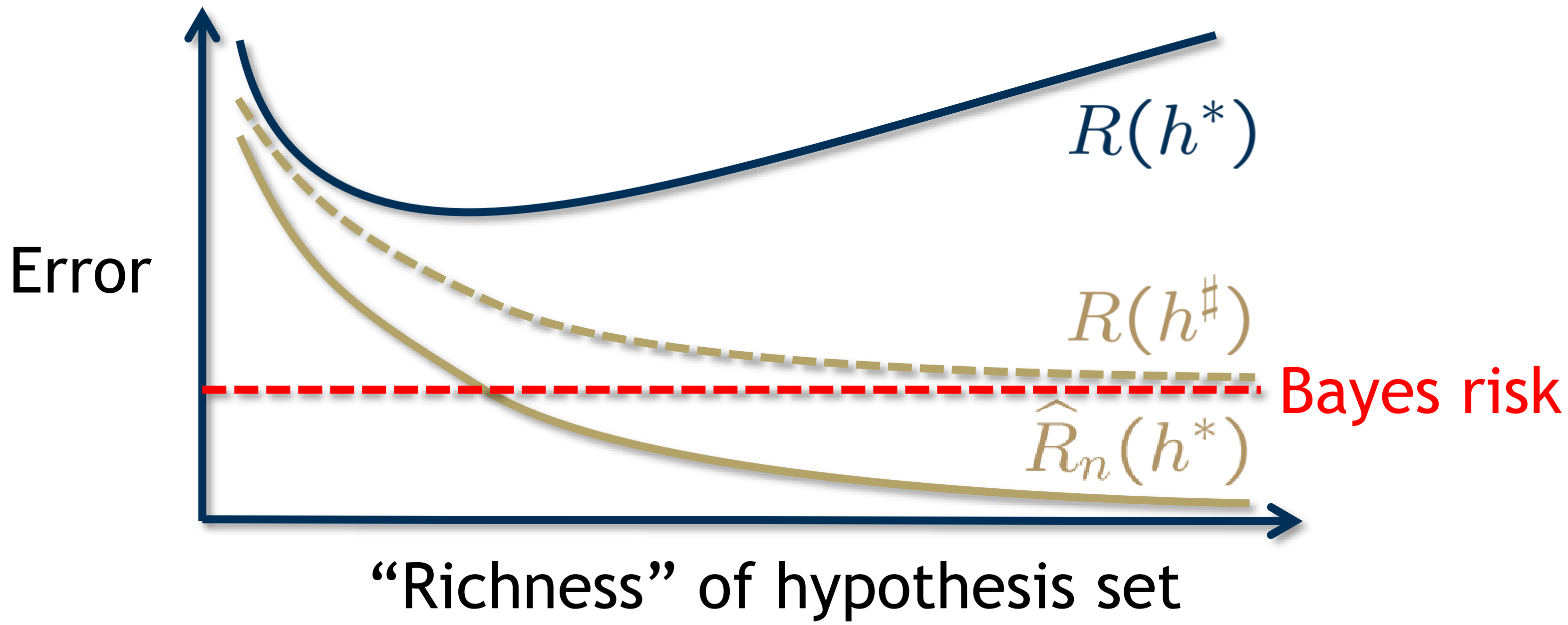
Tradeoffs in learning



Measuring “richness”

Today we will turn back to the question of when we can have confidence that $\widehat{R}_n(h^*) \approx R(h^*)$, but where h^* is chosen from an ***infinite*** set $\mathcal{H}$

To keep life (much) simpler, we will restrict our attention to binary classification, but an analogous theory can be developed for other supervised learning problems

- For a single hypothesis, we have

$$\mathbb{P}\left[\left|\widehat{R}_n(h) - R(h)\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

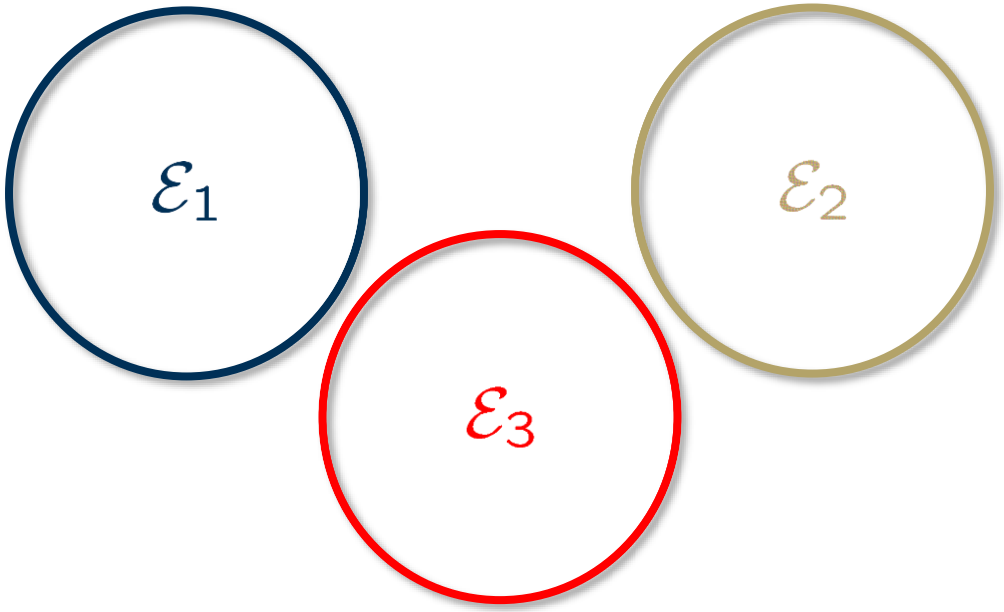
- For $m = |\mathcal{H}|$ hypotheses, and $h^* \in \mathcal{H}$, we have

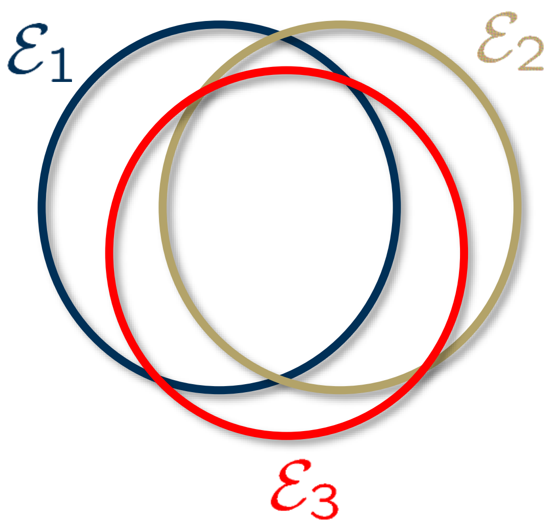
$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| > \epsilon\right] \leq 2me^{-2\epsilon^2 n}$$

Where did m come from?

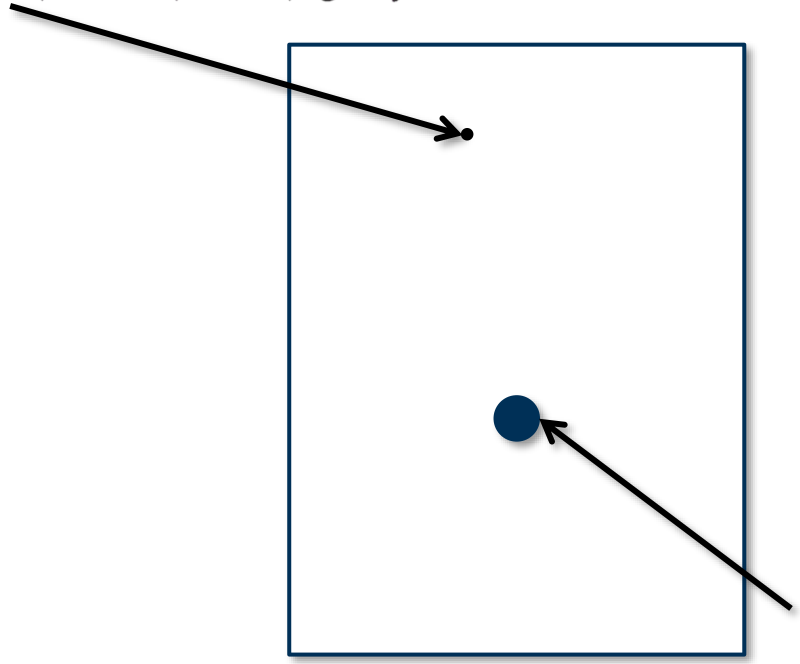
$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| > \epsilon\right] \leq \mathbb{P}\left[\max_{h_j \in \mathcal{H}} \left|\widehat{R}_n(h_j) - R(h_j)\right| \geq \epsilon\right]$$

$$\leq \sum_{j=1}^{m} \mathbb{P}\Big[\underbrace{\left|\widehat{R}_n(h_j) - R(h_j)\right| \geq \epsilon}_{\mathcal{E}_j}\Big]$$



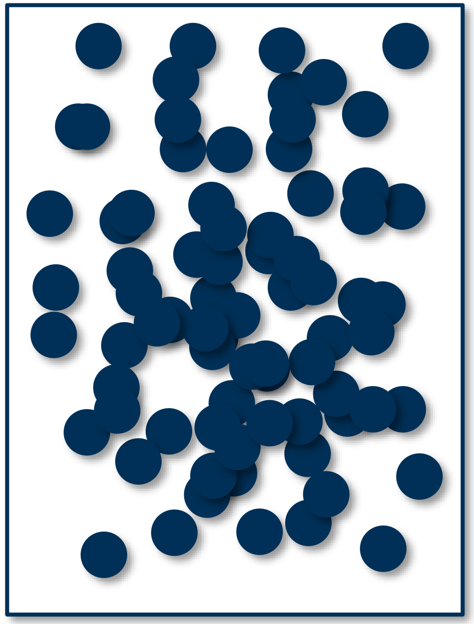


Visualizing Hoeffding

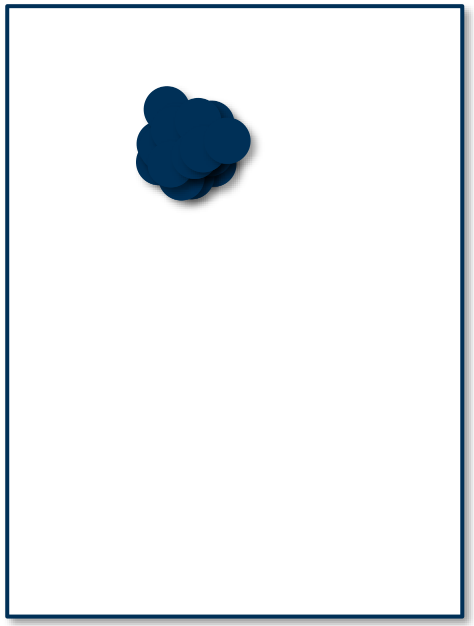


Union bound intuition

Consider many different h at once



An alternative picture



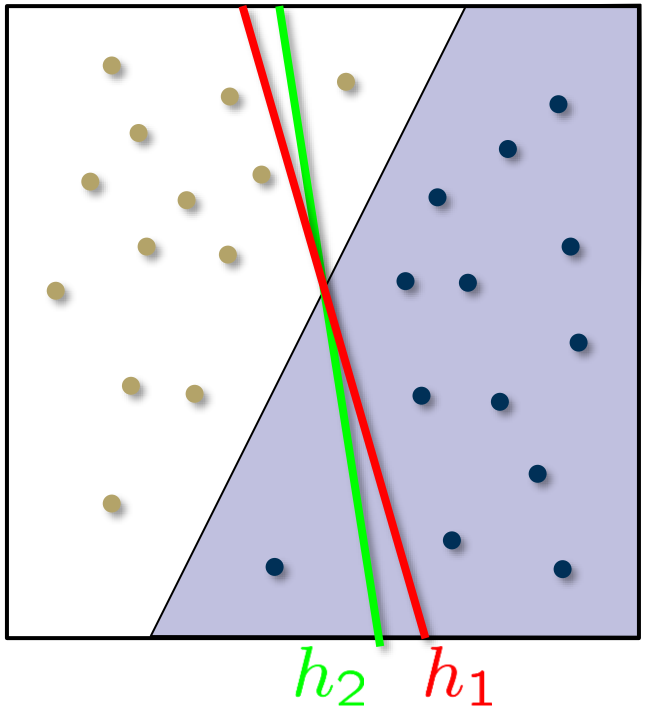
If all the “bad” datasets overlap, maybe we can handle much bigger $\mathcal{H}$ than the union bound suggests

Do “bad” datasets overlap?

Yes. There is (potentially) tremendous overlap!

$$R(h_1) \approx R(h_2)$$

$$\widehat{R}_n(h_1) \approx \widehat{R}_n(h_2)$$

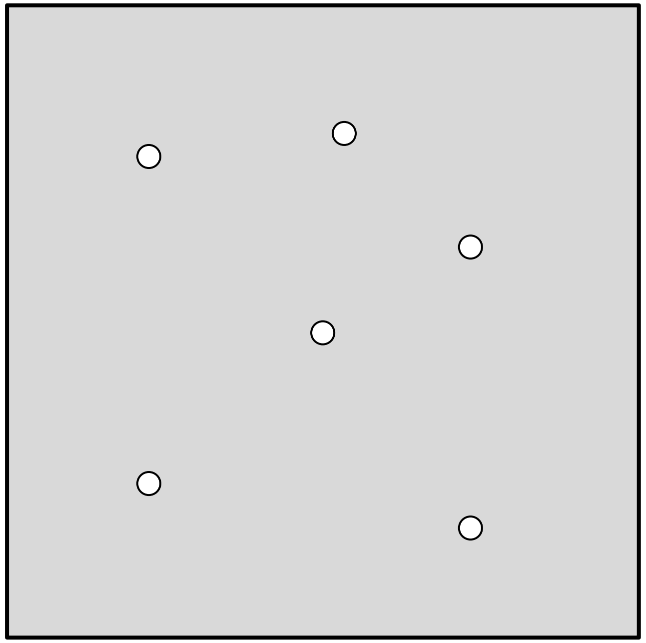


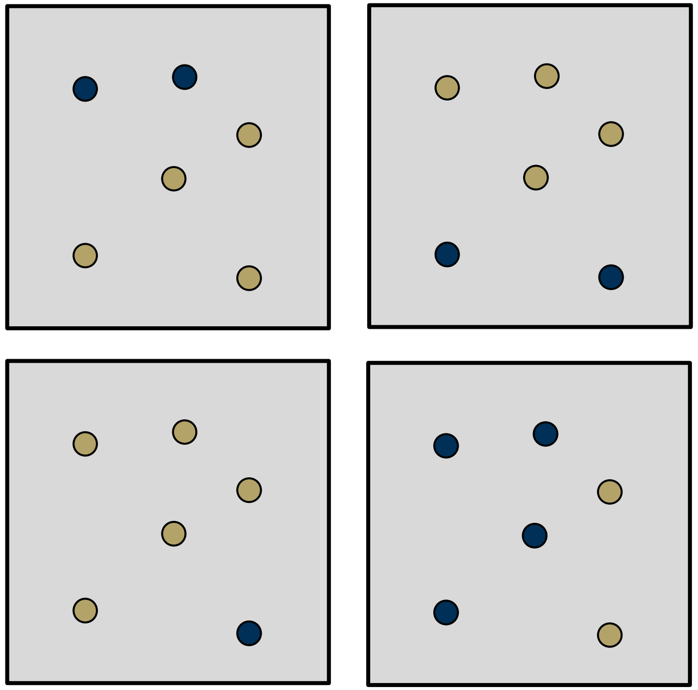
$$|\widehat{R}_n(h_1) - R(h_1)| \approx |\widehat{R}_n(h_2) - R(h_2)|$$

If not m, what?

Instead of considering all possible hypotheses in $\mathcal{H}$
we will consider a finite set of input points $\mathbf{x}_1, \ldots, \mathbf{x}_n$
and "combine" hypotheses that result in the same labeling

We will call a particular labeling of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ a ***dichotomy***





Hypotheses vs dichotomies

Hypotheses

- $h : \mathcal{X} \to \{-1, +1\}$
- Number of hypotheses $|\mathcal{H}|$ can be infinite

$|\mathcal{H}|$ (or m) is a poor way to measure "richness" of $\mathcal{H}$

Dichotomies

- $h : \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \to \{-1, +1\}$
- Number of dichotomies $|\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n)|$ is at most 2^n

Good candidate for replacing $|\mathcal{H}|$ as a measure of "richness"

The growth function

A dichotomy is defined in terms of a particular $\mathbf{x}_1, \ldots, \mathbf{x}_n$

We would like to be able to state results that hold no matter what $\mathbf{x}_1, \ldots, \mathbf{x}_n$ turn out to be

Define the ***growth function*** of $\mathcal{H}$ as

$$m_{\mathcal{H}}(n) := \max_{\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n)|$$

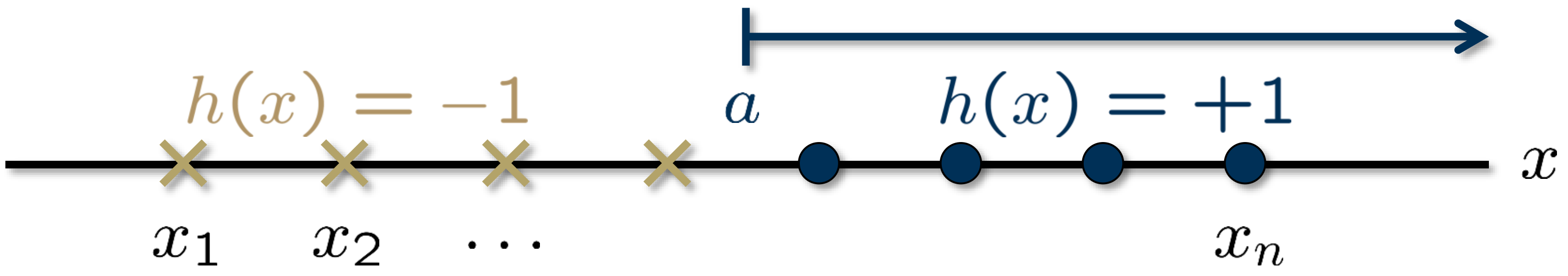
$m_{\mathcal{H}}(n)$ counts the ***most*** dichotomies that can possibly be generated on n points

It is easy to see that $m_{\mathcal{H}}(n) \leq 2^n$, but it can potentially be much smaller

Example 1: Positive rays

Candidate functions: $h : \mathbb{R} \to \{-1, +1\}$ such that

$h(x) = \text{sign}(x - a)$ for some $a \in \mathbb{R}$

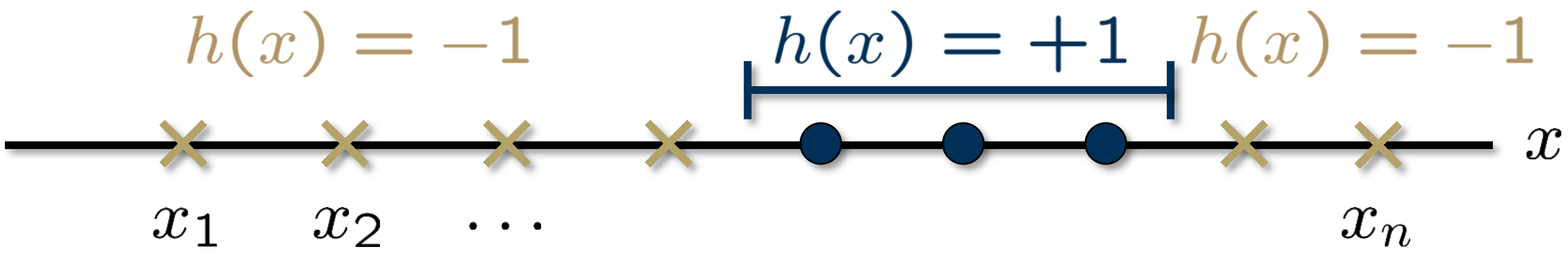


$$m_{\mathcal{H}}(n) = n + 1$$

Example 2: Positive intervals

Candidate functions: $h : \mathbb{R} \to \{-1, +1\}$ such that

$$h(x) = \begin{cases} +1 & \text{for } x \in [a, b] \\ -1 & \text{otherwise} \end{cases}$$

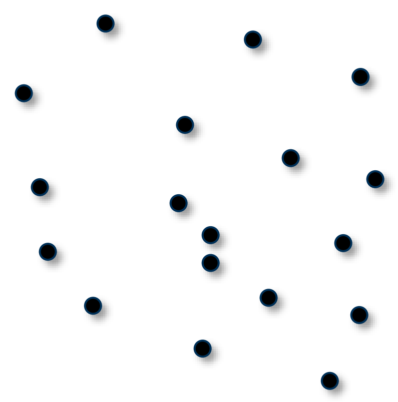


$$m_{\mathcal{H}}(n) = \binom{n+1}{2} + 1$$

$$= \tfrac{1}{2}n^2 + \tfrac{1}{2}n + 1$$

Example 3: Convex sets

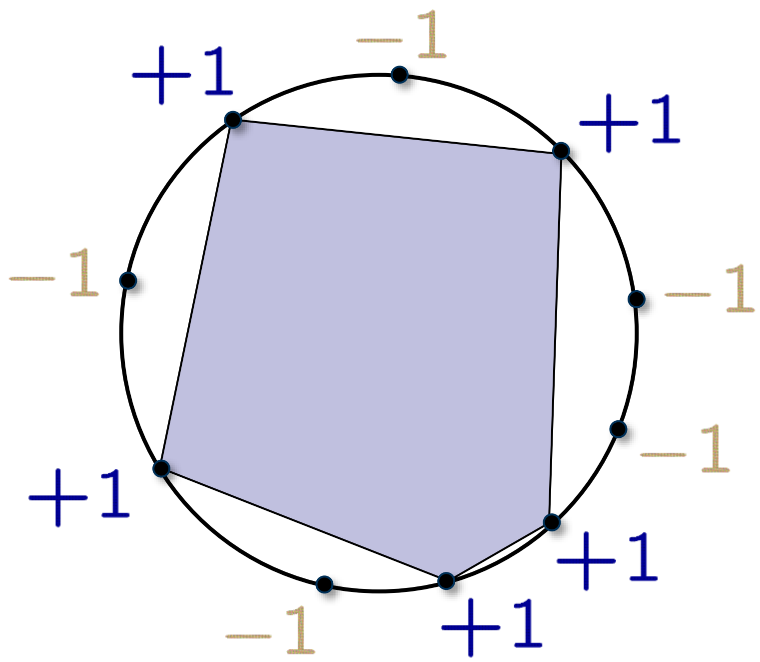
Candidate functions: $h : \mathbb{R}^2 \to \{-1, +1\}$ such that $\{\mathbf{x} : h(\mathbf{x}) = +1\}$ is convex



Example 3: Convex sets

Candidate functions: $h : \mathbb{R}^2 \to \{-1, +1\}$ such that

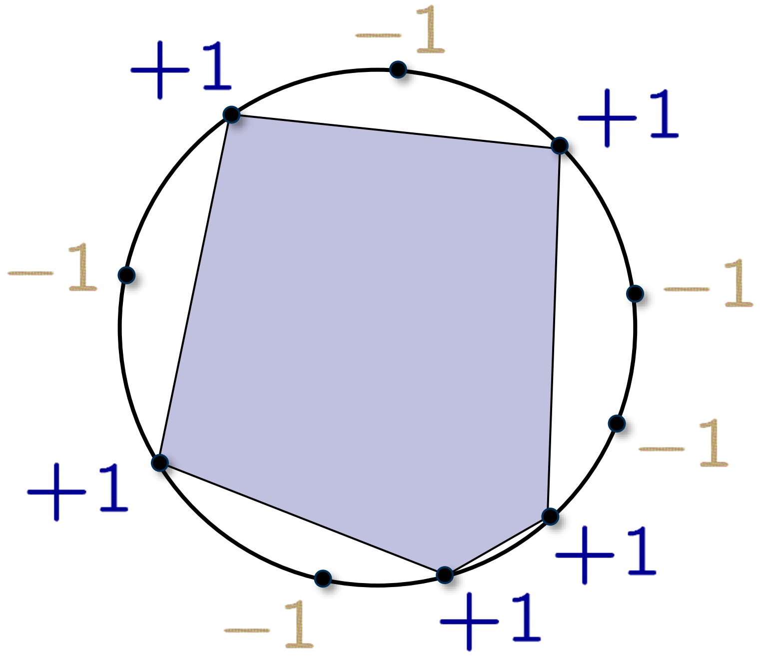
$\{\mathbf{x} : h(\mathbf{x}) = +1\}$ is convex



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

Example 3: Convex sets

Candidate functions: $h : \mathbb{R}^2 \to \{-1, +1\}$ such that $\{\mathbf{x} : h(\mathbf{x}) = +1\}$ is convex

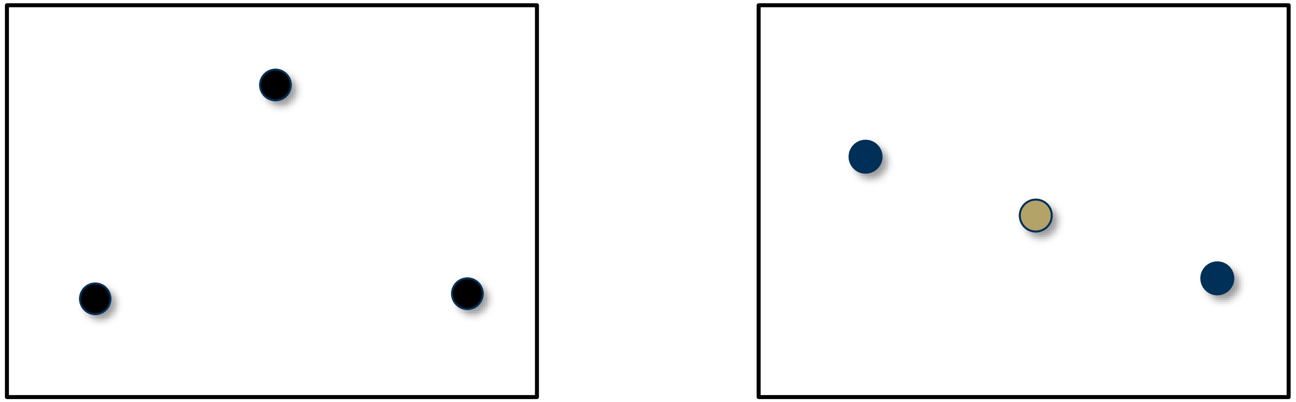


If $\mathcal{H}$ can generate all possible dichotomies on $\mathbf{x}_1, \ldots, \mathbf{x}_n$, then we say that $\mathcal{H}$ ***shatters*** $\mathbf{x}_1, \ldots, \mathbf{x}_n$

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

Example 4: Linear classifiers

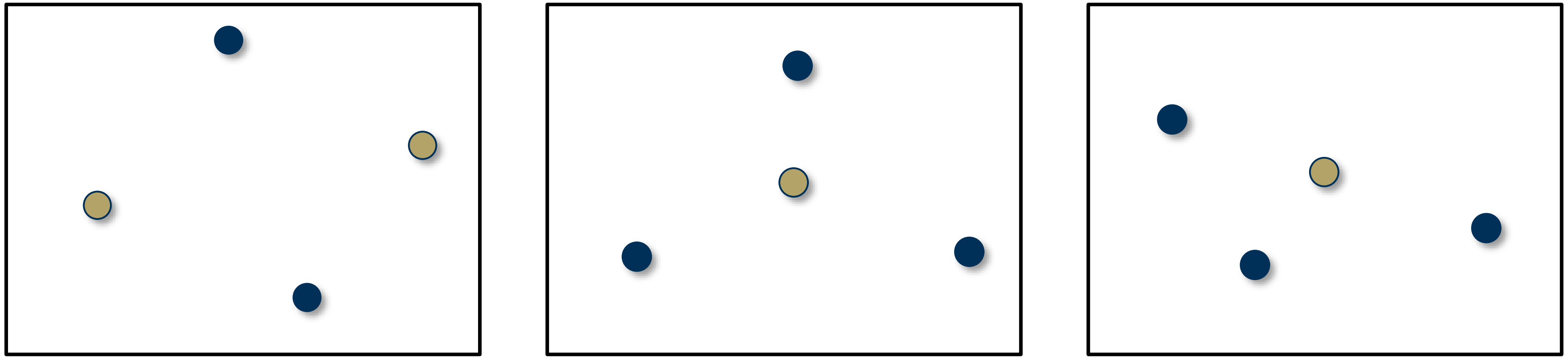
Candidate functions: $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$ such that
$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ for some
$\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$



$$m_{\mathcal{H}}(3) = 2^3$$

Example 4: Linear classifiers

Candidate functions: $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$ such that $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ for some $\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$



$$m_{\mathcal{H}}(4) = 14$$

Recap: Example growth functions

- Positive rays: $m_{\mathcal{H}}(n) = n + 1$
- Positive intervals: $m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1$
- Convex sets: $m_{\mathcal{H}}(n) = 2^n$
- Linear classifiers in $\mathbb{R}^2$: $m_{\mathcal{H}}(1) = 2$

$$m_{\mathcal{H}}(2) = 4$$
$$m_{\mathcal{H}}(3) = 8$$
$$m_{\mathcal{H}}(4) = 14$$
$$m_{\mathcal{H}}(n) = ?$$

Back to the big picture

Recall

$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| > \epsilon\right] \leq 2me^{-2\epsilon^2 n}$$

Another way to express this is that if you pick a δ, then we can guarantee that with probability at least $1 - \delta$

$$R(h^*) \leq \widehat{R}_n(h^*) + \sqrt{\frac{1}{2n} \log \frac{2m}{\delta}}$$

(Just set $2me^{-2\epsilon^2 n} = \delta$ and solve for ϵ)

If $m \propto e^n$, we have a problem…

No matter how big n gets, $\sqrt{\frac{1}{2n} \log \frac{2m}{\delta}}$ will never get any smaller…

What if... ?

What if we can replace m with $m_{\mathcal{H}}(n)$?

In particular, suppose that for any $\delta \in (0, 1)$, we can guarantee that with probability at least $1 - \delta$

$$R(h^*) \leq \widehat{R}_n(h^*) + \sqrt{\frac{1}{2n} \log \frac{2m_{\mathcal{H}}(n)}{\delta}}$$

- If $m_{\mathcal{H}}(n) = 2^n$, $\sqrt{\frac{1}{2n} \log \frac{2m_{\mathcal{H}}(n)}{\delta}}$ is a constant
- If $m_{\mathcal{H}}(n)$ is a polynomial in n, $\sqrt{\frac{1}{2n} \log \frac{2m_{\mathcal{H}}(n)}{\delta}}$ decays like $\sqrt{\frac{\log n}{n}}$

When is learning possible?

Assuming that we will indeed be allowed to substitute $m_{\mathcal{H}}(n)$ for m, we can argue that for a given set of hypotheses $\mathcal{H}$, learning is possible provided that is a polynomial $m_{\mathcal{H}}(n)$

Key idea: *Break points*

If no data set of size k can be shattered by $\mathcal{H}$, then k is a ***break point*** for $\mathcal{H}$

$$m_{\mathcal{H}}(k) < 2^k$$

If k is a break point, then so is any $k' > k$

Examples

- Positive rays: $m_{\mathcal{H}}(n) = n + 1$
  - break point: $k = 2$

- Positive intervals: $m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1$
  - break point: $k = 3$

- Convex sets: $m_{\mathcal{H}}(n) = 2^n$
  - break point: $k = \infty$
- Linear classifiers in $\mathbb{R}^2$: $m_{\mathcal{H}}(3) = 8$

  $m_{\mathcal{H}}(4) = 14$
  - break point: $k = 4$

So what?

If there exists any break point,
then $m_{\mathcal{H}}(n)$ is polynomial in n

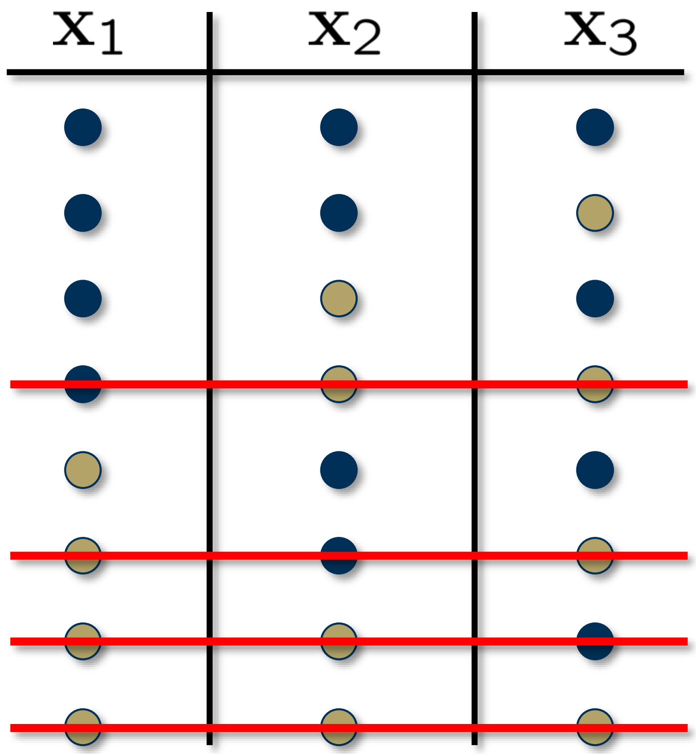
Also, if there are no break points, then $m_{\mathcal{H}}(n) = 2^n$

As soon as we have ***a single break point***, this starts eliminating tons of dichotomies

How many dichotomies?

You are given a hypothesis set which has a break point of 2

How many dichotomies can you get on 3 data points?



Bounding the growth function

We want to show that $m_{\mathcal{H}}(n)$ is polynomial in n

We will show that $m_{\mathcal{H}}(n) \leq$ ***some*** polynomial

Our approach will center around

$B(n,k) :=$ maximum number of dichotomies on n points such that no subset of size k can be shattered by these dichotomies

$B(n,k)$ is a purely combinatorial quantity

By definition, $m_{\mathcal{H}}(n) \leq B(n,k)$

Sauer's Lemma

Theorem If k is a break point, then

$$m_{\mathcal{H}}(n) \leq B(n,k) \leq \sum_{i=0}^{k-1} \binom{n}{i}$$

In fact, it is actually true that

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

but all we really need is the upper bound

Examples

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{k-1} \binom{n}{i}$$

- Positive rays: Break point of $k = 2$

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

- Positive intervals: Break point of $k = 3$

$$m_{\mathcal{H}}(n) = \tfrac{1}{2}n^2 + \tfrac{1}{2}n + 1 \leq \tfrac{1}{2}n^2 + \tfrac{1}{2}n + 1$$

- Linear classifiers in $\mathbb{R}^2$: Break point of $k = 4$

$$m_{\mathcal{H}}(n) = \ ? \ \leq \tfrac{1}{6}n^3 + \tfrac{5}{6}n + 1$$

Bottom line

For a given $\mathcal{H}$, all we need is for a break point to exist

$$m_{\mathcal{H}}(n) \leq \underbrace{\sum_{i=0}^{k-1} \binom{n}{i}}$$

polynomial with dominant term n^{k-1}

All that remains is to argue that we can actually replace $|\mathcal{H}|$ with $m_{\mathcal{H}}(n)$ to obtain an inequality along the lines of

$$R(h^*) \leq \widehat{R}_n(h^*) + \sqrt{\frac{1}{2n} \log \frac{2m_{\mathcal{H}}(n)}{\delta}}$$

VC generalization bound

We won't be able to quite show

$$R(h^*) \leq \widehat{R}_n(h^*) + \sqrt{\frac{1}{2n} \log \frac{2m_{\mathcal{H}}(n)}{\delta}}$$

For technical reasons (which we see next time), we will only be able to show that with probability $\geq 1 - \delta$

$$R(h^*) \leq \widehat{R}_n(h^*) + \sqrt{\frac{8}{n} \log \frac{4m_{\mathcal{H}}(2n)}{\delta}}$$

This is called the ***VC generalization bound***

Named after Vapnik and Chervonenkis, who proved it in 1971

Key difference

Using Hoeffding's inequality together with a union bound, we were able to show that

$$\mathbb{P}\left[\max_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \leq |\mathcal{H}| \cdot 2e^{-2\epsilon^2 n}$$

What the VC bound gives us is a generalization of the form

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \leq 2 \cdot m_{\mathcal{H}}(2n) \cdot 2e^{-\frac{1}{8}\epsilon^2 n}$$

supremum: maximum over an infinite set

Supremum

The ***supremum*** of a set $S \subset T$ is the least element of T that is greater than or equal to all elements of S

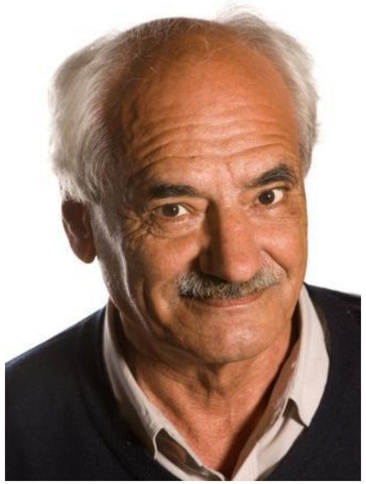
Sometimes called the ***least upper bound***

Examples

- $\sup\{1, 2, 3\} = 3$
- $\sup\{x : 0 \leq x \leq 1\} = 1$
- $\sup\{x : 0 < x < 1\} = 1$
- $\sup\{1 - 1/n : n > 0\} = 1$

Next time

Deep Neural Networks Hate Them!



They can turn the **supremum** over an **infinite** set into a **maximum** over a **finite** set *and* establish powerful generalization bounds using this **ONE WEIRD TRICK**