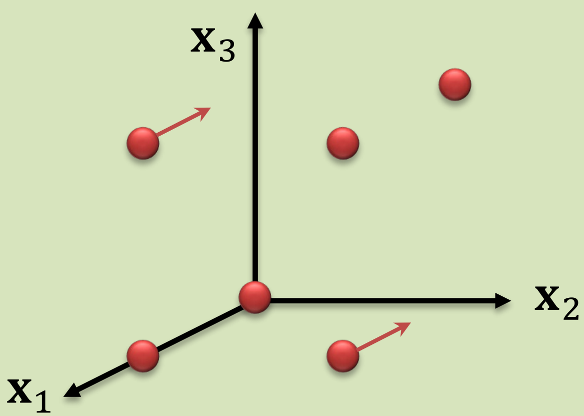
- To show that growth function $S_{\mathcal{F}}(m) \leq \sum_{i=0}^{k} \binom{m}{i}$ we prove the bound for any dichotomies $|\mathcal{F}(x_1, \ldots, x_m)|$ since $|\mathcal{F}(x_1, \ldots, x_m)| \leq S_{\mathcal{F}}(m)$
- Write $\mathbf{Y} = \mathcal{F}(x_1, \ldots, x_m) \subseteq \{0,1\}^m$, where $-1 \rightarrow 0$.
- Definition: Consider any column $1 \leq i \leq m$ and dichotomy $\mathbf{y} \in \mathbf{Y}$. The shift operator $H_i(\mathbf{y}; \mathbf{Y})$ returns $\mathbf{y}$ if there exists some $\mathbf{y}' \in \mathbf{Y}$ differing to $\mathbf{y}$ only in the i^{th} coordinate; otherwise it returns $\mathbf{y}$ with $y_i = 0$. Define $H_i(\mathbf{Y}) = \{H_i(\mathbf{y}; \mathbf{Y}) : \mathbf{y} \in \mathbf{Y}\}$ the shifting all dichotomies.
  * Intuition: Shifting along a column drops a +1 to 0 in that column so long as now other row would become duplicated.
- Definition: A set of dichotomies $\mathbf{V} \subseteq \{0,1\}^m$ is called closed below if for all $1 \leq i \leq m$, shifting does nothing $H_i(\mathbf{V}) = \mathbf{V}$.
  * Intuition: Every $\mathbf{v} \in \mathbf{V}$ has, for every $1 \leq i \leq m$ for which $v_i = 1$, some $\mathbf{u} \in \mathbf{V}$ the same as $\mathbf{v}$ except with $u_i = 1$.

Proof of Sauer-Shelah Lemma (by Haussler '95)

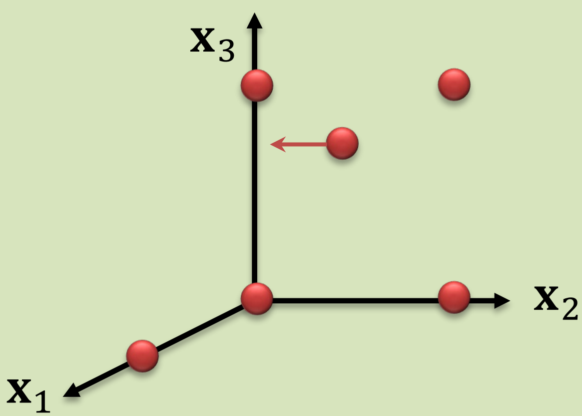
x_1	x_2	x_3
0	0	0
0	1	1
1	0	0
1	1	0
1	0	1
1	1	1



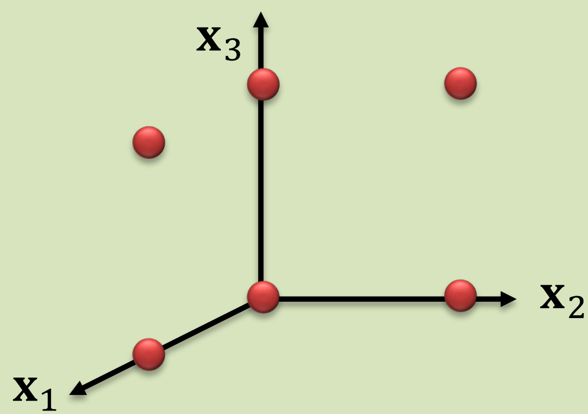
- Example set of 6 unique dichotomies on $m = 3$ pts with VC=2



Shift down along $i = 1$



Shift down along $i = 2$



Closed below

Proof of Sauer-Shelah Lemma (by Haussler '95)

- Goal: show that (1) shifting almost maintains VC dimension and cardinality all the way to a closed-below end, (2) closed-below sets have the desired Sauer-Shelah bound

- Shifting property 1: $|H_i(\mathbf{Y})| = |\mathbf{Y}|$ for any $\mathbf{Y}$.
  - Proof: no two dichotomies in $\mathbf{Y}$ shift to the same dichotomy

- Shifting property 2: $\mathrm{VC}(H_i(\mathbf{Y})) \leq \mathrm{VC}(\mathbf{Y})$ for any $i, \mathbf{Y}$.
  - Proof sketch: If $H_i(\mathbf{Y})$ shatters a subset of points, then so too does $\mathbf{Y}$

- Shifting property 3: if $\mathbf{Y}$ is closed below, then all dichotomies $\mathbf{y} \in \mathbf{Y}$ have at most $\mathrm{VC}(\mathbf{Y})$-many $y_i = 1$ (the rest 0).
  - Therefore: $|\mathbf{Y}| \leq \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{\mathrm{VC}(\mathbf{Y})}$ by counting
  - Proof sketch: if a $\mathbf{y} \in \mathbf{Y}$ had more 1s, all combinations would exist "below"

- Together: exists a shift sequence $i_1, \ldots, i_N$ to a closed below $H_{i_N}(\mathbf{Y})$:

$$|\mathbf{Y}| = \left|H_{i_1}(\mathbf{Y})\right| = \cdots = \left|H_{i_N}(\mathbf{Y})\right| \leq \sum_{i=0}^{\mathrm{VC}\left(H_{i_N}(\mathbf{Y})\right)} \binom{m}{i} \leq \cdots \leq \sum_{i=0}^{\mathrm{VC}(\mathbf{Y})} \binom{m}{i}$$

Mini Summary

- Linear classifiers in $\mathbb{R}^d$ have VC dimension $d + 1$
  - Lower bound VC-dim with specific points that are shattered
  - Upper bound VC-dim by lin. dependence of any $d + 2$ points
- Sauer-Shelah lemma bounds a family's growth function by a polynomial in VC dimension.
  - Ingenious shifting operator transforms sets of dichotomies into boundable closed-below sets
  - Along the way keeps cardinality and VC-dim controlled

Next time: Support vector machines