10-301/601: Introduction to Machine Learning Lecture 15 – Learning Theory (Infinite Case)

Henry Chai & Matt Gormley & Hoda Heidari

10/23/23

Front Matter

- Announcements
  - HW5 released 10/9, due 10/27 (Friday) at 11:59 PM
  - Exam 3 scheduled
    - Tuesday, December 12th from 5:30 PM to 8:30 PM
  - Sign up for peer tutoring! See Piazza for more details

Recall - Theorem 1: Finite, Realizable Case

- For a *finite* hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ (*realizable*) and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$N \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

Recall - Theorem 1: Finite, Realizable Case

- For a *finite* hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ (*realizable*) and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$N = \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

- Making the bound tight and solving for ϵ gives…

Statistical Learning Theory Corollary

- For a *finite* hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ (*realizable*) and arbitrary distribution p^*, given a training dataset S where $|S| = N$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$R(h) \leq \frac{1}{N}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

with probability at least $1 - \delta$.

Recall - Theorem 2: Finite, Agnostic Case

- For a *finite* hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$N \geq \frac{1}{2\epsilon^2}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ satisfy

$$\left|R(h) - \hat{R}(h)\right| \leq \epsilon$$

- Bound is inversely quadratic in ϵ, e.g., halving ϵ means we need four times as many labelled training data points

Statistical Learning Theory Corollary

- For a *finite* hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training dataset S where $|S| = N$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

What happens when $|\mathcal{H}| = \infty$?

- For a *finite* hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S where $|S| = N$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

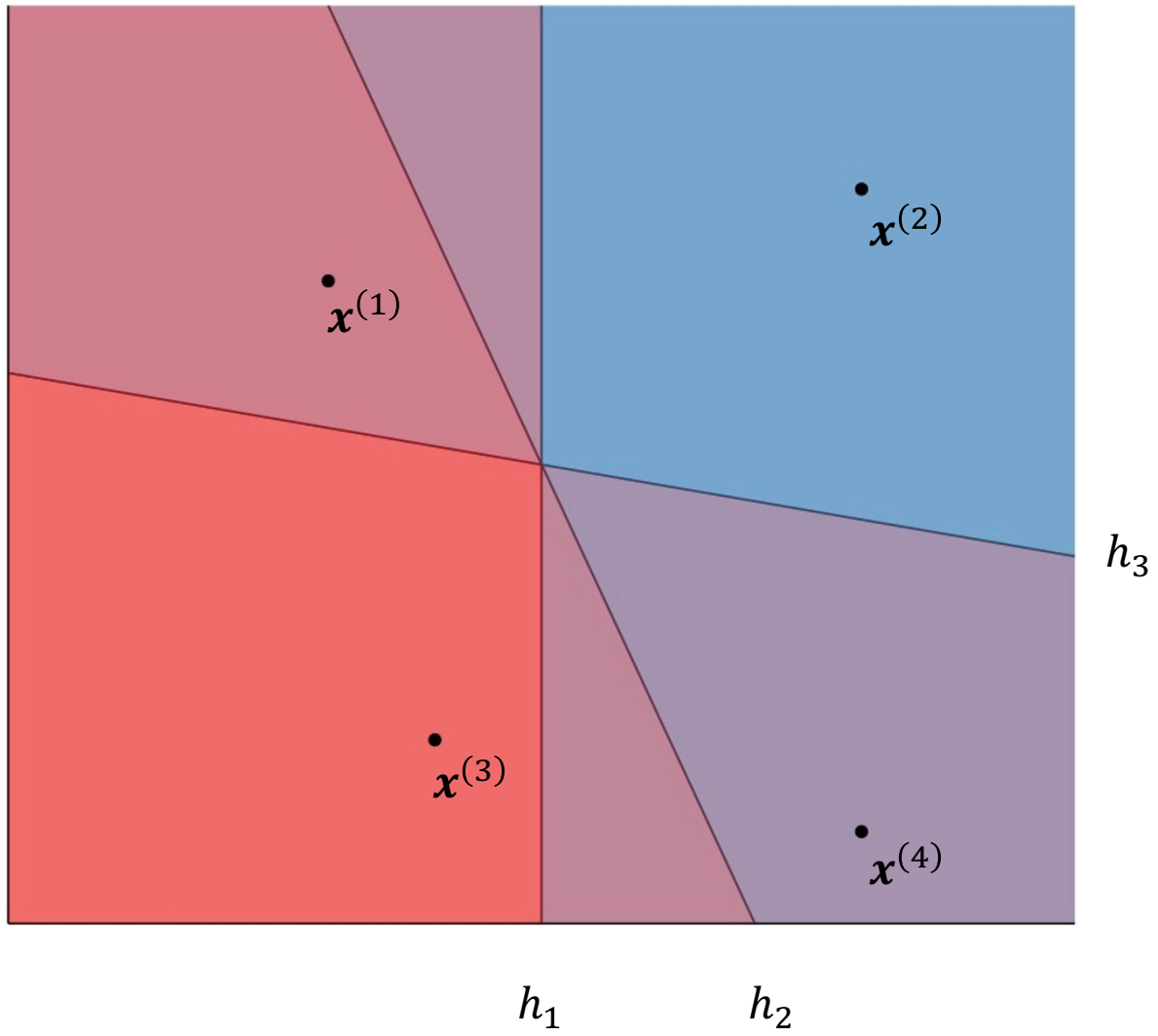
Labellings

- Given some finite set of data points $S = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ and some hypothesis $h \in \mathcal{H}$, applying h to each point in S results in a **labelling**
  - $[h(\boldsymbol{x}^{(1)}), \ldots, h(\boldsymbol{x}^{(N)})]$ is a vector of N +1's and -1's (recall: our discussion of PAC learning assumes binary classification)
- Given $S = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$, each hypothesis in $\mathcal{H}$ induces a labelling but not necessarily a unique labelling
  - The set of labellings induced by $\mathcal{H}$ on S is

$$\mathcal{H}(S) = \{[h(\boldsymbol{x}^{(1)}), \ldots, h(\boldsymbol{x}^{(N)})] \mid h \in \mathcal{H}\}$$

Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

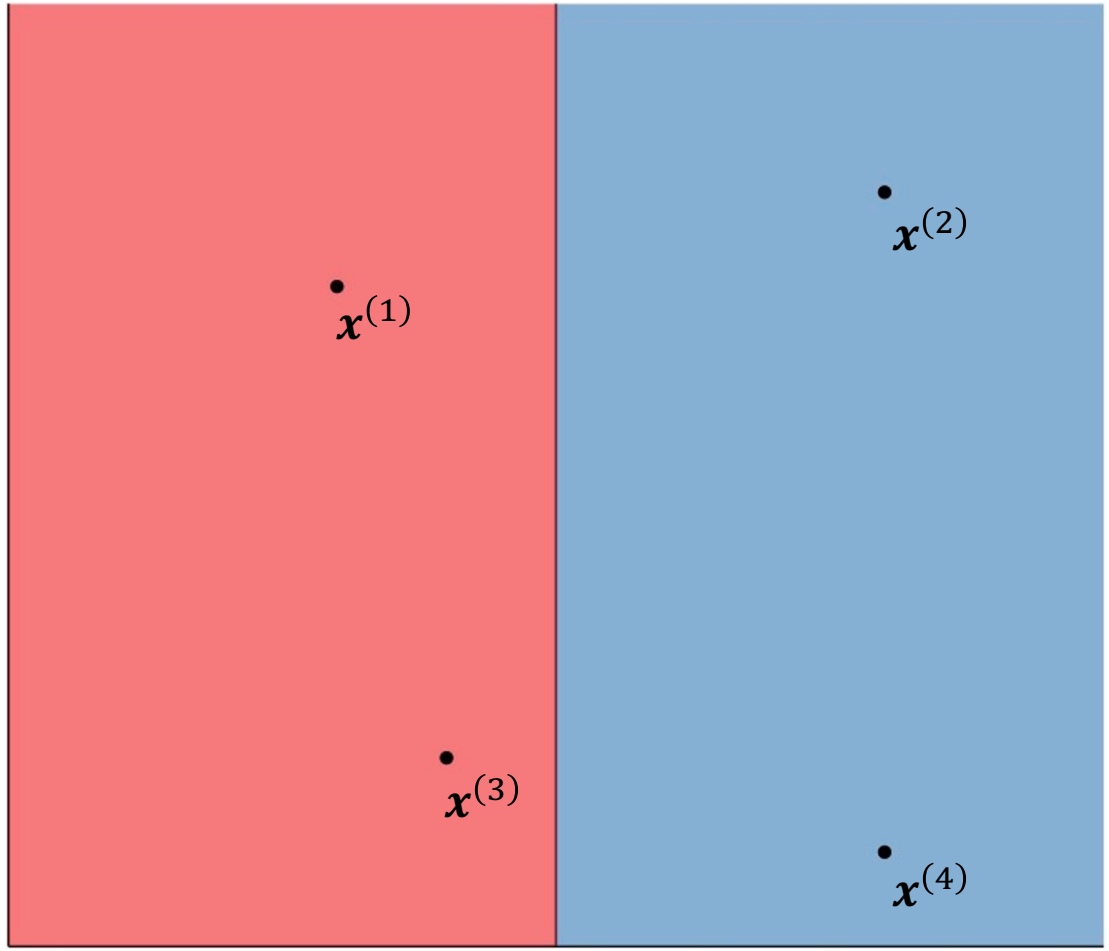


Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$[h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})]$

$= (-1, +1, -1, +1)$



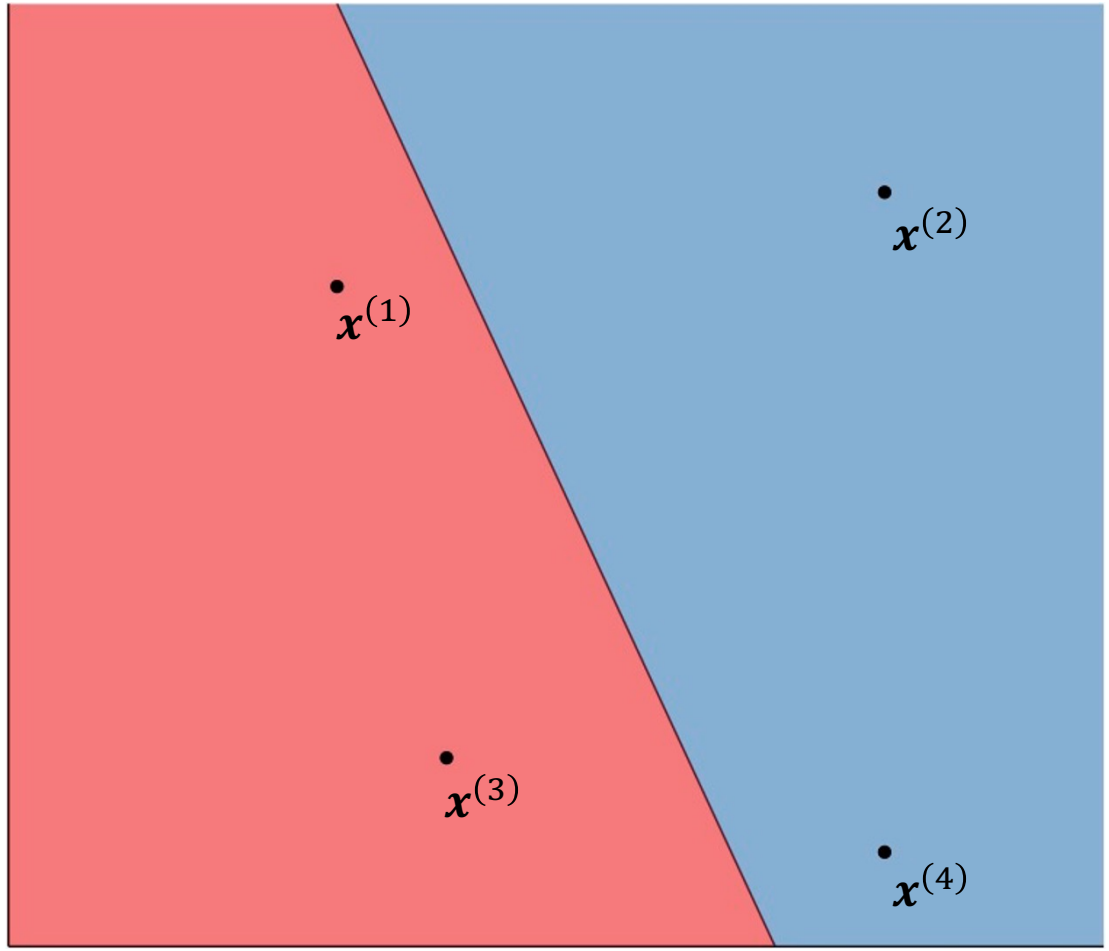
h_1

Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$[h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})]$

$= (-1, +1, -1, +1)$



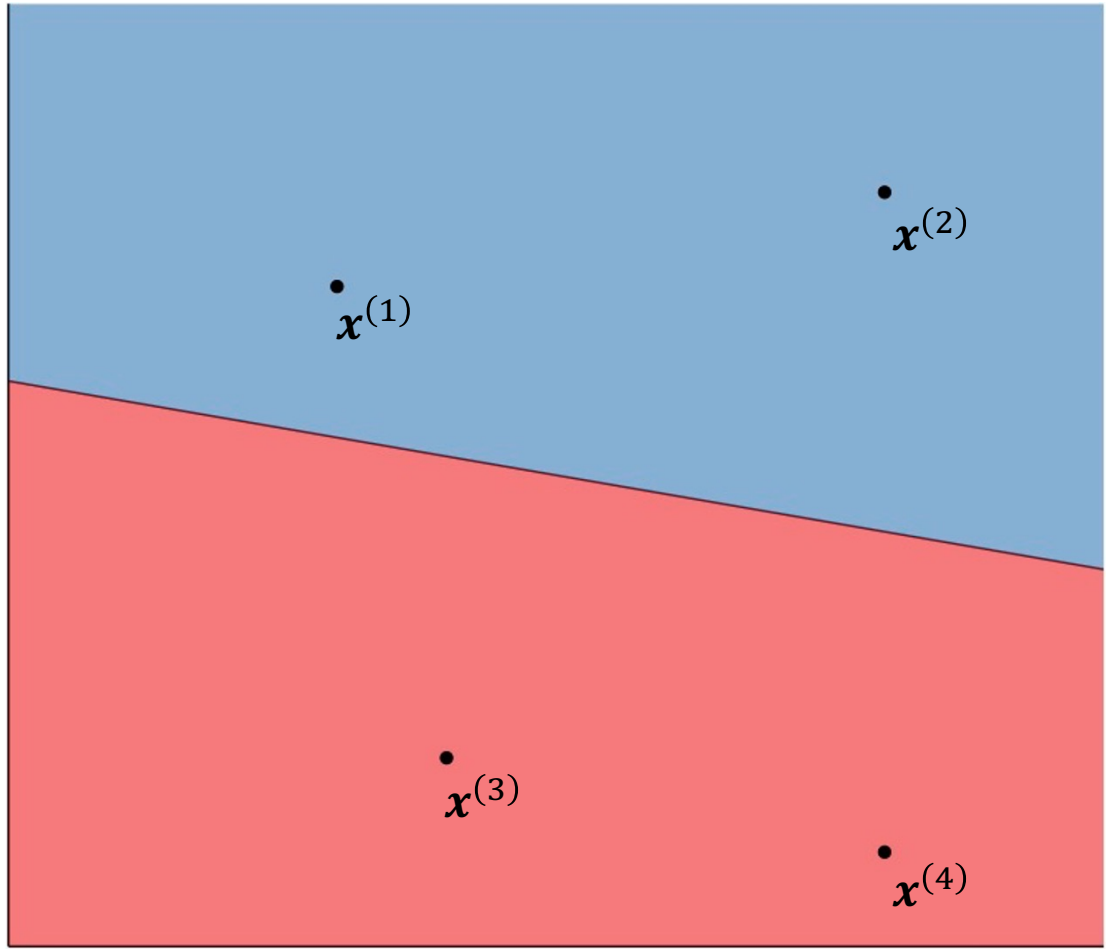
h_2

Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$[h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})]$

$= (+1, +1, -1, -1)$



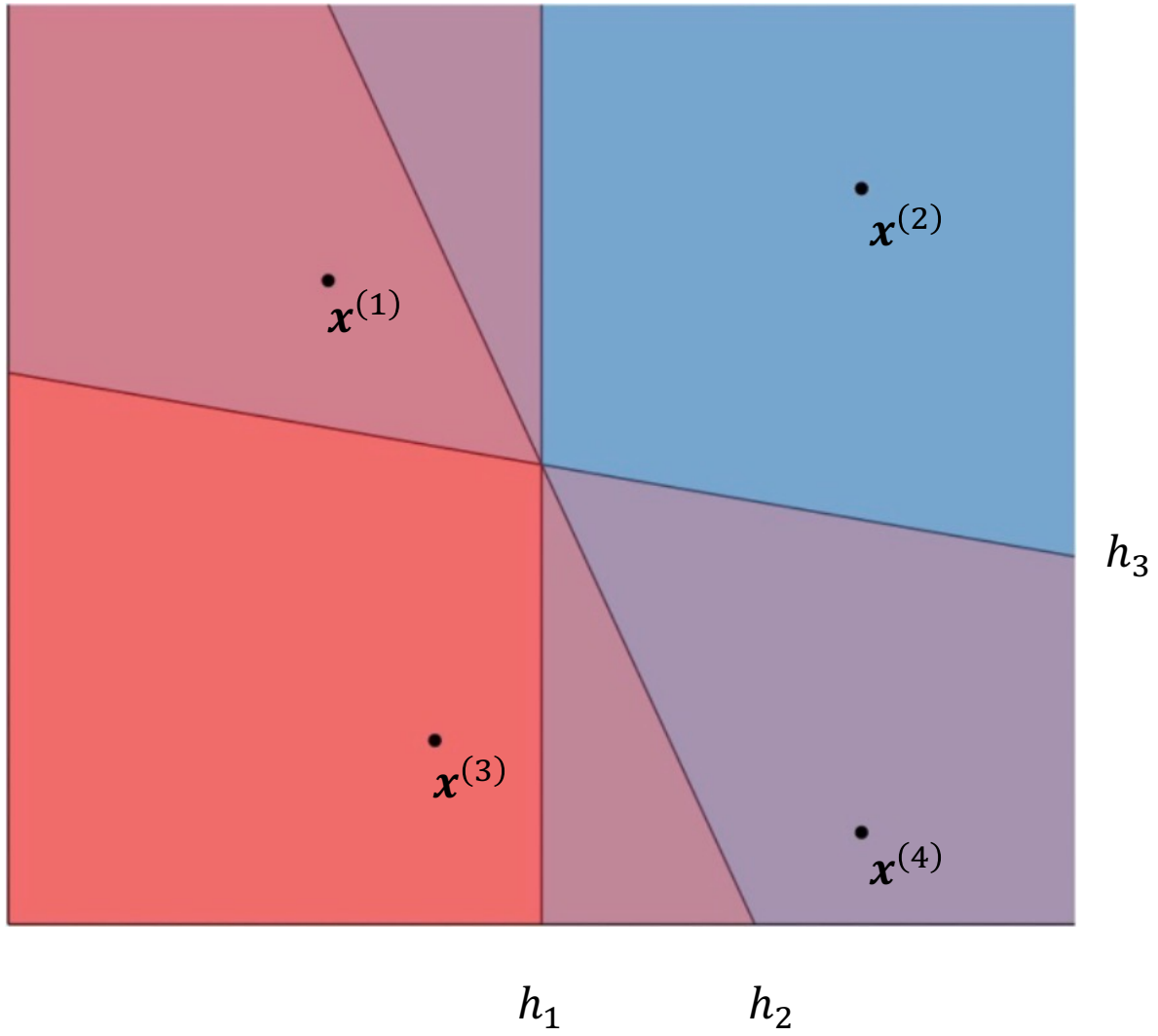
Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S)$

$= \{[+1, +1, -1, -1], [-1, +1, -1, +1]\}$

$|\mathcal{H}(S)| = 2$



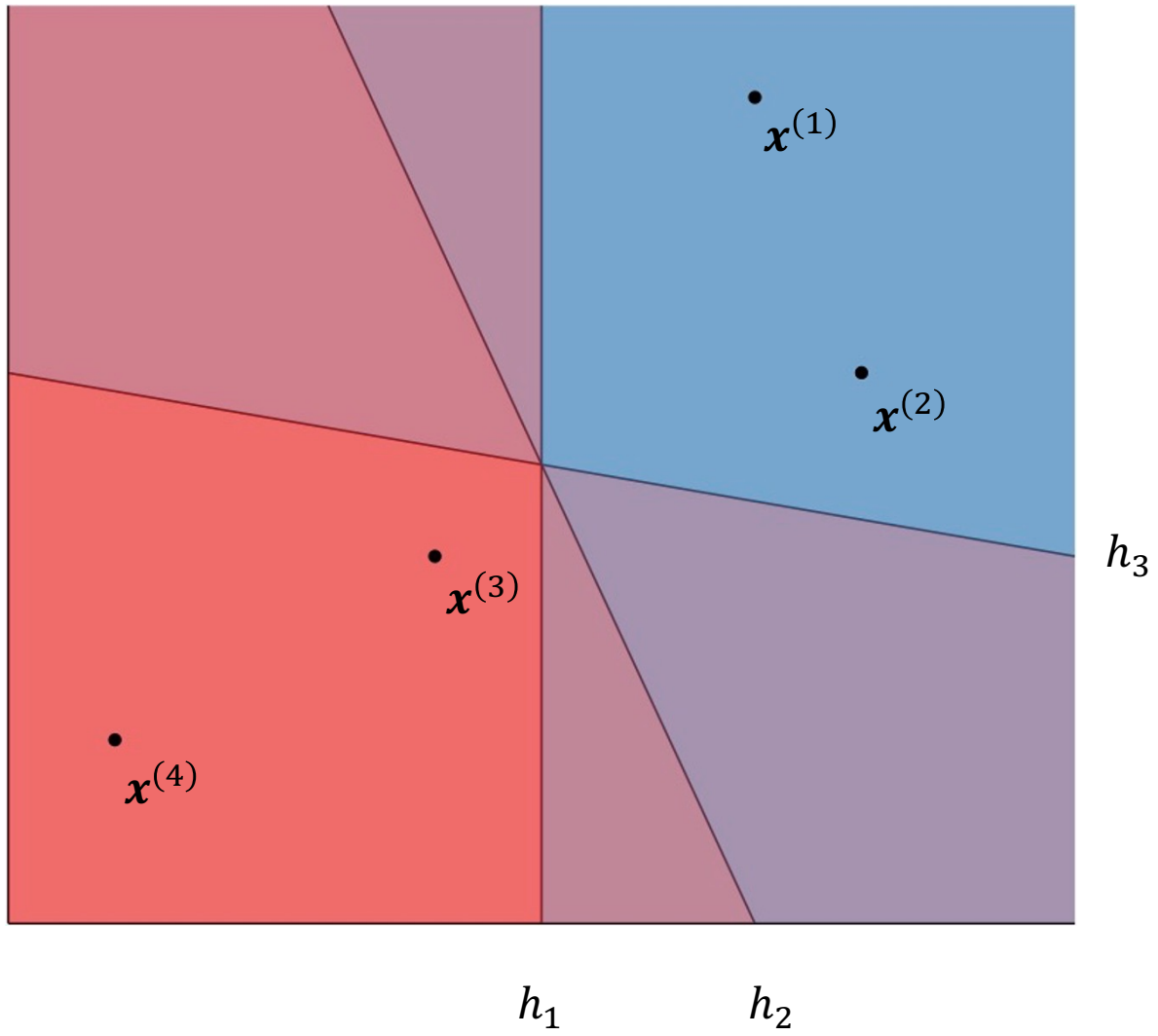
Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S)$

$= \{[+1, +1, -1, -1]\}$

$|\mathcal{H}(S)| = 1$

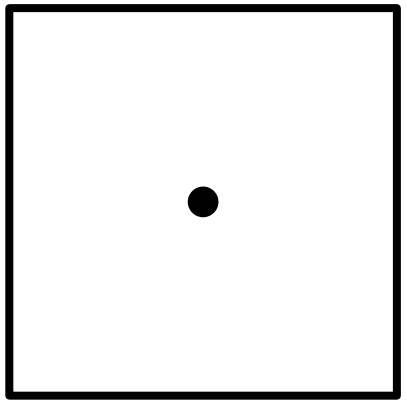


VC-Dimension

- $\mathcal{H}(S)$ is the set of all labellings induced by $\mathcal{H}$ on S
  - If $|S| = N$, then $|\mathcal{H}(S)| \leq 2^N$
  - $\mathcal{H}$ **shatters** S if $|\mathcal{H}(S)| = 2^N$
- The **VC-dimension** of $\mathcal{H}$, $VC(\mathcal{H})$, is the size of the largest set S that can be shattered by $\mathcal{H}$.
  - If $\mathcal{H}$ can shatter arbitrarily large finite sets, then $VC(\mathcal{H}) = \infty$
- To prove that $VC(\mathcal{H}) = d$, you need to show
  1. $\exists$ some set of d data points that $\mathcal{H}$ can shatter and
  2. $\nexists$ a set of $d + 1$ data points that $\mathcal{H}$ can shatter

VC-Dimension: Example

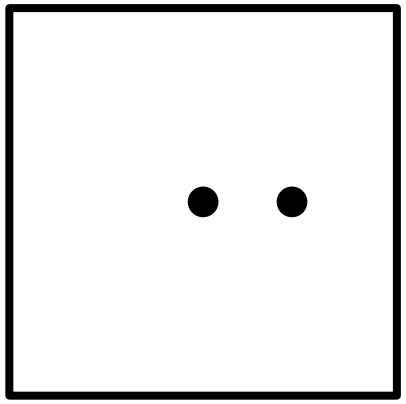
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?



S

VC-Dimension: Example

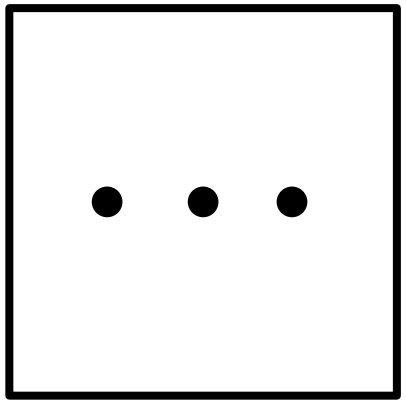
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?



S

VC-Dimension: Example

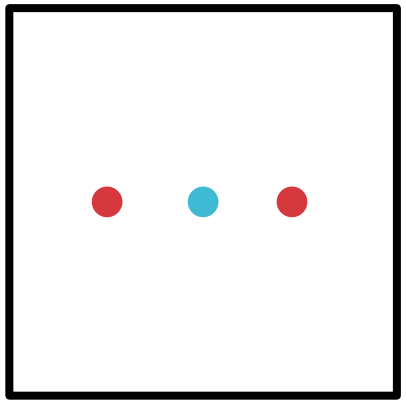
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

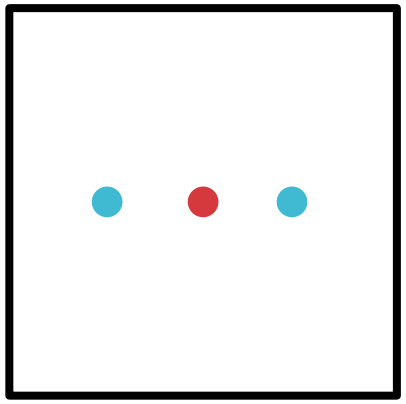
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

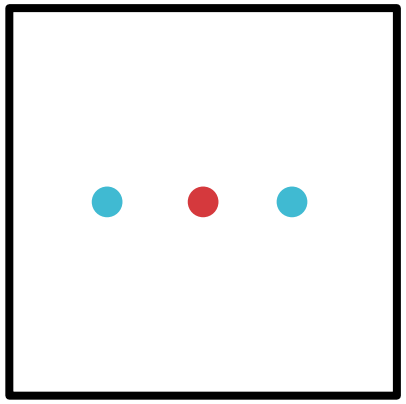
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

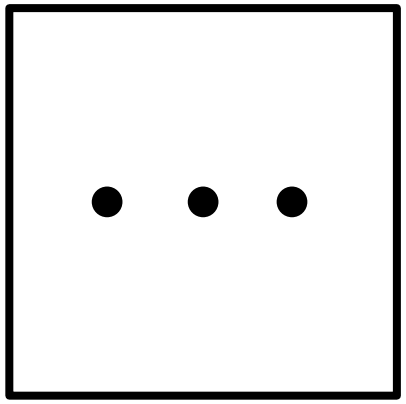
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter ***some*** set of 3 points?



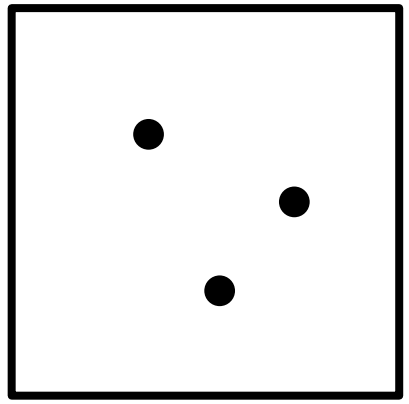
S

VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



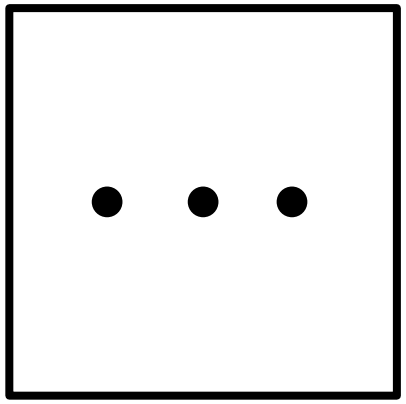
S_1



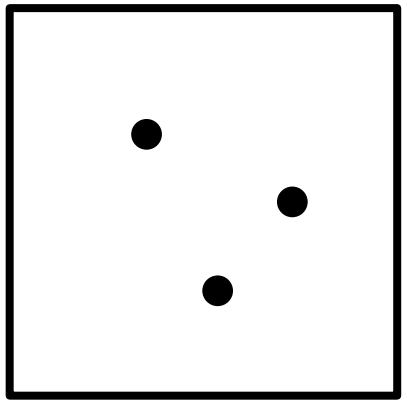
S_2

VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



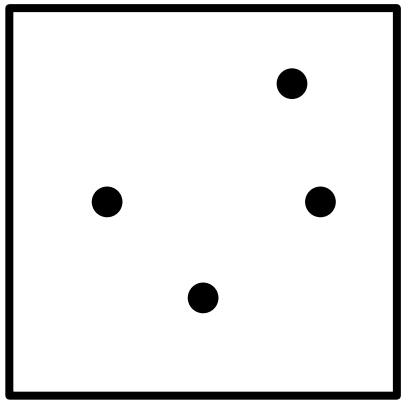
$|\mathcal{H}(S_1)| = 6$



$|\mathcal{H}(S_2)| = 8$

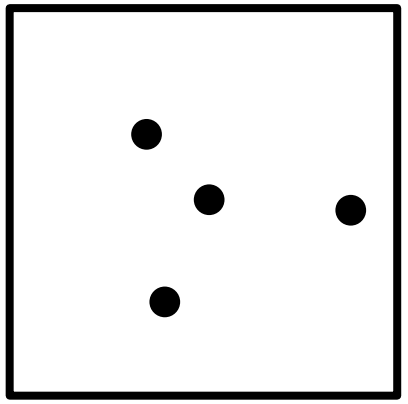
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



S_1

All points on the convex hull

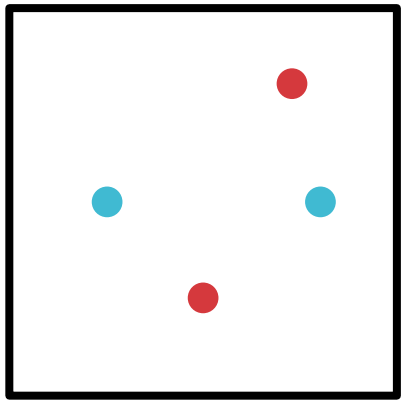


S_2

At least one point inside the convex hull

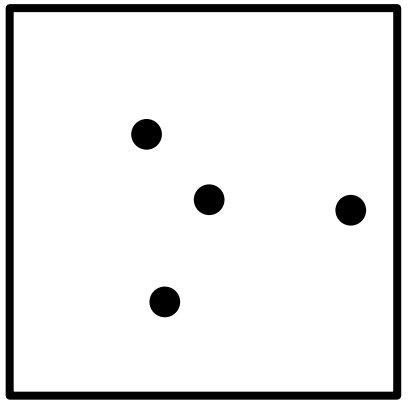
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



S_1

All points on the convex hull

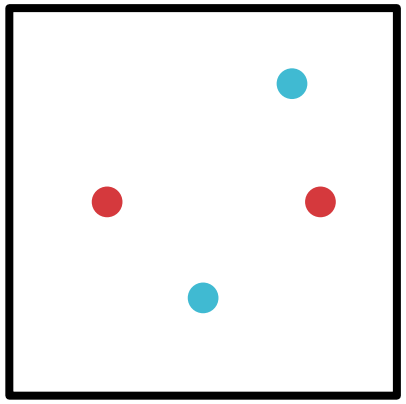


S_2

At least one point inside the convex hull

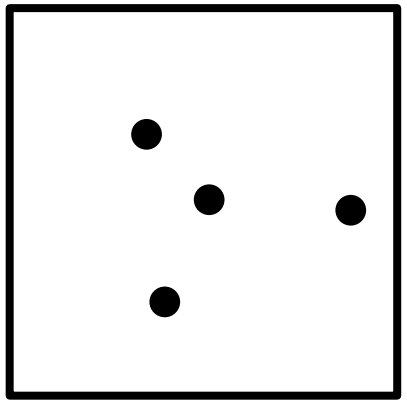
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



S_1

All points on the convex hull

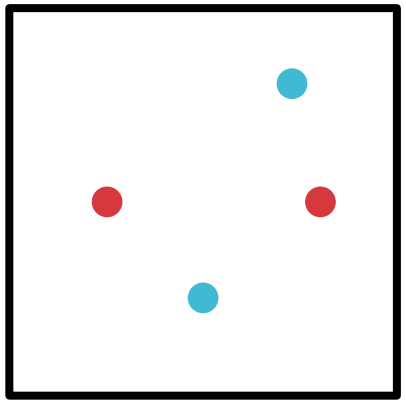


S_2

At least one point inside the convex hull

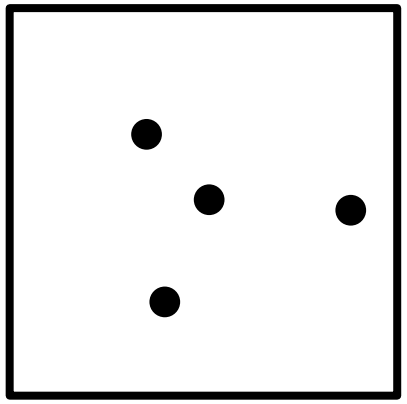
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

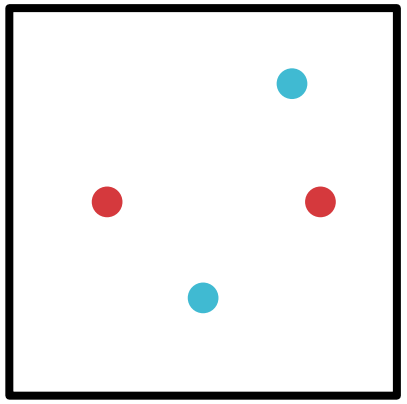


S_2

At least one point inside the convex hull

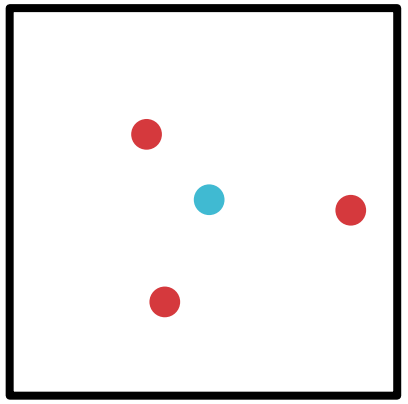
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

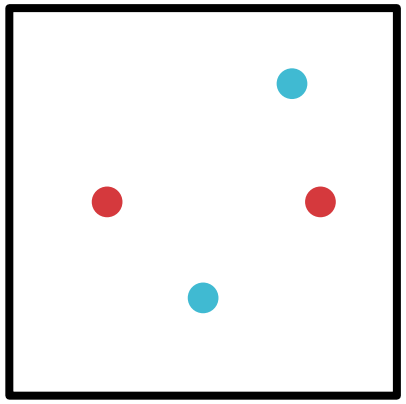


S_2

At least one point inside the convex hull

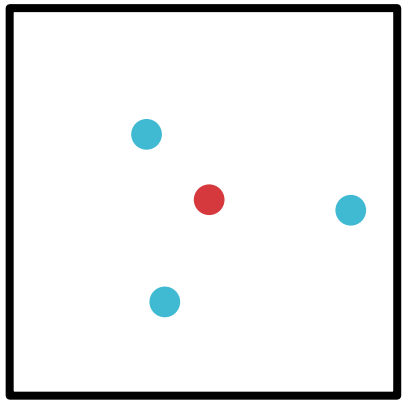
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

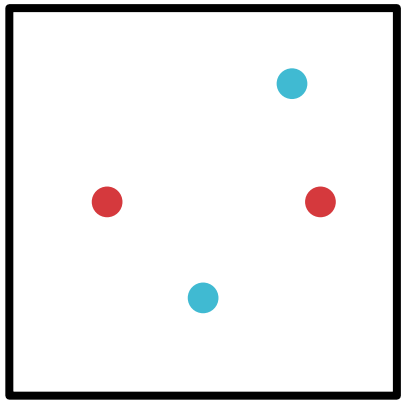


S_2

At least one point inside the convex hull

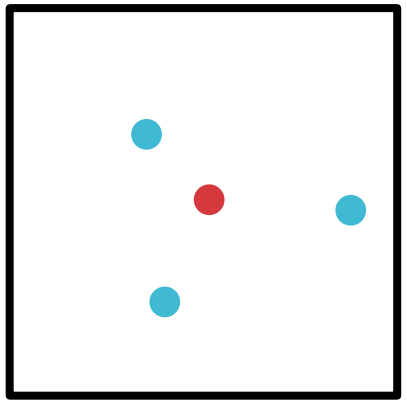
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

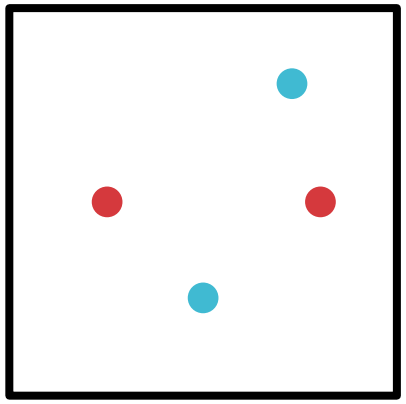


$|\mathcal{H}(S_2)| = 14$

At least one point inside the convex hull

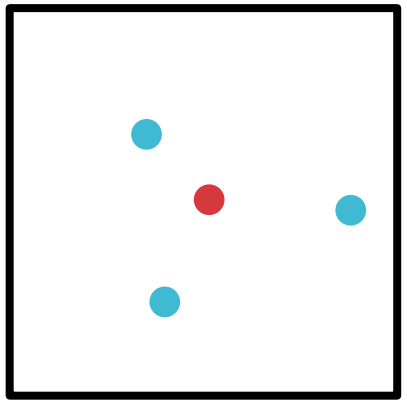
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- $VC(\mathcal{H}) = 3$
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull



$|\mathcal{H}(S_2)| = 14$

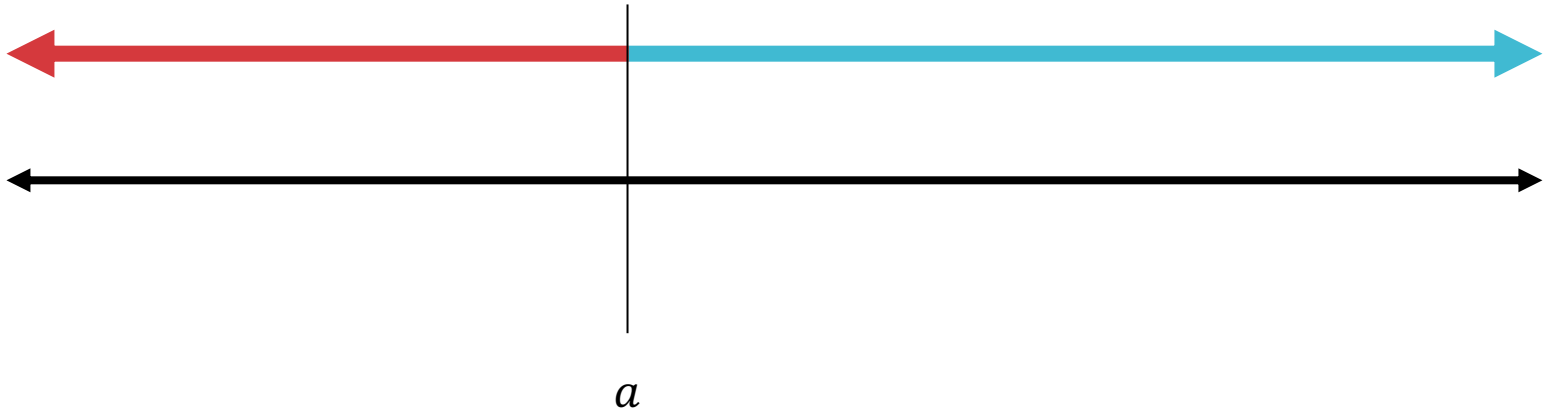
At least one point inside the convex hull

VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^d$ and $\mathcal{H} =$ all d-dimensional linear separators
- $VC(\mathcal{H}) = d + 1$

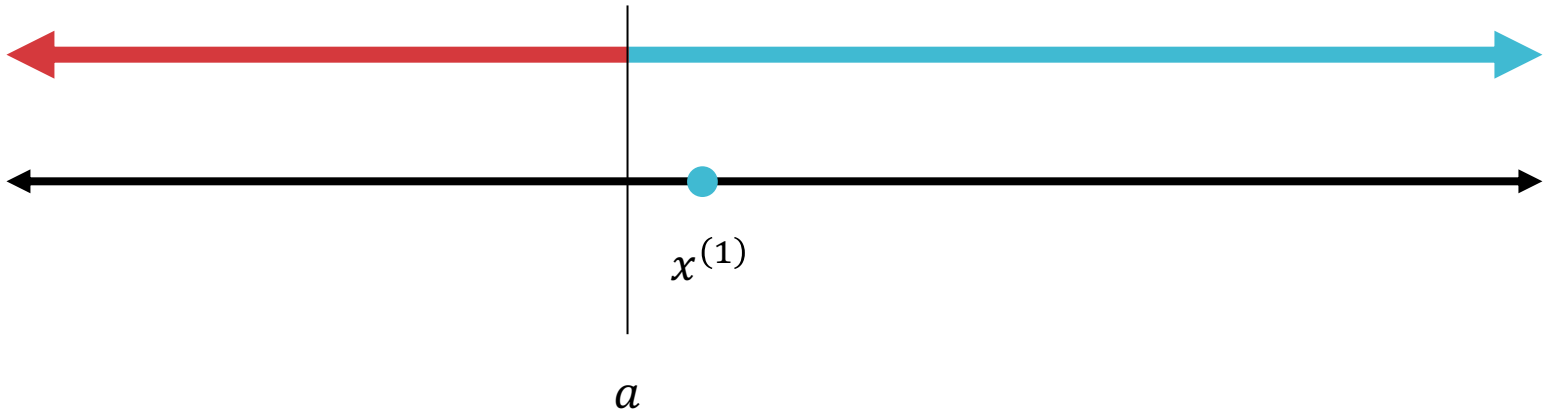
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



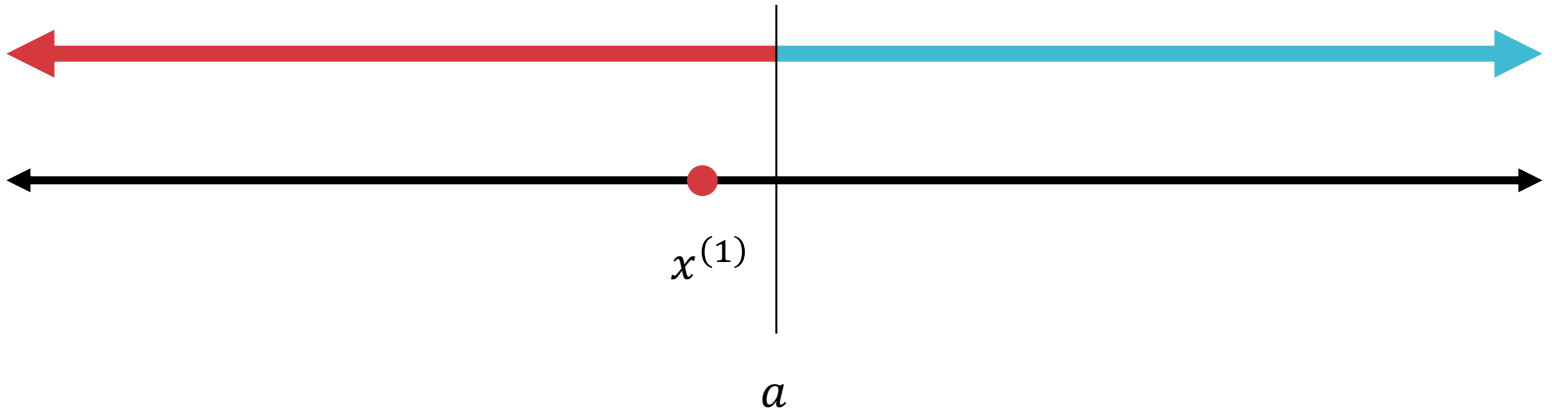
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



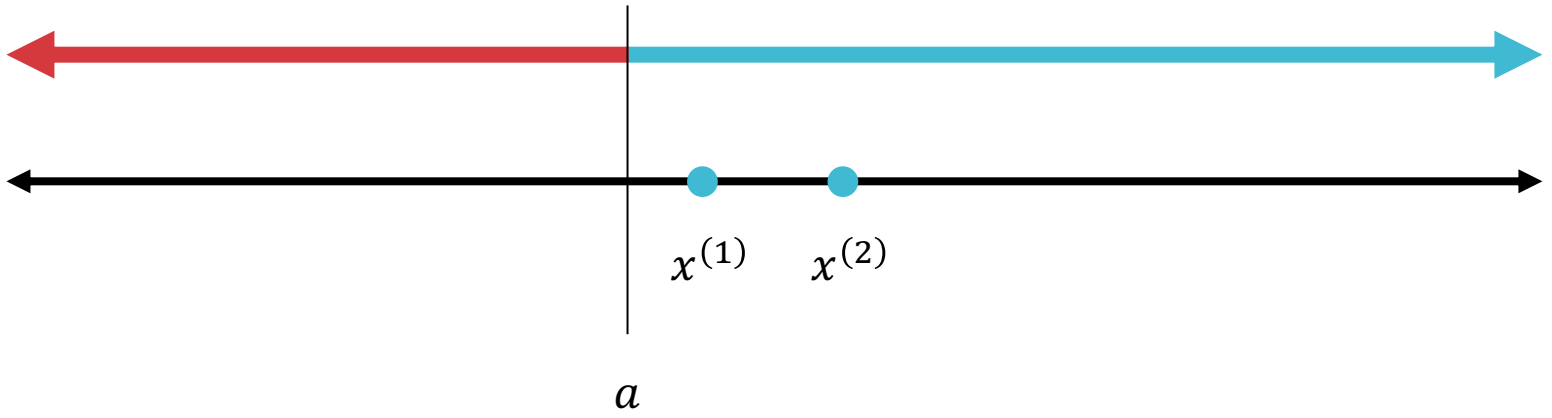
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



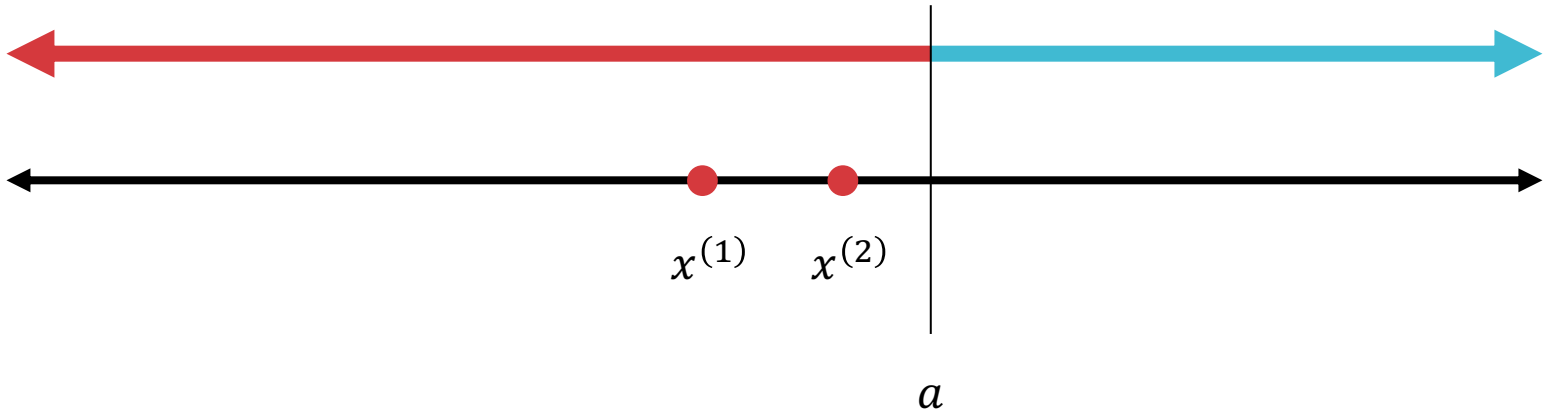
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



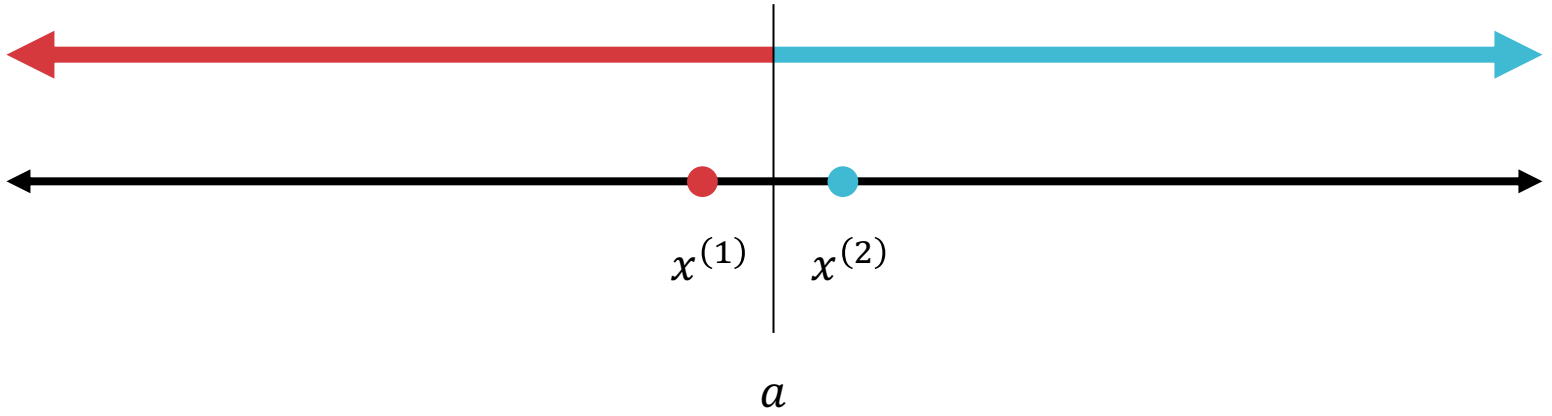
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



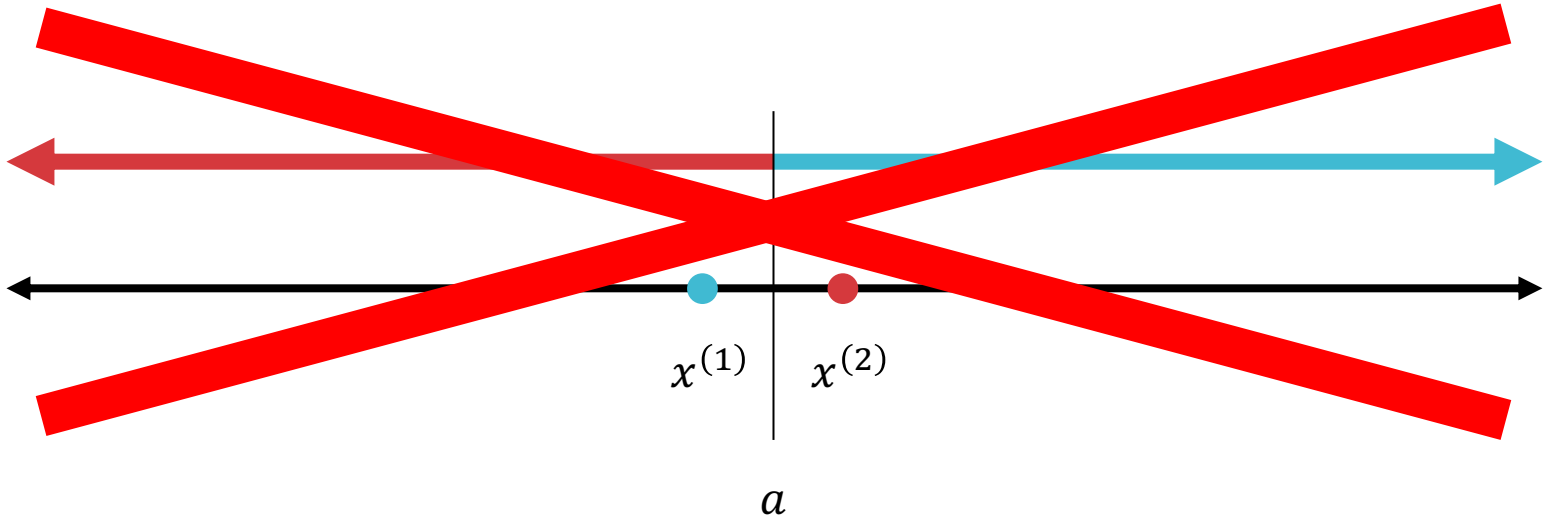
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



VC-Dimension: Example

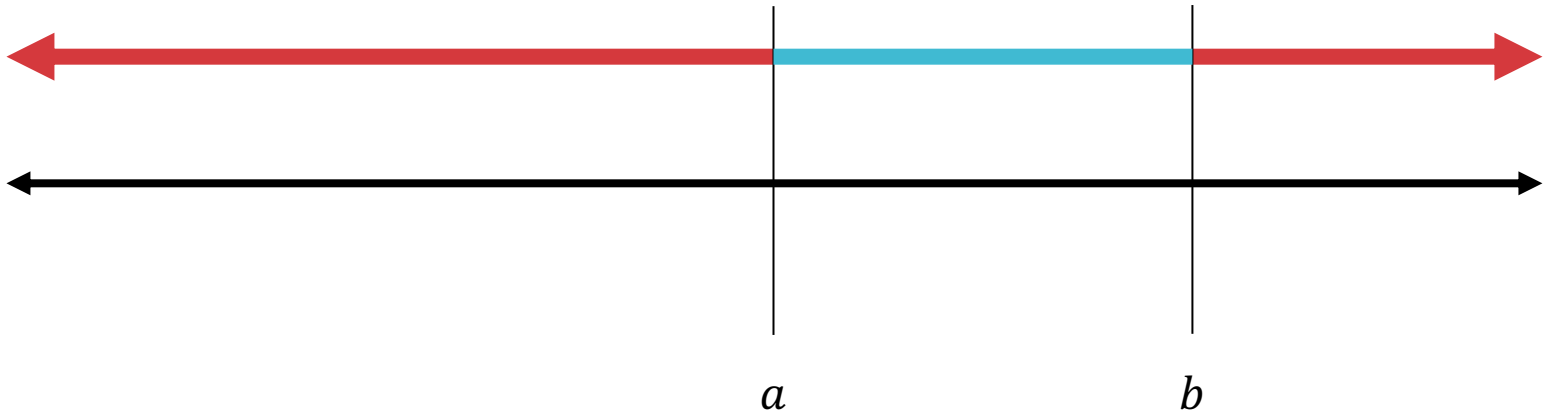
- $x \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- $VC(\mathcal{H}) = 1$

VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals

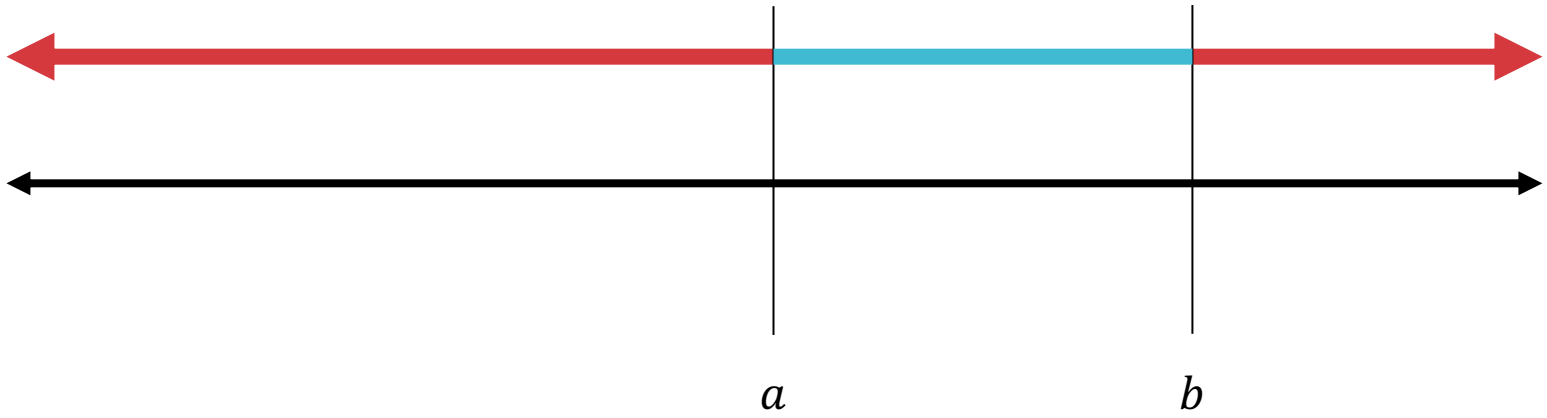


Poll Question 1:

What is $VC(\mathcal{H})$?

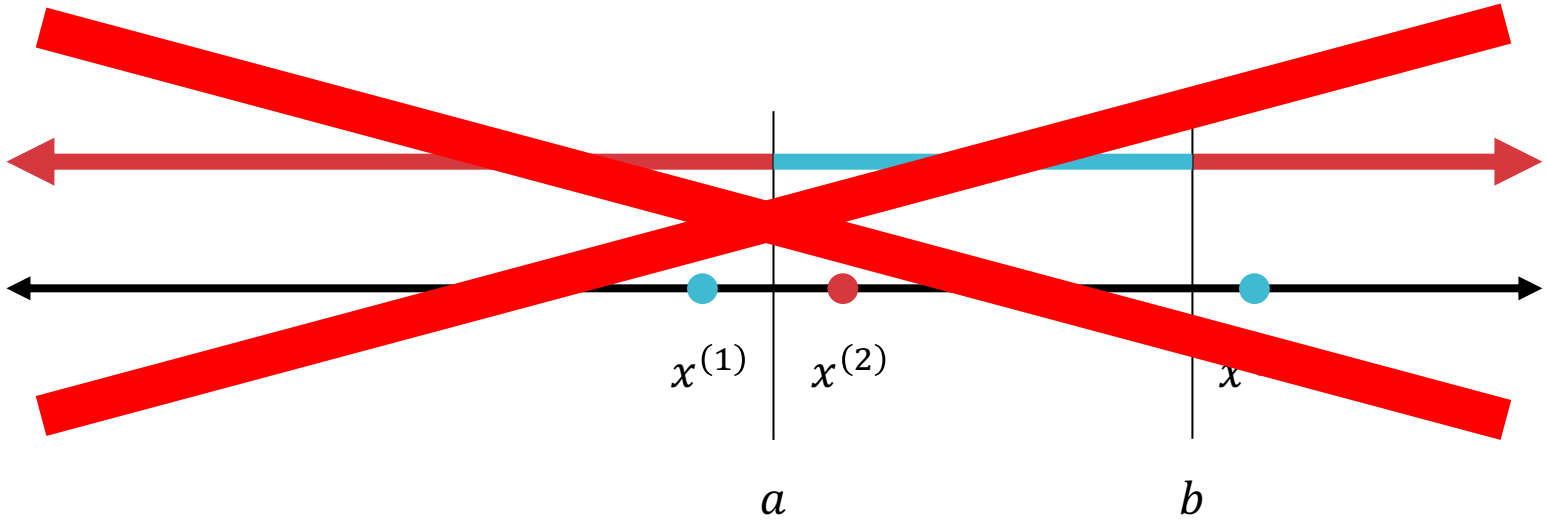
A. 0
B. 1
C. 1.5 **(TOXIC)**
D. 2
E. 3

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



- $VC(\mathcal{H}) = 2$

Theorem 3: Vapnik-Chervonenkis (VC)-Bound

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M = O\left(\frac{1}{\epsilon}\left(VC(\mathcal{H})\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

Statistical Learning Theory Corollary 3

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ and arbitrary distribution p^*, given a training dataset S where $|S| = N$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$R(h) \leq O\left(\frac{1}{N}\left(VC(\mathcal{H}) \log\left(\frac{N}{VC(\mathcal{H})}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

with probability at least $1 - \delta$.

Theorem 4: Vapnik-Chervonenkis (VC)-Bound

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$N = O\left(\frac{1}{\epsilon^2}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ have

$$\left|R(h) - \hat{R}(h)\right| \leq \epsilon$$

Statistical Learning Theory Corollary 4

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training dataset S where $|S| = N$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

with probability at least $1 - \delta$.

Approximation Generalization Tradeoff

How well does h generalize?

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

How well does h approximate c^*?

Approximation Generalization Tradeoff

Increases as $VC(\mathcal{H})$ increases

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

Decreases as $VC(\mathcal{H})$ increases

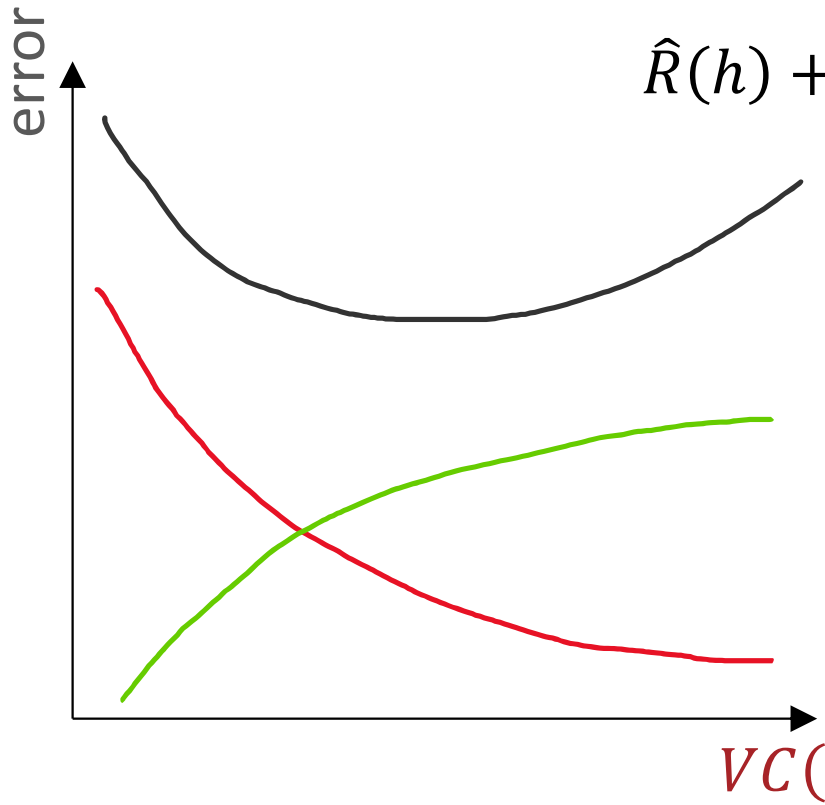
Can we use this corollary to guide model selection?

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training dataset S where $|S| = N$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

with probability at least $1 - \delta$.

Learning Theory and Model Selection

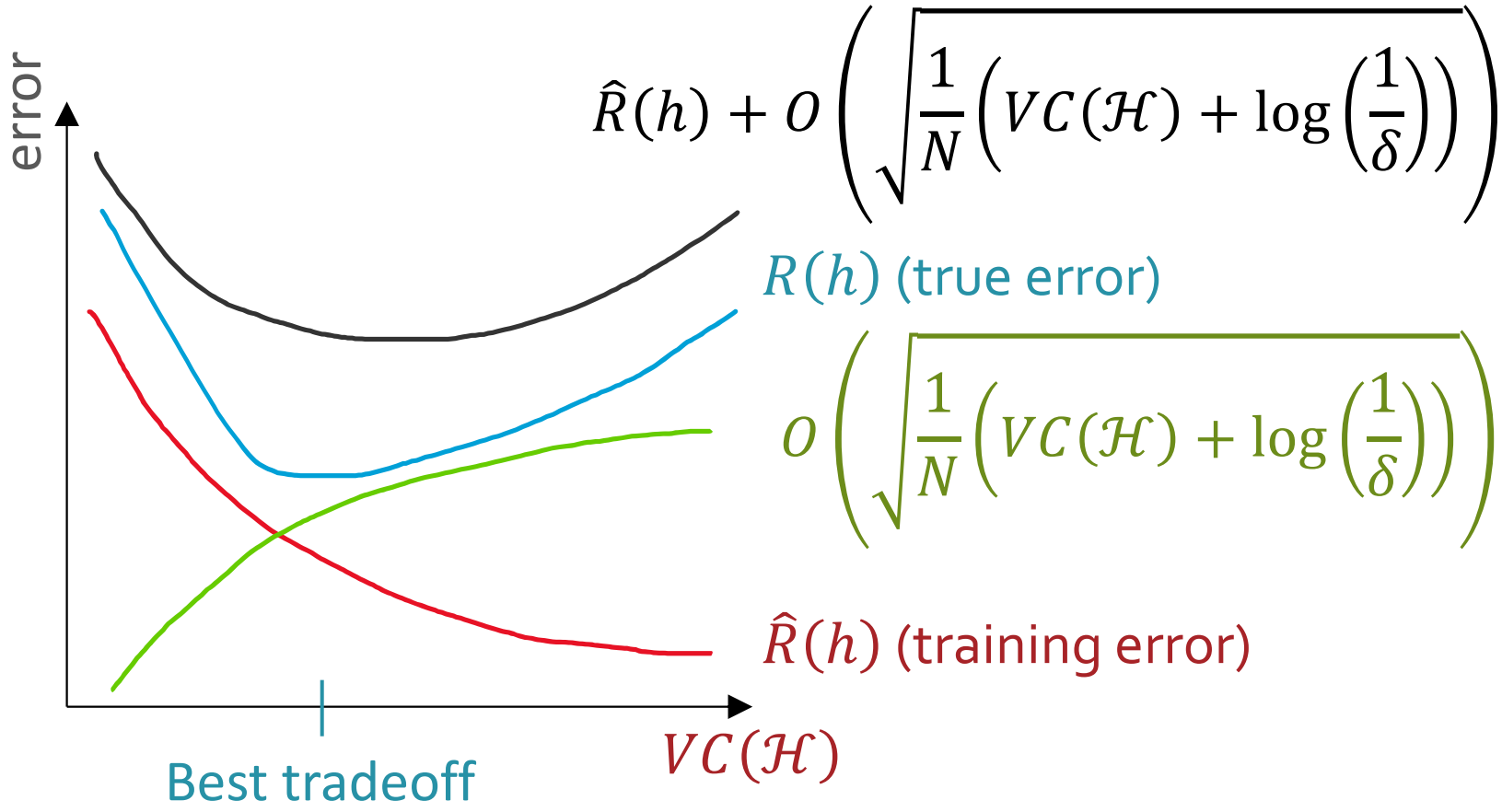


$$\hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

$$O\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

$\hat{R}(h)$ (training error)

Learning Theory and Model Selection



- How can we find this "best tradeoff" for linear separators?
- Use a regularizer! By (effectively) reducing the number of features our model considers, we reduce its VC-dimension.

Learning Theory Learning Objectives

You should be able to…

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world machine learning examples
- Theoretically motivate regularization

10-301/601: Introduction to Machine Learning Lecture 15 – Societal Impacts of ML

Henry Chai & Matt Gormley & Hoda Heidari

10/23/23

ML in Societal Applications

Deep learning is being used to predict critical COVID-19 cases

8 WAYS MACHINE LEARNING WILL IMPROVE EDUCATION

BY MATTHEW LYNCH / JUNE 12, 2018 / 5

Artificial Intelligence and Accessibility: Examples of a Technology that Serves People with Disabilities

SCI-FI VISIONS

Your Future Doctor May Not be Human. This Is the Rise of AI in Medicine.

From mental health apps to robot surgeons, artificial intelligence is already changing the practice of medicine.

techworld FROM IDG Features Technology Innovation Partner Zone the techies

Home 〉 Features 〉 Emerging tech & innovation Features

Researcher explains how algorithms can create a fairer legal system

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

By Claire Cain Miller

20 JAN 2017 | Insight

Kevin Petrasic | Benjamin Saul

Algorithms and bias: What lenders need to know

The algorithms that power fintech may discriminat
can be difficult to anticipate—and financial institut
accountable even when alleged discrimination is c
unintentional.

HOME > STRATEGY

Artificial intelligence is slated to disrupt 4.5 million jobs for African Americans, who have a 10% greater likelihood of automation-based job loss than other workers

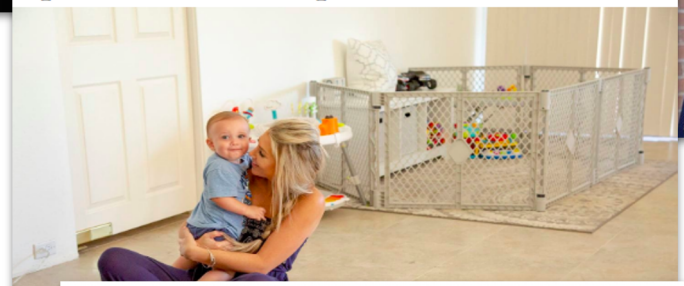
Allana Akhtar Oct 7, 2019, 12:57 PM

Misinformation on coronavirus is proving highly contagious

By DAVID KLEPPER July 29, 2020

The Switch

Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.



ACLU

Email address ZIP code GO

BECOME A MEMBER / RENEW / TAKE ACT

ISSUES KNOW YOUR RIGHTS DEFENDING OUR RIGHTS BLOGS ABOUT

SPEAK FREELY

How Facebook Is Giving Sex Discrimination in Employment Ads a New Life

By Galen Sherwin, ACLU Women's Rights Project
SEPTEMBER 18, 2018 | 10:00 AM

The New York Times

I.R.S. Changes Audit Practice That Discriminated Against Black Taxpayers

The agency will overhaul how it scrutinizes returns that claim the earned-income tax credit, which is aimed at alleviating poverty.

MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

By Robert David Hart · July 10, 2017

PROPUBLICA Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Societal Goals

Foster:

- Productivity and efficiency gains
- Innovation and economic growth
- Due process
  - Consistency
  - Traceability
  - Making choices & biases evident
- …

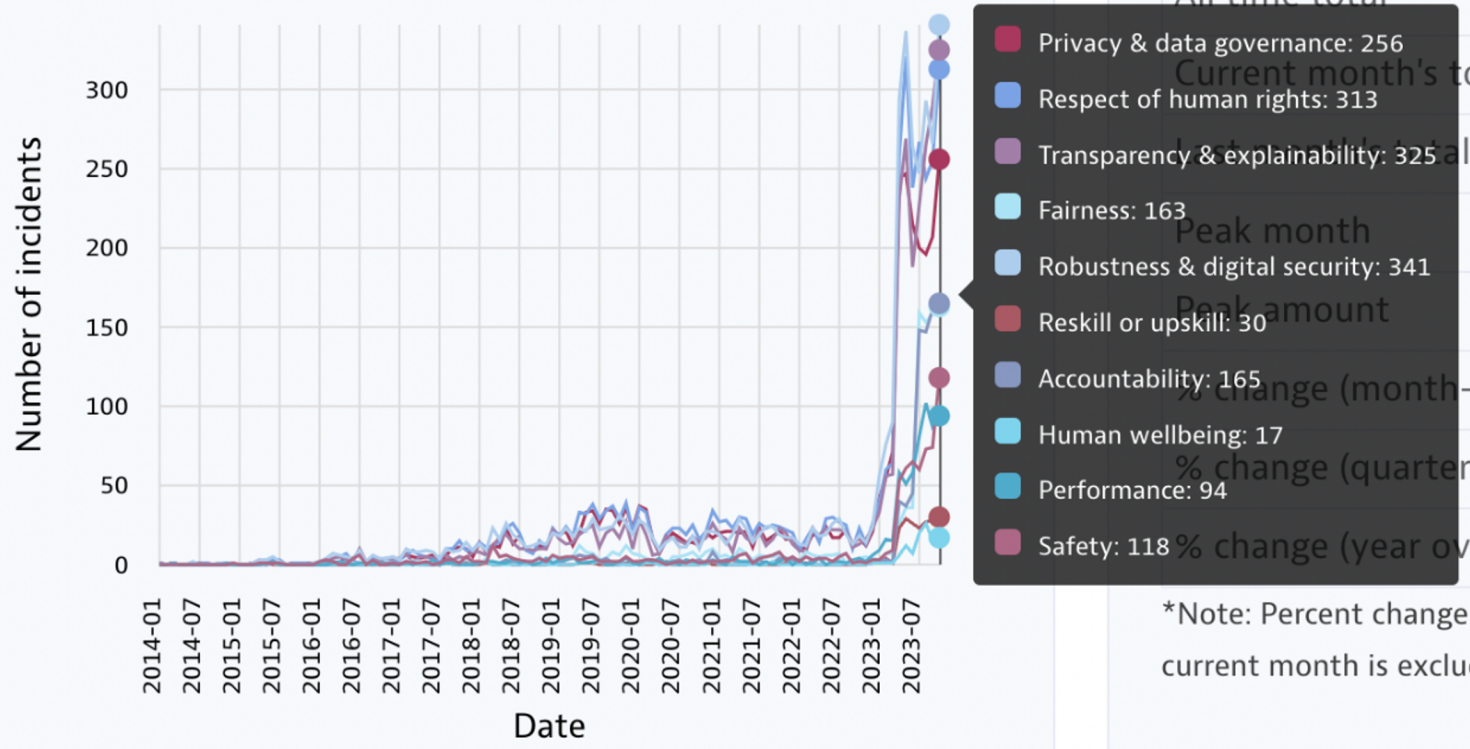
Mitigate:

- Violations of human rights
  - Justice, equity, and non-discrimination
  - Privacy and non-surveillance
  - Freedom of communication and expression
  - Economic freedom
- Negative impact on human flourishing and wellbeing
  - Loss of human sovereignty and control
  - Human cognitive abilities
  - …

AI Incidents on the Rise

Summary visualisations

Evolution of incidents by AI principle



Summary statistics

	Incidents	Articles
All time total	6264	36345
Current month's total	317	1768
Last month's total	616	3227
Peak month	2023-10	2023-10
Peak amount	616	3227
% change (month-over-month)	23.2	51.22
% change (quarter-over-quarter)	13.01	13.87
% change (year over year)	961.58	690.9

*Note: Percent change is calculated based on preceding full months (i.e. the current month is excluded).

Principles

- Fairness
- Accountability
- Transparency
- Safety and reliability
- Privacy
- ...

Beyond Principles

Concerns around **impact**:

- Economic (IP, Antitrust, labor market effects)
- Sustainability and environmental
- Eroding democratic values
  - misinformation and disinformation

Concerns around the **process**:

- Human sovereignty, autonomy, agency, self-determination
  - Participation
  - Recourse / appeal
  - Mental health
- ...

Unfairness and Discrimination

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

(Outcome) Unfairness

Formal Principle of Distributive Justice:

"Equals should be treated equally, and unequals unequally, in proportion to relevant similarities and differences." [Aristotle, …, Feinberg'1973]

Working Definition of Outcome Unfairness:

Disparate or unequal allocation of harm/benefit across socially salient, but morally irrelevant groups of people.

Mathematical Notions of Fairness

- **Group** notions
  - Statistical parity
  - Equality of accuracy
  - Equality of false positive/false negative rates
  - Equality of positive/negative predictive value

- **Individual** notions
  - Treat similar individuals similarly.

- **Counterfactual** notions

Statistical/Demographic Parity

- Equal **selection rate** across different groups:

$$P[\hat{Y} = 1 \mid S = s_1] = P[\hat{Y} = 1 \mid S = s_2]$$

- Equal Employment Opportunity Commission:

> *"A selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of [discrimination]."*

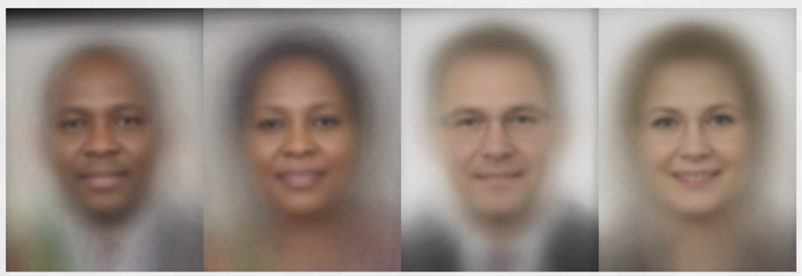
Equality of Accuracy

- Equality of the prediction accuracy (L) across groups:

$$E[L(\hat{y}, y) \mid S = s_1] = E[L(\hat{y}, y) \mid S = s_2]$$

- **Example:** Gender shades (Buolamwini et al.'18)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Equality of FPR/FNR

- Equality of the **False Positive Rate (FPR)** across groups:

$$P[\hat{Y}=1 \mid Y=0, S=s_1]=P[\hat{Y}=1 \mid Y=0, S=s_2]$$

- Equality of the **False Negative Rate (FNR)** across groups:

$$P[\hat{Y}=0 \mid Y=1, S=s_1]=P[\hat{Y}=0 \mid Y=1, S=s_2]$$

- Equality of **Odds**: equal FNR and FPR simultaneously

Equality of PPV/NPV

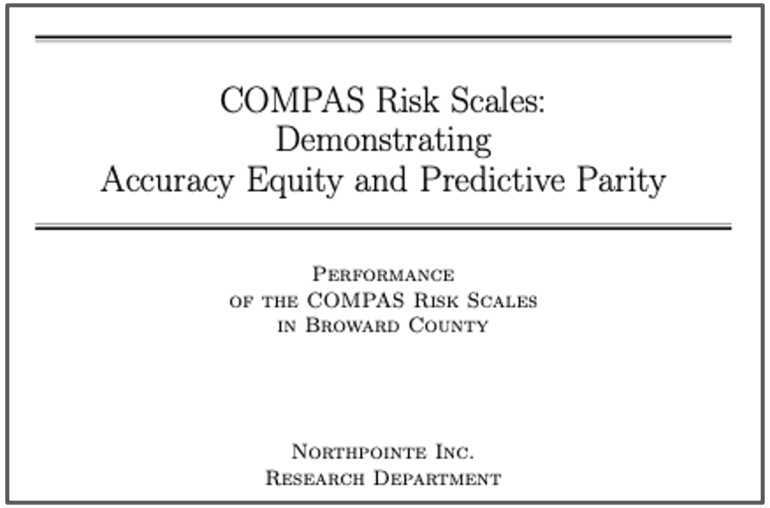
- Equality of the **Positive Predictive Value (PPV)**

$$P[Y=1 \mid \hat{Y}=1, S=s_1]=P[Y=1 \mid \hat{Y}=1, S=s_2]$$

- Equality of the **Negative Predictive Value (NPV)**

$$P[Y=0 \mid \hat{Y}=0, S=s_1]=P[Y=0 \mid \hat{Y}=0, S=s_2]$$

- **Predictive Value Parity (PVP):** equal PPV and NPV simultaneously



COMPAS Risk Scales:
Demonstrating
Accuracy Equity and Predictive Parity

PERFORMANCE
OF THE COMPAS RISK SCALES
IN BROWARD COUNTY

NORTHPOINTE INC.
RESEARCH DEPARTMENT

Common Pros and Cons

- Ignores possible correlation between Y and S.
- Allows for trading off different types of error.
- Allows laziness.
- Doesn't consider practical considerations.
  - e.g., High accuracy difficult to attain for small groups
- …

Summary of Fairness Notions w. Confusion Matrix

For each group s, form:

	$\hat{Y} = 0$	$\hat{Y} = 1$
Y=0	a (true negative)	b (false positive)
Y=1	c (false negative)	d (true positive)

- Statistical parity = Equality of $\frac{b+d}{a+b+c+d}$
- Equality of accuracy = Equality of $\frac{a+d}{a+b+c+d}$
- Equality of FPR/FNR = Equality of $\frac{b}{a+b} / \frac{c}{c+d}$
- Equality of PPV/NPV = Equality of $\frac{d}{d+b} / \frac{a}{a+c}$

across all s.

Individual vs. Group Fairness

- Treating people as individuals, regardless of their group membership.
- Disparate Treatment:

  "Similarly situated individuals must be treated similarly."

- Similarity must be defined *with respect to the task at hand.*

  Example: movie casting vs. employment decisions in tech sector

Formalizing Individual Fairness

(Dwork et al. 2012):

- $d(\mathbf{x}_i,\mathbf{x}_j)$: a metric defining distance between two individuals
- D: a measure of distance between distributions
- A randomized classifier h mapping $\mathbf{x}$ to $\Delta_h(\mathbf{x})$ satisfies the (D, d)-Lipschitz property if $\forall \mathbf{x}_i , \mathbf{x}_j ,$

$$D(\Delta_h(\mathbf{x}_i), \Delta_h(\mathbf{x}_j)) \leq d(\mathbf{x}_i , \mathbf{x}_j).$$

Several problems with the Formulation

- Does not treat **dis**similar individuals **differently**.
- How should we pick d and D?
- Applicable to probabilistic models, only.
- Computationally expensive ($O(n^2)$ pairwise constraints)
- …