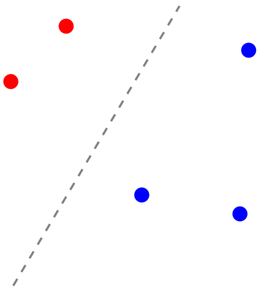
Cover's theorem(s)

Han Xiao

November 30, 2022

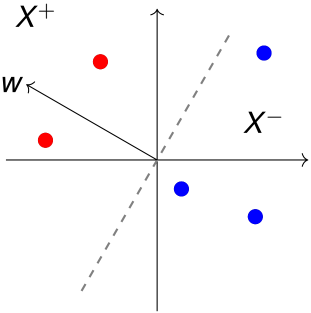
separating two sets of points



setting

- given a set of points $\mathcal{X} \in \mathbb{R}^d$
- a dichotomy (X^+, X^-) of $\mathcal{X}$ is *homogeneously linearly separable*
- iff there is a vector $w \in \mathbb{R}^d$ s.t.

$$x \cdot w > 0 \text{ if } x \in X^+$$
$$x \cdot w < 0 \text{ if } x \in X^-$$

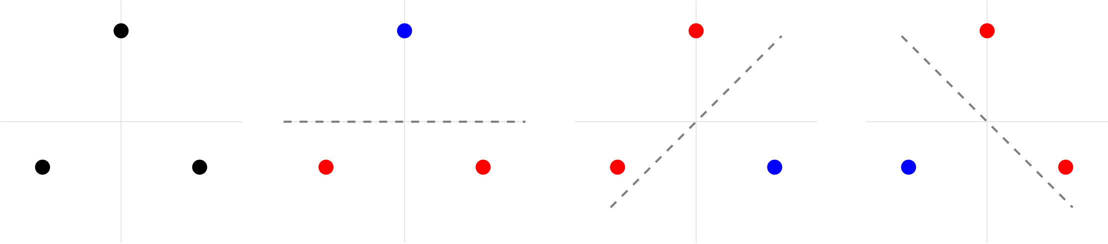


two separable dichotomies
$(\{\textcolor{blue}{\bullet}\}, \{\textcolor{red}{\bullet}\})$ and $(\{\textcolor{red}{\bullet}\}, \{\textcolor{blue}{\bullet}\})$

- the separating hyperplane is the $(d-1)$ orthogonal subspace to w
- the hyperplane must pass through the origin

warm-up 1/2

how many separable dichotomies?

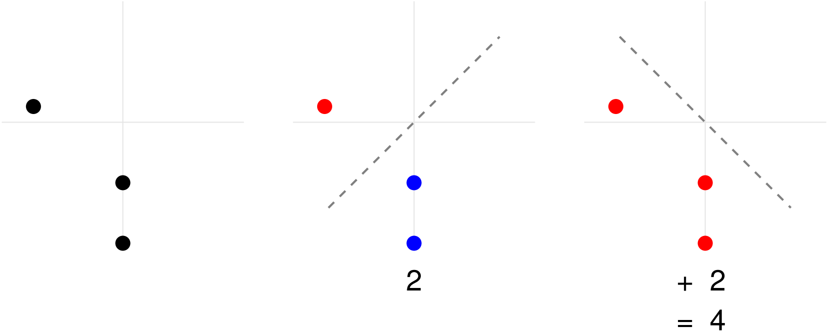


2 + 2 + 2

= 6

warm-up 2/2

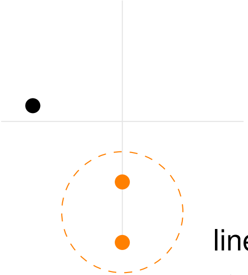
how many separable dichotomies?



Fewer than the previous example.

assumption: points are in general position

Why 4 not 6?



linearly dependent
$\Rightarrow$ **inseparable**
(by any hyperplane through origin)

$\mathcal{X}$ are in *general position* if

- every subset of size *d* or fewer are *linearly independent*

Why assuming so?

- we can analyze the **upper bound** on the number of separable dichotomies
- points are likely to be in general position (e.g., if points are uniformly distributed)

main character – $C(N, d)$

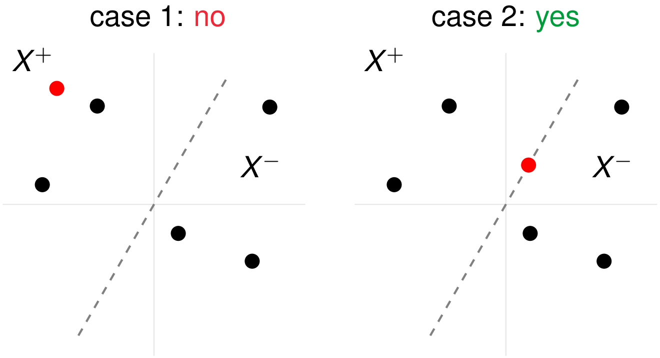
$C(N, d) = \#$ of linearly separable dichotomies of N in $\mathbb{R}^d$

(assuming $\mathcal{X}$ in general position)

Question: does $C(N, d)$ have a closed-form formula?

analysis by induction 1/6

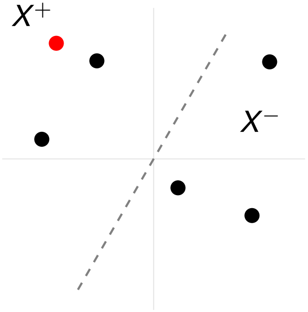
- assume we have N points $\mathcal{X}$ and add a new point y
- for each separable (X^+, X^-) of $\mathcal{X}$, there are two possibilities of y, depending on
- if there is a separating hyperplane w of (X^+, X^-) s.t. $w \cdot y = 0$



analysis by induction 2/6

case 1

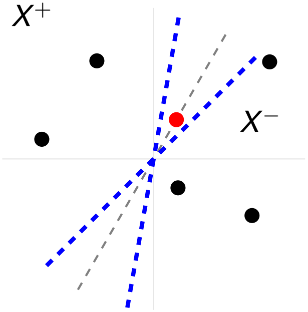
- i.e., there is **no** such separating hyperplane w s.t. $w \cdot y = 0$
- either $(X^+ \cup \{y\}, X^-)$ or $(X^+, X^- \cup \{y\})$ is separable
- $\Rightarrow$ 1 separable dichotomy for $\mathcal{X} \cup \{y\}$.



analysis by induction 3/6

case 2

- i.e., there is such separating hyperplane w s.t. $w \cdot y = 0$
- **both** $(X^+ \cup \{y\}, X^-)$ and $(X^+, X^- \cup \{y\})$ are separable
- $\Rightarrow$ 2 separable dichotomies for $\mathcal{X} \cup \{y\}$.



analysis 4/6: combining case 1 and case 2

- let D be the number of "case-2" dichotomies.

$$C(N+1,d) = \underbrace{C(N,d) - D}_{1\times\#\text{ of case 1 dich.}} + \underbrace{2\,D}_{2\times\#\text{ of case 2 dich.}}$$
$$= C(N,d) + D$$

- question: what is D?

analysis 5/6: what is D?

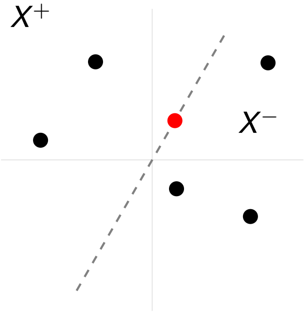
recall D is the number of (X^+, X^-) for which there is a separating w s.t. $w \cdot y = 0$.

Lemma

$(X^+ \cup \{y\}, X^-)$ and $(X^+, X^- \cup \{y\})$ are both separable in $\mathbb{R}^d$

$$\Leftrightarrow$$

(X^+, X^-) is separable by a $(d-1)$-dimensional space containing y

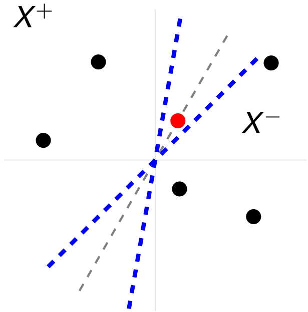


$\Rightarrow D \; = \; C(N, d-1)$ $\leftarrow$ the key!

proof of lemma 1/2

$(X^+ \cup \{y\}, X^-)$ and $(X^+, X^- \cup \{y\})$ are separable $\Leftarrow$ (X^+, X^-) is separable by a $(d-1)$-dimensional space containing y

Simple, recall that the hyperplane can be shifted either way



proof of lemma 2/2

$(X^+ \cup \{y\}, X^-)$ and $(X^+, X^- \cup \{y\})$ are separable $\Rightarrow$ (X^+, X^-) is separable by a $(d-1)$-dimensional space containing y

- let w_1 by a hyperplane that separates $(X^+ \cup \{y\}, X^-)$
  - $w_1 \cdot y > 0$
- let w_2 by a hyperplane that separates $(X^+, X^- \cup \{y\})$
  - $w_2 \cdot y < 0$
- let $w^* = (-w_2 \cdot y)w_1 + (w_1 \cdot y)w_2$
- fact 1: $w^* \cdot y = 0$

  $\rightarrow$ y is contained by the hyperplane w^*

  $\rightarrow$ what is this hyperplane? the subspace orthogonal to y!
- fact 2: $w^* \cdot x > 0$ for $x \in X^+$ and $w^* \cdot x < 0$ for $x \in X^-$

  $\rightarrow$ (X^+, X^-) is separated by w^*

analysis 6/6

$$C(N, d) = C(N-1, d) + C(N-1, d-1)$$

...... expand recursively ......

$$= \sum_{k=0}^{N-1} \binom{N-1}{k} C(1, d-k)$$

note that:

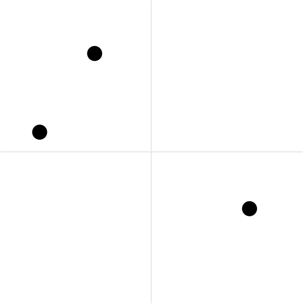
$$C(1, m) = \begin{cases} 2, & m \geq 1 \\ 0, & m < 1 \end{cases}$$

$$\Rightarrow \quad C(N, d) = 2 \sum_{k=0}^{d-1} \binom{N-1}{k} \quad \longleftarrow \quad \text{Cover's theorem}$$

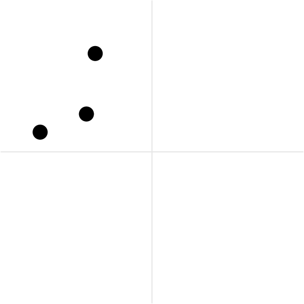
verify it!

- $C(N, d) = \sum_{k=0}^{N-1} \binom{N-1}{k} C(1, d-k)$
- $C(3, 2) = \binom{2}{1} C(1, 1) + \binom{2}{0} C(1, 2) = 3 \times 2 = 6$
- $C(4, 2) = \binom{3}{1} C(1, 1) + \binom{3}{0} C(1, 2) = 4 \times 2 = 8$

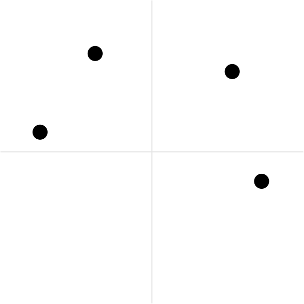
$d = 2, N = 3$



$d = 2, N = 3$



$d = 2, N = 4$



generalization of Cover's theorem

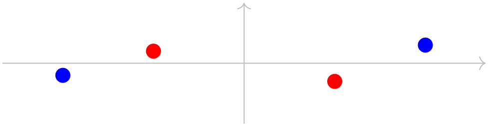
Theorem

if the hyperplane is constrained to contain k linearly independent points $\{y_1, \ldots, y_k\}$, then there are $C(N, d-k)$ separable dichotomies of $\mathcal{X}$.

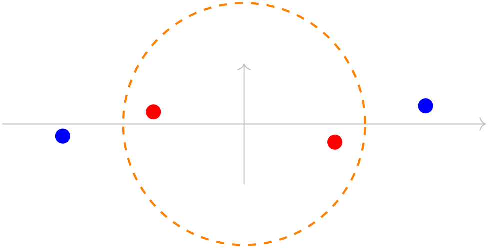
an assumption: projection of $\mathcal{X}$ onto the orthogonal subspace of $S(\{y_1, \ldots, y_k\})$ are in general position, where $S(\{\cdot\})$ is the space spanned by $\{\cdot\}$.

generalization to arbitrary surfaces

- say $\mathcal{X}$ are in $\mathbb{R}^m$, for now, we only considered *linear separability* in $\mathbb{R}^m$.
- what if $\mathcal{X}$ are not linearly separable in $\mathbb{R}^m$?



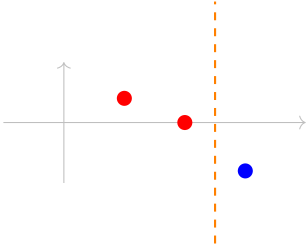
- can we transform them into *a higher-dimensional space* $\mathbb{R}^d$ so that they *are* linearly separable?
- the separating surface in $\mathbb{R}^m$ is possibly *non-linear*.



generalization to arbitrary surfaces

- assume each point $x \in \mathbb{R}^m$ is transformed into $\mathbb{R}^d$ where $d > m$
- by some function $\phi : \mathbb{R}^m \to \mathbb{R}^d$

arbitrary surface examples 1/2

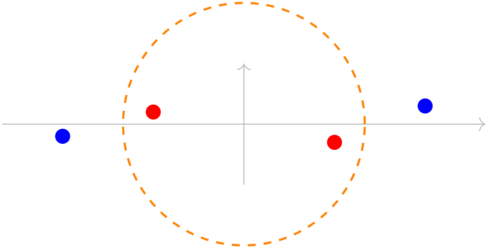


What if $\phi(x) = (1, x)$?

$\Rightarrow$

separating surfaces = hyperplanes not necessarily passing through origin

degree of freedom: $d + 1$



What if $\phi(x) = (1, x, ||x||^2)$?

$\Rightarrow$

separating surfaces = hyperspheres

degree of freedom: $d + 2$

arbitrary surface examples 2/2

- let ϕ be all r-wise products of x_i
- i.e., $\phi(x) = \left(1, x_1, x_2, \ldots, x_m, (x_1)^2, x_1 x_2, \ldots, x_i x_j, \ldots, x_m^r\right)$
- the surface is called a rational rth-order variety
- degree of freedom: $\binom{m+r}{r}$

generalization to arbitrary surfaces: a summary

def. of $\phi(x)$	separating surface	degree of freedom	number of separable dichotomies	separating capacity
(x)	hyperplane through origin	m	$C(N, m)$	$2m$
$(1, x)$	hyperplane	$m+1$	$C(N, m+1)$	$2\,(m+1)$
$(1, x, \|\|x\|\|_2)$	hypersphere	$m+2$	$C(N, m+2)$	$2\,(m+2)$
$(x, \|\|x\|\|_2)$	hypercone	$m+1$	$C(N, m+1)$	$2\,(m+1)$
all r-wise products of x_i	rational r-order variety	$\binom{m+r}{r}$	$C(N, \binom{m+r}{r})$	$2\left(\binom{m+r}{r}\right)$

let’s go back to $\mathbb{R}^d$ and consider linear separability again.

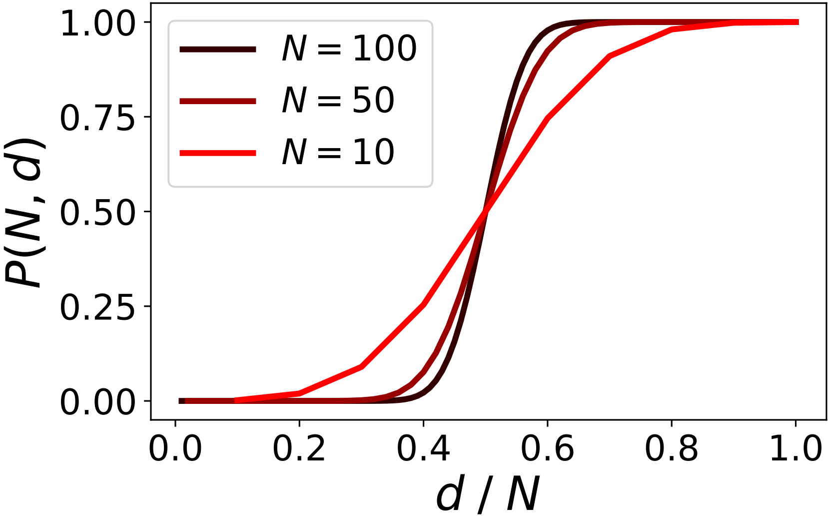
probability of being separable

- let $P(N, d)$ be the probability of a random dichotomy being linearly separable
- further assume each dichotomy has equal chance of being drawn$\rightarrow$ $1/2^N$

$$
\begin{aligned}
P(N, d) &= \sum_{(X^+, X^-)} \left(\frac{1}{2}\right)^N \mathbb{1}\left[(X^+, X^-) \text{ is separable}\right] \\
&= \left(\frac{1}{2}\right)^N C(N, d) \\
&= \left(\frac{1}{2}\right)^{N-1} \sum_{k=0}^{d-1} \binom{N-1}{k}
\end{aligned}
$$

- What is $P(N, d)$?
- $\Rightarrow$ cumulative binomial distribution
  i.e., $N-1$ flips of a fair coin resulting in $d-1$ or fewer heads.

What does $P(N, d)$ look like?

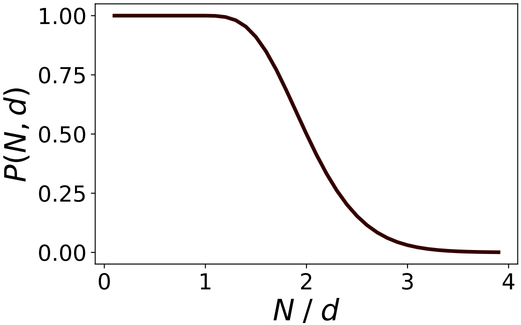


how many points can a d-dimensional space naturally have

s.t. $P(N, d)$ is high?

$\Rightarrow$ separating capacity

separating capacity 1/2



- given some integer n, consider

$$P(n, d) - P(n+1, d) = \left(\frac{1}{2}\right)^n \binom{n-1}{d-1}$$

- interpretation: speed of change w.r.t. n at n ("derivative")
- $\rightarrow$ negative binomial distribution!

Separating capacity 2/2

- asymptotic behaviour, for $\epsilon > 0$

$$\lim_{d\to\infty} P(2d(1+\epsilon), d) = 0$$
$$\lim_{d\to\infty} P(2d, d) = \frac{1}{2}$$
$$\lim_{d\to\infty} P(2d(1+\epsilon), d) = 1$$

- $2d$ is the **separating capacity** of a surface family having d *degrees of freedom*.

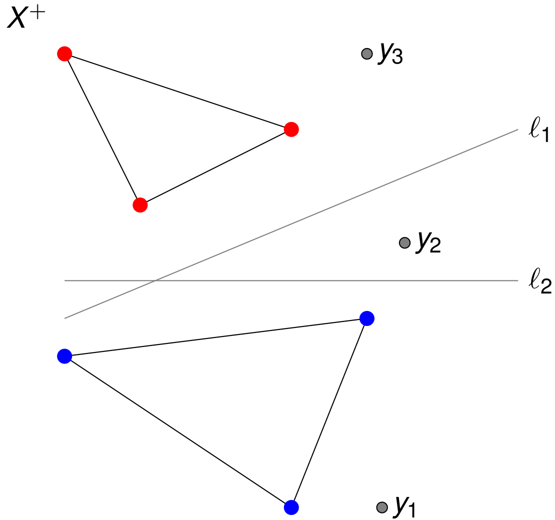
Implications on classification – generalization and learning

generalization and learning: setting

- assume we have a binary classification problem
- with training set (X^+, X^-)
- a new point y is said to be ambiguous w.r.t. a family of surfaces
- if there exists one surface inducing the dichotomy $(X^+ \cup \{y\}, X^-)$
- and there exists another surface inducing the dichotomy $(X^+, X^- \cup \{y\})$

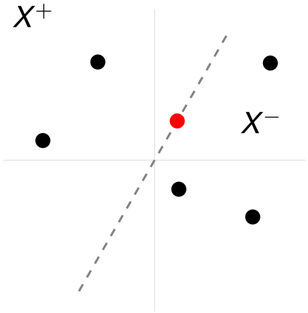
two classifiers trained on (X^+, X^-) give different predictions on y.

an example of being ambiguous



condition of ambiguity

- when is y ambiguous?
- when there exists a separating surface containing y!



probability of ambiguity

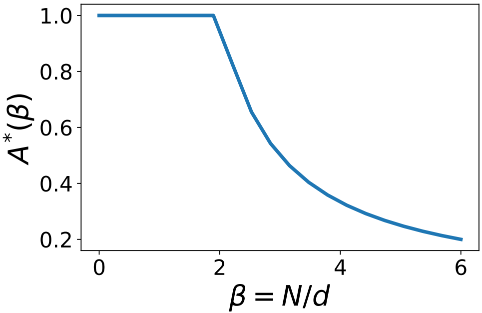
- what is the probability that y is ambiguous, given a training set of N points in $\mathbb{R}^d$?
- let $A(N, d)$ denote this probability

$$A(N, d) = \frac{\text{\# of separable dich. containing } y}{\text{\# of separable dich.}}$$
$$= \frac{C(N, d-1)}{C(N, d)}$$

asymptotic behaviour of $A(N, d)$

- let $\beta = \frac{N}{d}$
- let $A^*(\beta) = \lim_{N=\beta d, d\to\infty} A(N, d)$
- (after some analysis)

$$A^*(\beta) = \begin{cases} 1, & 0 \leq \beta \leq 2 \\ \frac{1}{\beta-1} & \beta \geq 2 \end{cases}$$



Implications

- more data $\Rightarrow$ less ambiguity
- one manifestation of "curse of dimensionality"

  as $d \uparrow$, more data is need to *generalize unambiguously*

Big Data To Good Data: Andrew Ng Urges ML Community To Be More Data-Centric And Less Model-Centric [1]

06/04/2021

summary

- Cover's theorem (proved by induction):

$$C(N, d) = 2 \sum_{k=0}^{d-1} \binom{N-1}{k}$$

- separating capacity of a family of surfaces having d degree of freedom $\Rightarrow 2d$
- implications
  - transforming data into a higher-dimensional space $\rightarrow$ linear separability (kernel SVM, neural networks, etc)
  - the need for more data for classifiers to generalize unambiguously
  - more?

reference

Cover, Thomas M. “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.” *IEEE transactions on electronic computers* (1965)