Recap

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

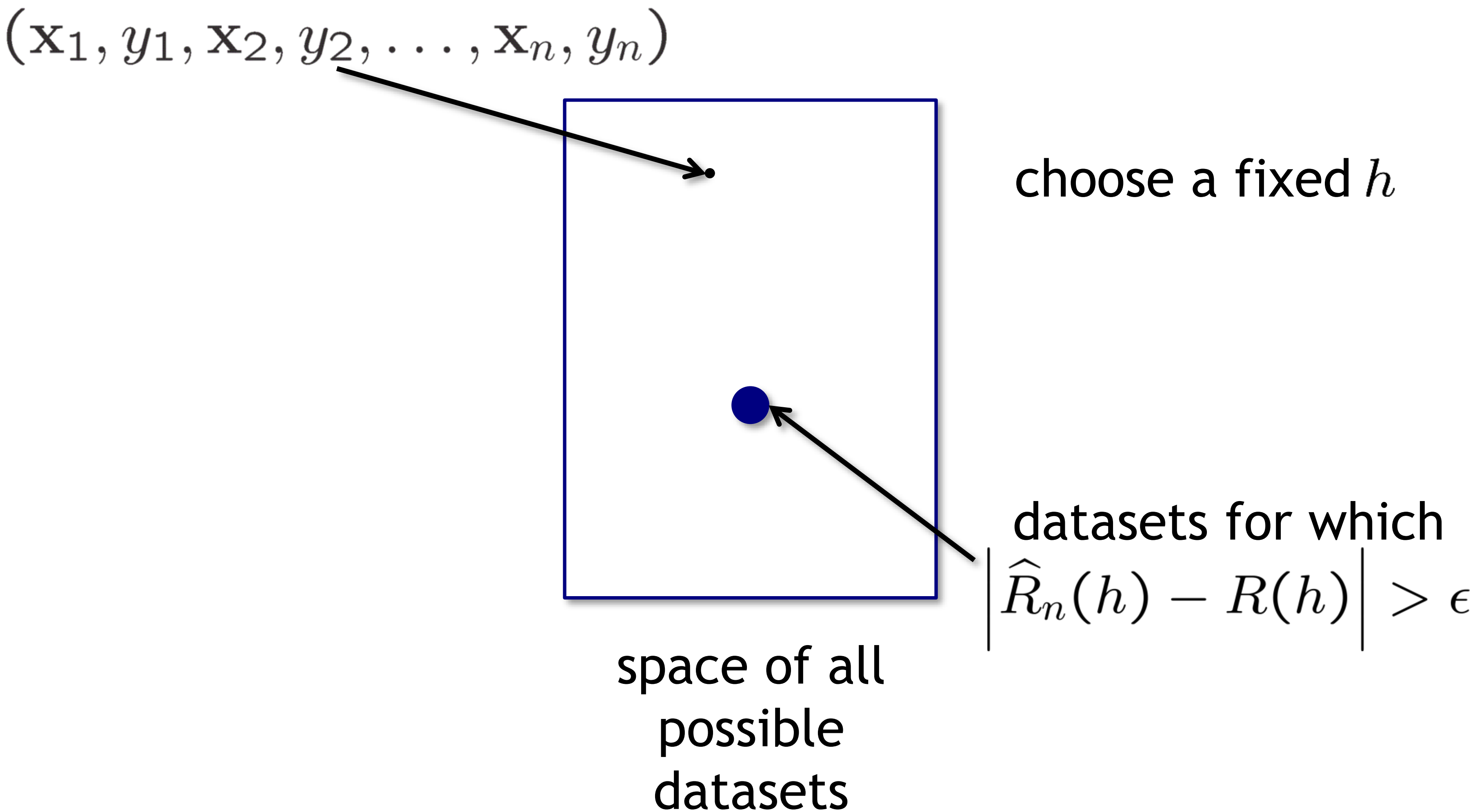
$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| > \epsilon\right] \leq 2me^{-2\epsilon^2 n}$$

or equivalently, that with probability at least $1 - \delta$

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

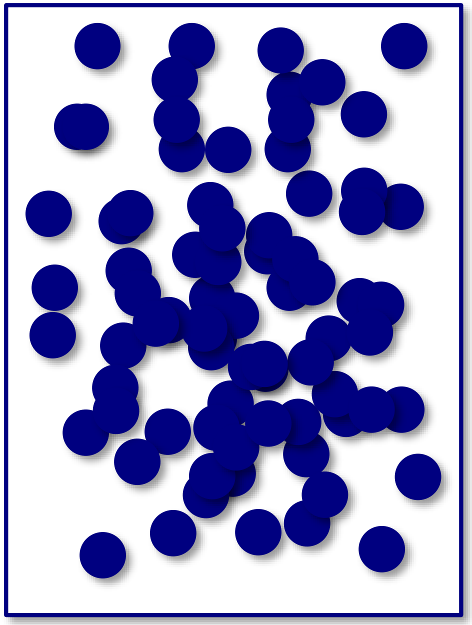
Bound becomes meaningless when $|\mathcal{H}| = \infty$

Hoeffding's inequality

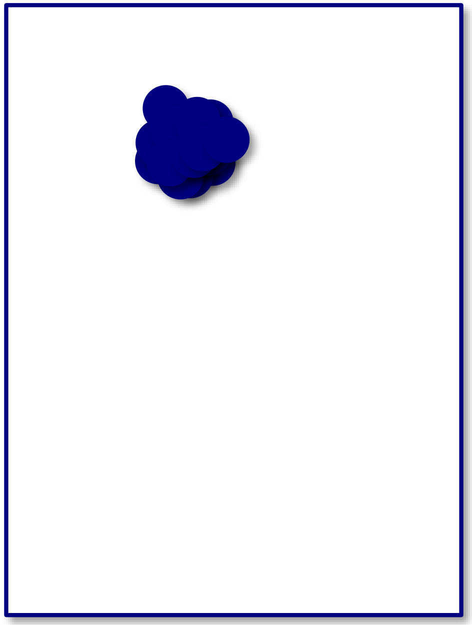


Union bound intuition

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

Big idea from last time

Rather than measuring the "size" of $\mathcal{H}$ with $|\mathcal{H}|$, we can instead think about:

Using $\mathcal{H}$, how many ways can we label a dataset?

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

Using this language, we can answer our question via the growth function $m_{\mathcal{H}}(n)$, which counts the ***most*** dichotomies that $\mathcal{H}$ could ever generate on n points

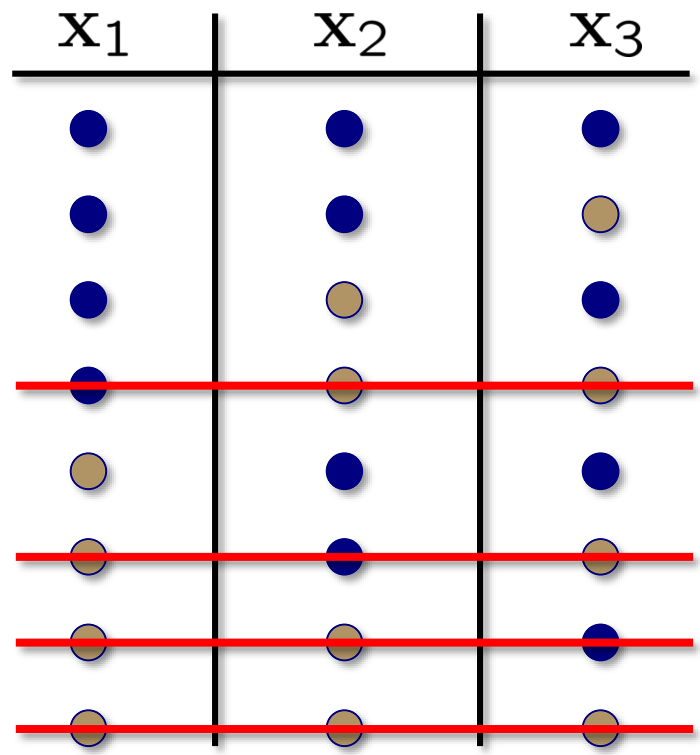
It is easy to see that $m_{\mathcal{H}}(n) \leq 2^n$

- If $m_{\mathcal{H}}(n) = 2^n$, we say that $\mathcal{H}$ can ***shatter*** a set of size n
- If no set of size k can be shattered by $\mathcal{H}$ ($m_{\mathcal{H}}(k) < 2^k$) then k is a ***break point***

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?



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 will soon see), 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

Mathematical statement

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

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

The magic in the proof of the VC bound is to realize that we can relate the ***supremum*** over ***all*** $h \in \mathcal{H}$ to the ***maximum*** over a finite number of $h \in \mathcal{H}$ using a really cool trick!

Role of the growth function

We aim to get a bound on $\mathbb{P}[|\widehat{R}_n(h) - R(h)| > \epsilon]$ that holds for any $h \in \mathcal{H}$, i.e., a bound on

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right]$$

Perhaps it is not surprising that we can understand $\widehat{R}_n(h)$ using the growth function…

There may be infinitely many $h \in \mathcal{H}$, but $\mathcal{H}$ can only generate $m_{\mathcal{H}}(n)$ ***unique dichotomies***

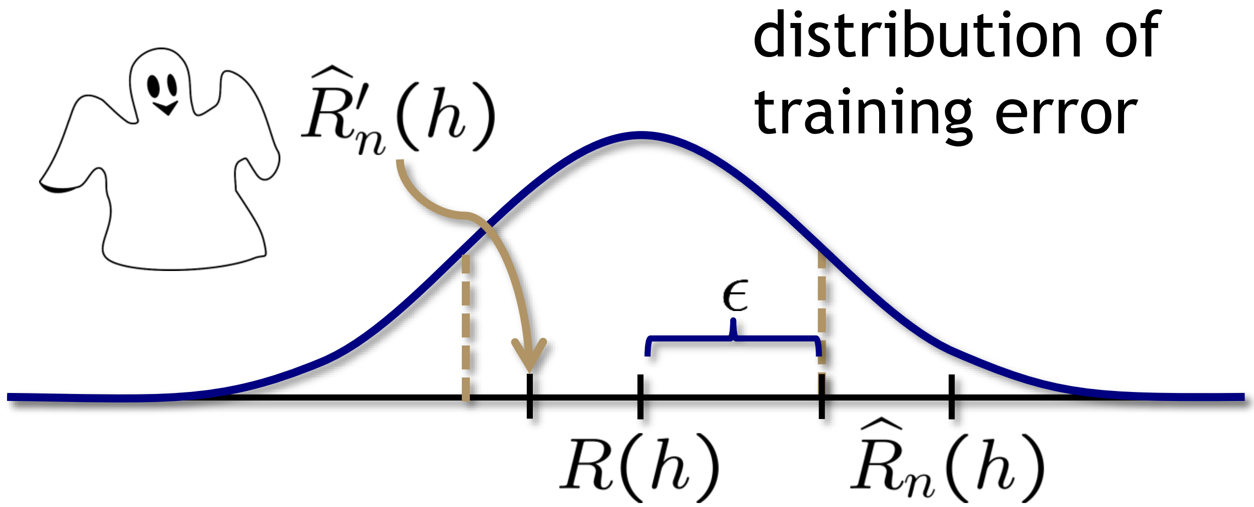
Thus, $\widehat{R}_n(h)$ can only take at most $m_{\mathcal{H}}(n)$ different values

Unfortunately, $R(h)$ can still take infinitely many different values, and so there are infinitely many $|\widehat{R}_n(h) - R(h)|$

Fundamental insight

The key insight (or trick) is to consider ***two*** datasets!

We will imagine that in addition to our training data, we have access to a second independent dataset (of size n), which we call the ***ghost dataset***



Can we relate $\mathbb{P}[|\widehat{R}_n(h) - R(h)| > \epsilon]$ to something like $\mathbb{P}[|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \epsilon]$?

Using the ghost dataset

Suppose (for the moment) that the empirical estimates $\widehat{R}_n(h)$ and $\widehat{R}'_n(h)$ are random variables that are drawn from a *symmetric* distribution with mean (and median) $R(h)$

Consider the following events:

- A : the event that $|\widehat{R}_n(h) - R(h)| > \epsilon$
- B : the event that $|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \epsilon$

Claim: $\mathbb{P}[B|A] \geq \frac{1}{2}$

Thus $\mathbb{P}[B] = \mathbb{P}[B|A] \cdot \mathbb{P}[A] \geq \frac{1}{2}\mathbb{P}[A]$

$\Rightarrow$ $\mathbb{P}[|\widehat{R}_n(h) - R(h)| > \epsilon] \leq 2\mathbb{P}[|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \epsilon]$

Using the ghost dataset

Unfortunately, the distribution of $\widehat{R}_n(h)$ and $\widehat{R}'_n(h)$ is binomial (not symmetric) so this exact statement doesn't hold in general, but the intuition is valid

Instead, we have the following bound:

Lemma 1 (Ghost dataset)

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \leq 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\right]$$

Bounding the worst-case deviation

Lemma 2 (Where the magic happens)

Let $S = \{(\mathbf{x}_i, y_i), i = 1, \ldots, 2n\}$. Then

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} |\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\right]$$

$$\leq m_{\mathcal{H}}(2n) \cdot \sup_{S} \sup_{h \in \mathcal{H}} \underbrace{\mathbb{P}\left[|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2} \middle| S\right]}$$

here, the dataset S is fixed
the probability is with respect to a random partition of S into two training sets of size n

Proof of Lemma 2

It is straightforward to show that

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\right]$$
$$\leq \sup_{S} \mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\middle| S\right]$$

Note that in the probability on the right-hand side, the dataset S is fixed.

Thus, there are only a finite number of dichotomies that $\mathcal{H}$ can generate on S. Call this number $m_{\mathcal{H}}(S)$.

Let $h_1, \ldots, h_{m_{\mathcal{H}}(S)}$ be the classifiers giving rise to these dichotomies

Proof of Lemma 2

Using this observation, we have

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-\widehat{R}'_n(h)|>\frac{\epsilon}{2}\middle|S\right]$$

$$=\mathbb{P}\left[\max_{h_1,\dots,h_{m_{\mathcal{H}}(S)}}|\widehat{R}_n(h_i)-\widehat{R}'_n(h_i)|>\frac{\epsilon}{2}\middle|S\right]$$

$$\leq\sum_{i=1}^{m_{\mathcal{H}}(S)}\mathbb{P}\left[|\widehat{R}_n(h_i)-\widehat{R}'_n(h_i)|>\frac{\epsilon}{2}\middle|S\right]$$

$$\leq m_{\mathcal{H}}(S)\max_{h_1,\dots,h_{m_{\mathcal{H}}(S)}}\mathbb{P}\left[|\widehat{R}_n(h_i)-\widehat{R}'_n(h_i)|>\frac{\epsilon}{2}\middle|S\right]$$

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

Final step

At this point, we have shown

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|\widehat{R}_n(h) - R(h)| > \epsilon\right]$$

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

Lemma 3 (Random partitions)

For ***any*** h and ***any*** S,

$$\mathbb{P}\left[|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\middle|S\right] \leq 2e^{-\frac{1}{8}\epsilon^2 n}$$

Proof follows from a simple lemma (also by Hoeffding)

Putting it all together

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-R(h)|>\epsilon\right]$$

$$\leq 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-\widehat{R}'_n(h)|>\frac{\epsilon}{2}\right]$$

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

$$\leq 2\cdot m_{\mathcal{H}}(2n)\cdot 2e^{-\frac{1}{8}\epsilon^2 n}$$

Thus, for any $h\in\mathcal{H}$, we have that with probability $\geq 1-\delta$

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

Using the VC bound: The VC dimension

We went to a lot of trouble to show that if k is a break point for $\mathcal{H}$, then $m_{\mathcal{H}}(n) \leq \sum_{i=0}^{k-1} \binom{n}{i} \leq n^{k-1} + 1$

$$\Rightarrow R(h) \leq \widehat{R}_n(h) + \sqrt{\frac{8}{n} \log \frac{4((2n)^{k-1}+1)}{\delta}}$$

$$\lessapprox \widehat{R}_n(h) + \sqrt{\frac{8(k-1)}{n} \log \frac{8n}{\delta}}$$

True for $k \geq 3$

The ***VC dimension*** of a hypothesis set $\mathcal{H}$, denoted $d_{\mathrm{VC}}(\mathcal{H})$, is the largest n for which $m_{\mathcal{H}}(n) = 2^n$

- $d_{\mathrm{VC}}(\mathcal{H})$ is the most points that $\mathcal{H}$ can shatter
- $d_{\mathrm{VC}}(\mathcal{H})$ is 1 less than the smallest break point

$$\Rightarrow R(h) \lessapprox \widehat{R}_n(h) + \sqrt{\frac{8 d_{\mathrm{VC}}}{n} \log \frac{8n}{\delta}}$$

Examples

- Positive rays:

$$d_{\mathsf{VC}} = 1$$

- Positive intervals:

$$d_{\mathsf{VC}} = 2$$

- Convex sets:

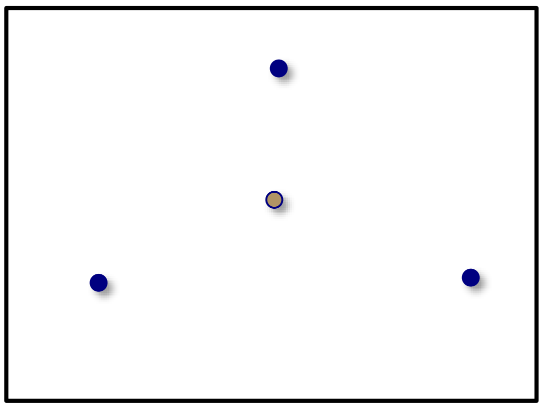
$$d_{\mathsf{VC}} = \infty$$

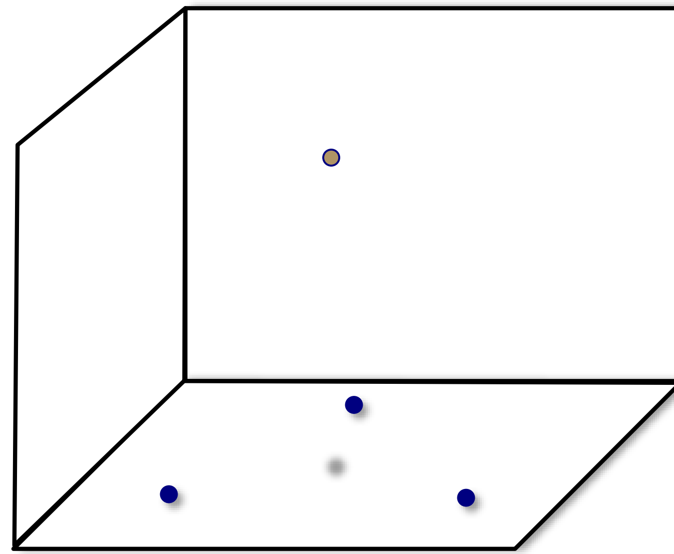
- Linear classifiers in $\mathbb{R}^2$:

$$d_{\mathsf{VC}} = 3$$

VC dimension of general linear classifiers

For $d = 2$, $d_{\text{VC}} = 3$





In general $d_{\text{VC}} = d + 1$

We will prove this by showing that $d_{\text{VC}} \leq d + 1$ and $d_{\text{VC}} \geq d + 1$

One direction

Lets first show that there exists a set of $d+1$ points in $\mathbb{R}^d$ that are shattered

$$\mathbf{X} = \underbrace{\begin{bmatrix} -\widetilde{\mathbf{x}}_1^T- \\ -\widetilde{\mathbf{x}}_2^T- \\ \vdots \\ -\widetilde{\mathbf{x}}_{d+1}^T- \end{bmatrix}}_{d+1} = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

One can show that $\mathbf{X}$ is invertible

Can we shatter this data set?

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find a vector

satisfying $\text{sign}(\mathbf{X}\boldsymbol{\theta}) = \mathbf{y}$?

Easy! Just make $\boldsymbol{\theta} = \mathbf{X}^{-1}\mathbf{y}$ and we have

$$\text{sign}(\mathbf{X}\boldsymbol{\theta}) = \text{sign}(\mathbf{y}) = \mathbf{y}$$

We can shatter a set of $d + 1$ points

What does this prove?

a) $d_{\text{VC}} = d + 1$

b) $d_{\text{VC}} \geq d + 1$ ✓

c) $d_{\text{VC}} \leq d + 1$

d) None of the above

To finish the proof

In order to show that $d_{\text{VC}} \leq d+1$, we need to show

a) There are $d+1$ points we cannot shatter

b) There are $d+2$ points we cannot shatter

c) We cannot shatter any set of $d+1$ points

d) We cannot shatter any set of $d+2$ points ✔

The other direction

Take any $d+2$ points $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_{d+2}$

More points than dimensions, so there must be some j for which

$$\widetilde{\mathbf{x}}_j = \sum_{i \neq j} \alpha_i \widetilde{\mathbf{x}}_i$$

where not all $\alpha_i = 0$

Consider the dichotomy where the $\widetilde{\mathbf{x}}_i$ with $\alpha_i \neq 0$ are labeled $y_i = \text{sign}(\alpha_i)$, and $y_j = -1$

No linear classifier can implement such a dichotomy!

Why not?

$$\widetilde{\mathbf{x}}_j = \sum_{i \neq j} \alpha_i \widetilde{\mathbf{x}}_i \quad \Rightarrow \quad \boldsymbol{\theta}^T \widetilde{\mathbf{x}}_j = \sum_{i \neq j} \alpha_i \boldsymbol{\theta}^T \widetilde{\mathbf{x}}_i$$

If $y_i = \text{sign}(\boldsymbol{\theta}^T \widetilde{\mathbf{x}}_i) = \text{sign}(\alpha_i)$, then $\alpha_i \boldsymbol{\theta}^T \widetilde{\mathbf{x}}_i > 0$

This means that $\boldsymbol{\theta}^T \widetilde{\mathbf{x}}_j = \sum_{i \neq j} \alpha_i \boldsymbol{\theta}^T \widetilde{\mathbf{x}}_i > 0$

Thus $y_j = \text{sign}(\boldsymbol{\theta}^T \widetilde{\mathbf{x}}_j) = +1$

Interpreting the VC dimension

We have just shown that for a linear classifier in $\mathbb{R}^d$

$$d_{\text{VC}} \geq d + 1$$

$$d_{\text{VC}} \leq d + 1$$

➡ $$d_{\text{VC}} = d + 1$$

How many parameters does a linear classifier in $\mathbb{R}^d$ have?

$$\mathbf{w} \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

➡ $$d + 1$$

The usual examples

- Positive rays
  - $d_{\text{VC}} = 1$
  - 1 parameter
- Positive intervals
  - $d_{\text{VC}} = 2$
  - 2 parameters
- Convex sets
  - $d_{\text{VC}} = \infty$
  - as many parameters as you want

Effective number of parameters

Additional parameters do not always contribute additional degrees of freedom

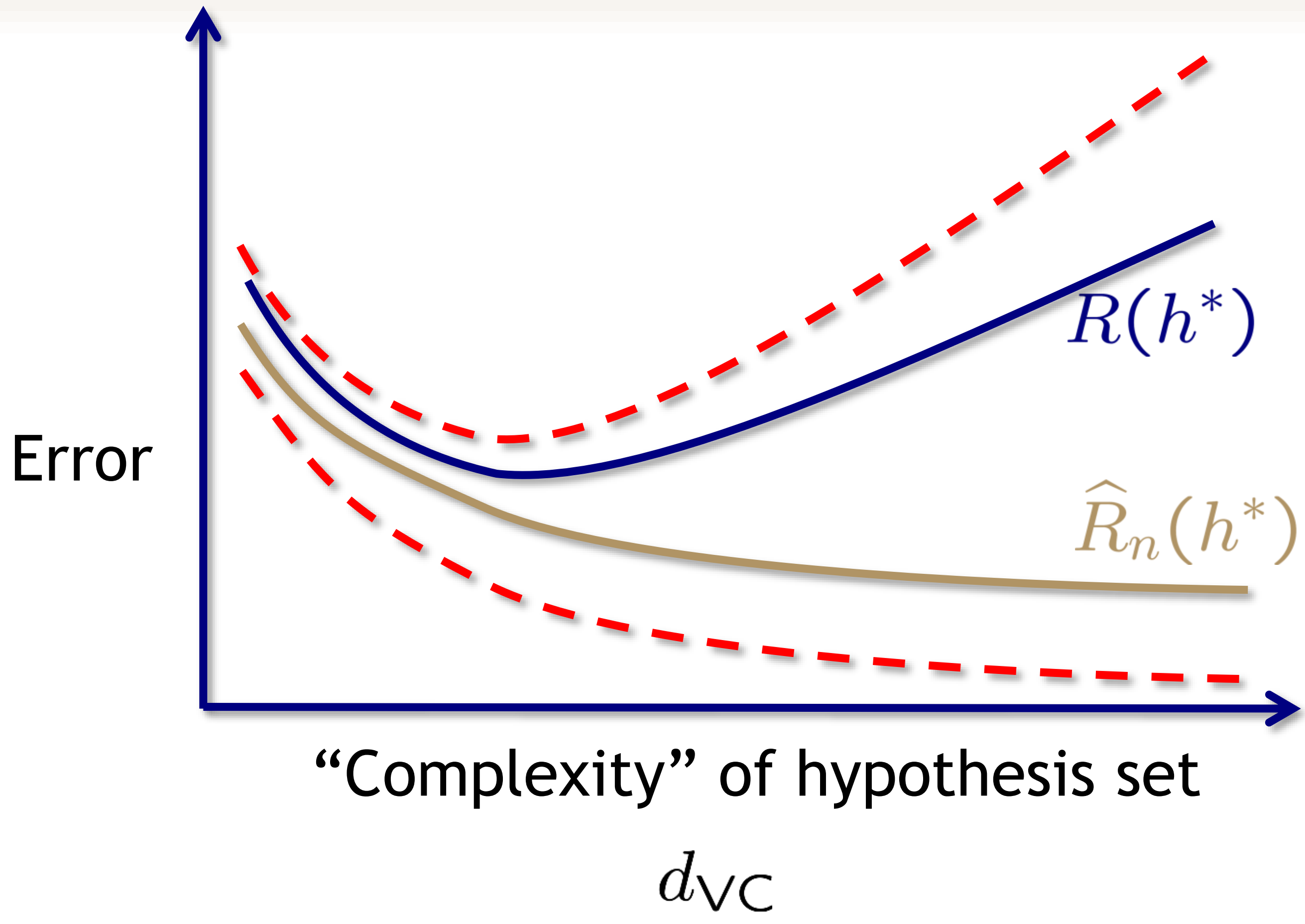
Example

Take the output of a linear classifier, and then feed this into another linear classifier

$$y_i = \text{sign}\left(w'\left(\text{sign}(\boldsymbol{\theta}^T\mathbf{x}_i)\right) + b'\right)$$

The parameters w' and b' are totally redundant (they do not allow us to create any new classifiers/dichotomies)

Interpreting the VC bound



VC bound in action

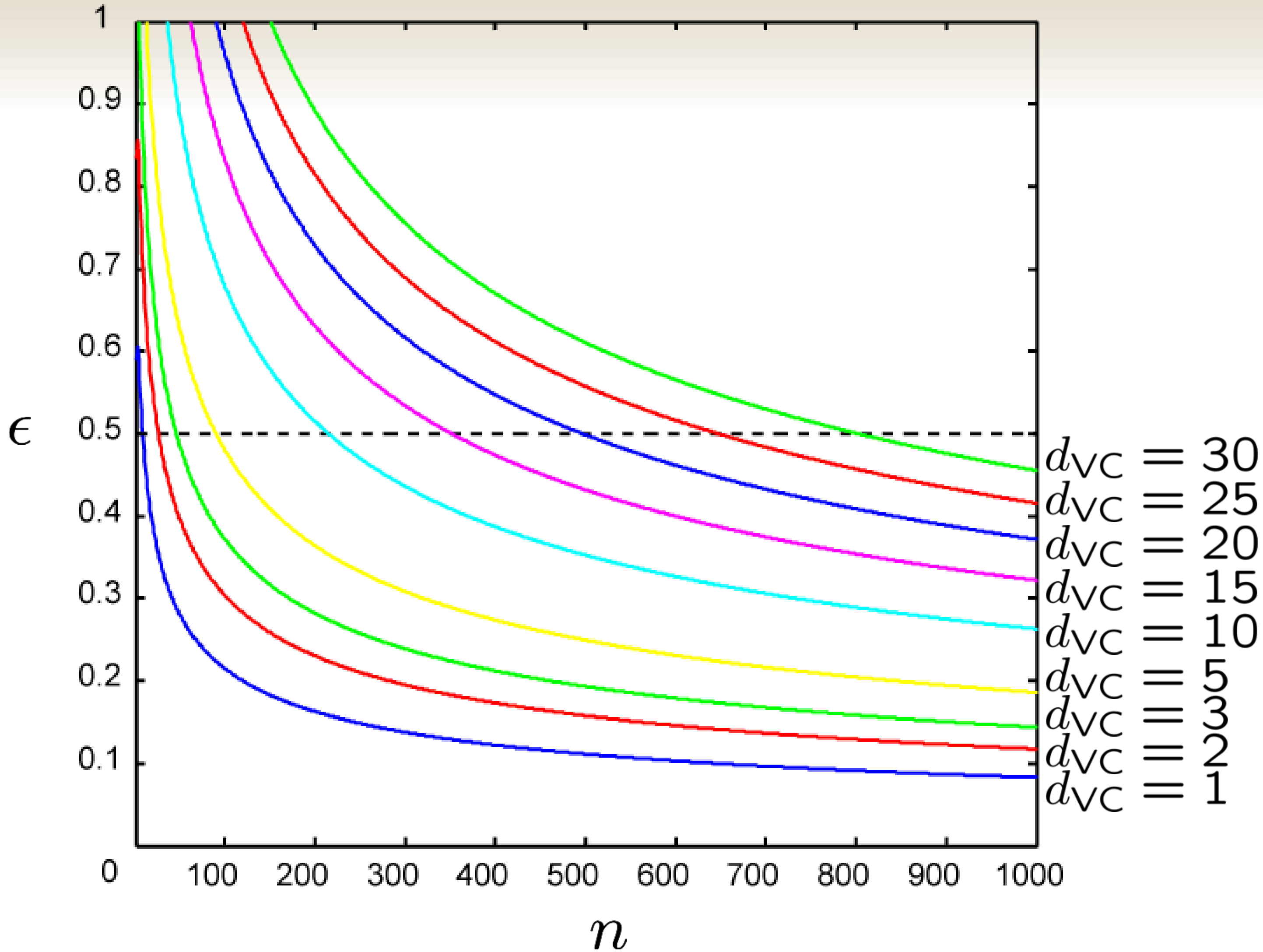
How big does our training set need to be?

$$R(h) \lessapprox \widehat{R}_n(h) + \underbrace{\sqrt{\frac{8d_{\mathrm{VC}}}{n} \log \frac{8n}{\delta}}}_{\epsilon}$$

Just to see how this behaves, let's ignore the constants and suppose that

$$\epsilon \sim \sqrt{\frac{d_{\mathrm{VC}}}{n} \log n}$$

VC bound in action



RULE OF THUMB: $n \geq 10 d_{VC}$