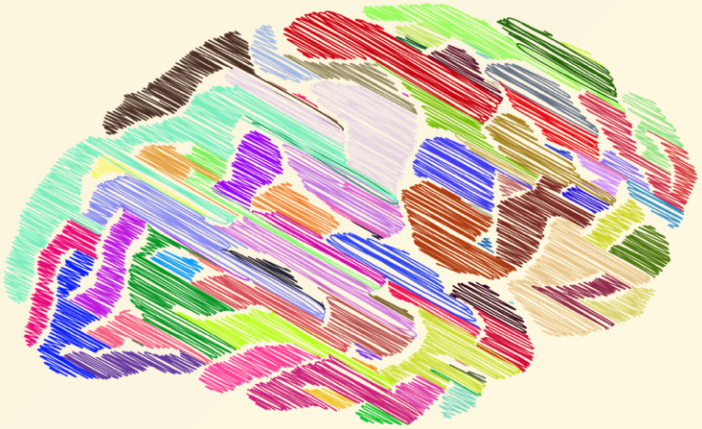
Statistical Learning Theory



Jeremy Bernstein
bernstein@caltech.edu

Agenda for today

1. What is generalisation?
2. How does VC theory work?
3. Inequalities that we'll need
4. The main result
5. What does VC theory miss?

Data source

We have a source of iid data samples.

We shake the tree and it gives us an input x and a binary label y

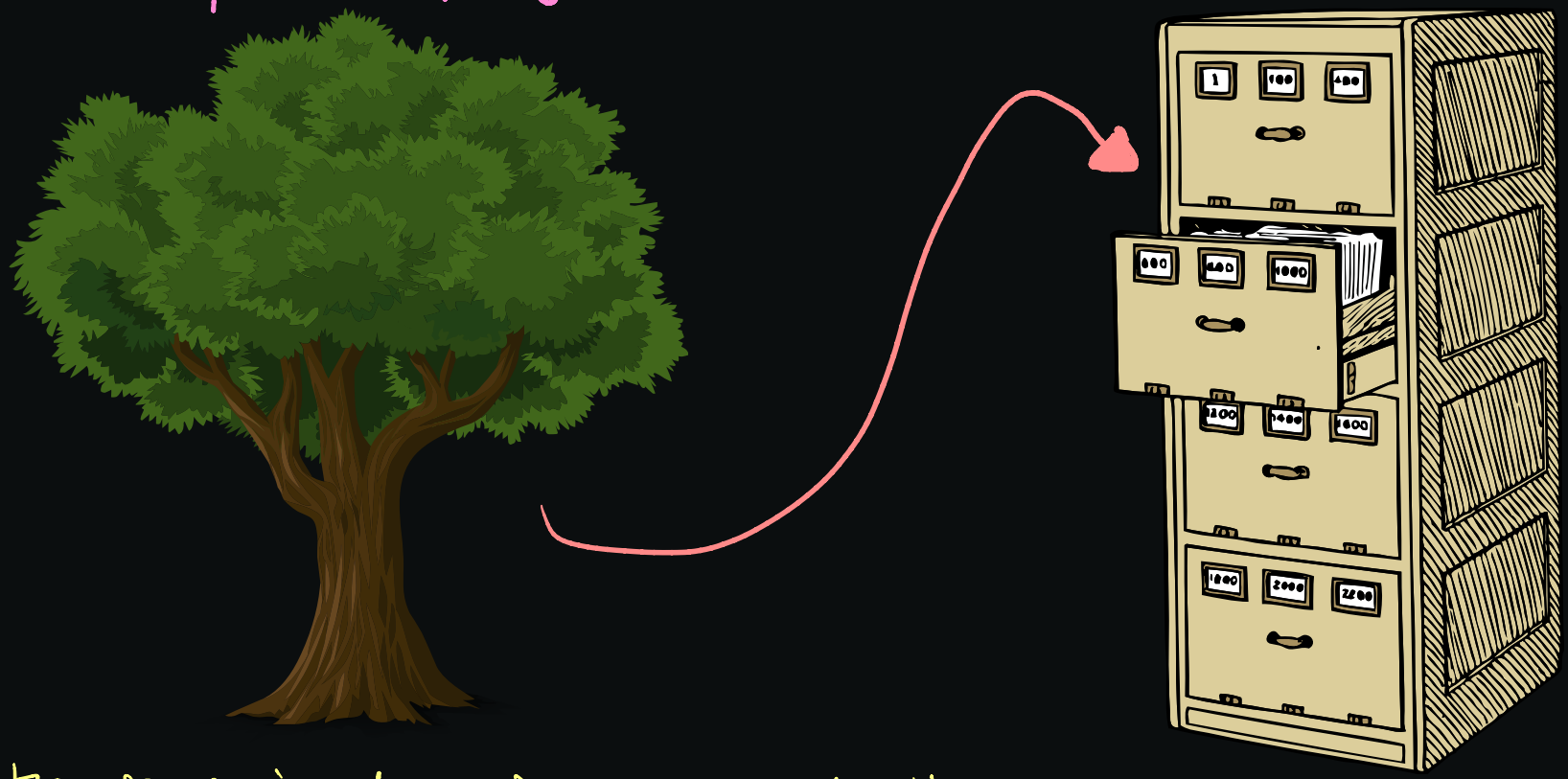
$$x, y \overset{iid}{\sim} p(x, y)$$

data distribution

Data source

Filing cabinet

Imagine a "learning machine" that puts each data sample (x, y) into a shelf of a filing cabinet.



If the same input x is encountered, it can perfectly classify it just by opening the corresponding shelf.

BUT: could it classify an input x it hasn't seen before? No!

The train / test split

Suppose a machine learning system can fit a

training sample $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

that was once upon a time drawn iid from $p(x,y)$  .

Then best practice is to evaluate the system on a

fresh iid test sample $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

to check it didn't just "memorise" the training labels.

In this lecture, the train and test samples will both contain m datapoints.

Test versus population error

The motivation for a test set is the law of large numbers.

For a large test set that is drawn fresh iid, it is highly unlikely for the test error to deviate strongly from the population error.

$\mathbb{I}$ is the indicator fn.

f is our classifier

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left[\operatorname{sign} f(x^{(i)}) \neq y^{(i)}\right]$$

test error

$$P_{x,y\sim p(x,y)}\left[\operatorname{sign} f(x) \neq y\right]$$

population error

Agenda for today

1. What is generalisation?
2. How does VC theory work?
3. Inequalities that we'll need
4. The main result
5. What does VC theory miss?

intuition first

rigour will follow.

Finite function classes

If we pick a classifier f and evaluate it on a large, fresh iid train sample, will we get a good estimate of the population error?
Yes! With high probability, by the law of large numbers.

If we pick 10 classifiers $f_1, f_2, \ldots, f_{10}$ and evaluate all of them on a large, fresh iid train sample, will we get a good estimate of all their population errors?
Yes! With high probability, by the law of large numbers.

Same question, but for 100^{100} classifiers.
Hmm, I'm not sure. LLN will likely fail for one classifier.

Infinite function classes

The previous slide illustrates the basic mechanism of VC-style generalisation theory.

If the amount of train data dwarfs the number of classifiers in the function class, then it is highly unlikely for <u>any</u> classifier to overfit. In this case:

finding the classifier that minimises train error $\approx$ finding the classifier that minimises population error

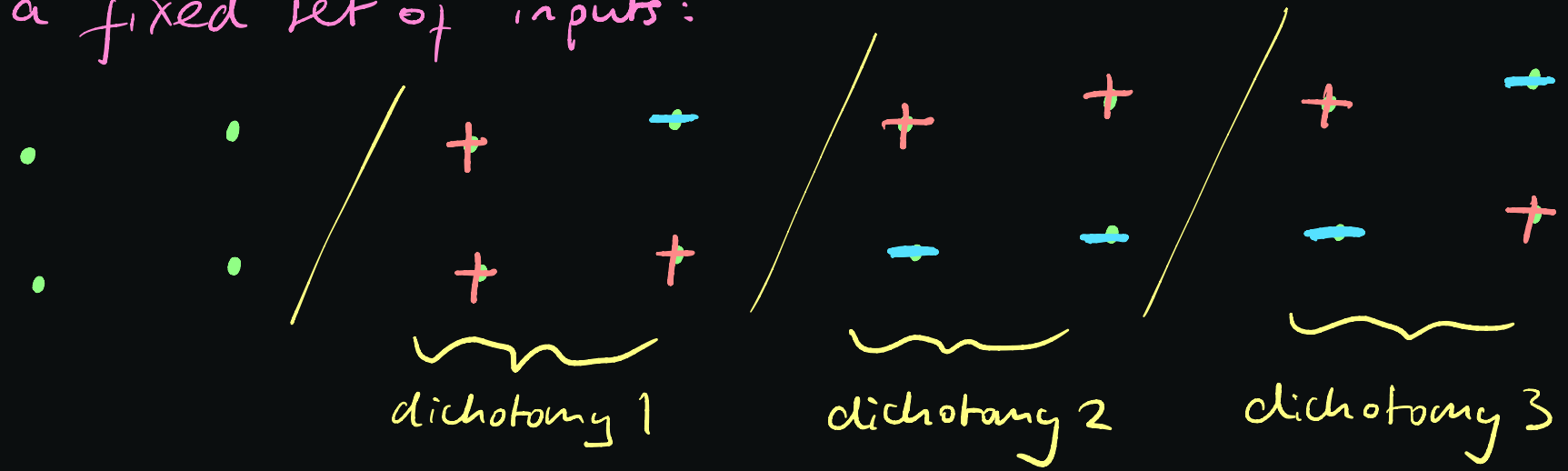
But what if there are an infinite number of classifiers in the function class?

—— this is the last hurdle.

Counting dichotomies

a "dichotomy" is a binary labelling

For a fixed set of inputs:



even if the function class contains infinitely many classifiers, they can only realise a finite number of dichotomies on that set of points.

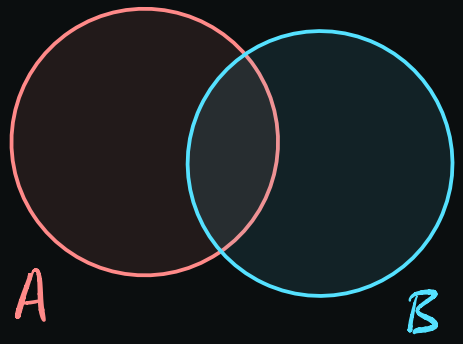
Let $N(F, X)$ denote the number of dichotomies that function class F can realise on a set of inputs X.

VC theory says that $N(F, X)$ is more important than $|F|$.

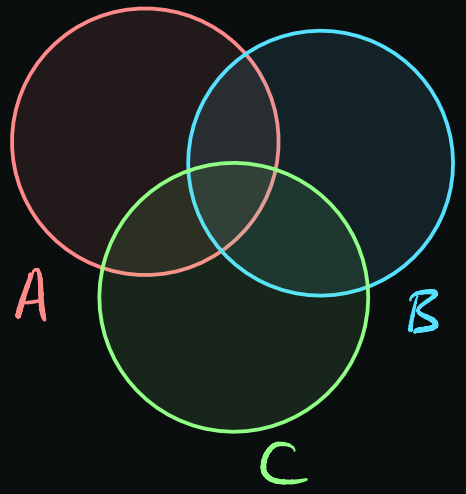
Agenda for today

1. What is generalisation?
2. How does VC theory work?
3. Inequalities that we'll need
4. The main result
5. What does VC theory miss?

The union bound



$$\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$$



$$\mathbb{P}[A \cup B \cup C] \leq \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C]$$

... and so on.

How to think about this: the probability of <u>any</u> of k events with individual probability p occurring $\leq k \times p$.

The Chernoff bound

Let $b_1, b_2, \ldots, b_m \overset{iid}{\sim} \text{Bernoulli}(p)$.

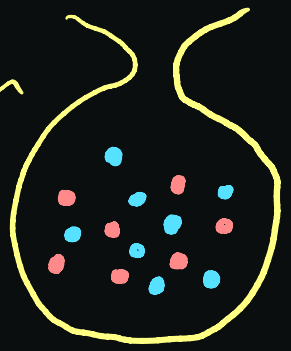
Then $$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} b_i - \mathbb{E}[b_1]\right| \geq \varepsilon\right] \leq 2e^{-2m\varepsilon^2}$$

In words, the probability that the average of a large number of iid Bernoulli trials differs from the true mean is exponentially small.

Chernoff bound++

The same bound holds for sampling without replacement.

Suppose you have a bag containing a fraction p blue balls and $1-p$ red balls.



You draw m balls without replacement. Let $b_i = 1$ indicate that the ith ball was blue and $b_i = 0$ means the ball was red. Then,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} b_i - \mathbb{E}[b_1]\right| \geq \varepsilon\right] \leq 2e^{-2m\varepsilon^2}$$

Agenda for today

1. What is generalisation?
2. How does VC theory work?
3. Inequalities that we'll need
4. The main result
5. What does VC theory miss?

VC theory

THEORY OF PROBABILITY
AND ITS APPLICATIONS

Volume XVI *Number 2*

1971

ON THE UNIFORM CONVERGENCE OF RELATIVE FREQUENCIES OF EVENTS TO THEIR PROBABILITIES

V. N. VAPNIK AND A. YA. CHERVONENKIS

our simplified version.

Statement of the result

Let X denote the combined set of m train inputs and m test inputs. We will prove that:

$$\mathbb{P}\left[\text{for any } f \in F \;\middle|\; \text{train error} - \text{test error}\right| \geq \varepsilon\Big] \leq \frac{2\,\mathbb{E}_X[N(F, X)]}{\exp\left(\frac{m\varepsilon^2}{2}\right)}$$

Looking at the RHS, the numerator measures the number of dichotomies that the function class F can realise, averaged over all possible data samples X.

For large enough m, all classifiers $f \in F$ will attain similar train and test errors with high probability.

Main idea of the proof

Break up the data sampling into two stages:

1. Draw $2m$ datapoints iid. — refer to them as X.
2. Randomly choose m to be the "train set".
   The other m will be the "test set".

Think of X as a bag of $2m$ datapoints: $X =$ 

Once we have X fixed, we can consider the dichotomies that F realises on X.

e.g. 

For a particular dichotomy, we draw m points from the bag without replacement, and compare the error on those points to the error on all of X.

In symbols

$$\mathbb{P}\left[\begin{array}{l}\text{for any} \\ f \in F\end{array} \;\middle|\; \begin{array}{l}\text{train} \\ \text{error}\end{array} - \begin{array}{l}\text{test} \\ \text{error}\end{array} \right| \geq \varepsilon \Big]$$

$$= \int d\mathbb{P}(X)\, \mathbb{P}\left[\begin{array}{l}\text{for any} \\ f \in F\end{array} \;\middle|\; \begin{array}{l}\text{train} \\ \text{error}\end{array} - \begin{array}{l}\text{error} \\ \text{on } X\end{array} \right| \geq \frac{\varepsilon}{2} \,\Big|\, X \Big]$$

$$\overset{\text{(union bound)}}{\leq} \int d\mathbb{P}(X) \sum_{\substack{\text{dichotomies of } X \\ \text{realisable by } F}} \mathbb{P}\left[\left| \begin{array}{l}\text{train} \\ \text{error}\end{array} - \begin{array}{l}\text{error} \\ \text{on } X\end{array} \right| \geq \frac{\varepsilon}{2} \,\middle|\, X, \text{dichotomy} \right]$$

$$\overset{\text{(C++)}}{\leq} \int d\mathbb{P}(X) \sum_{\substack{\text{dichotomies of } X \\ \text{realisable by } F}} 2 e^{-m\varepsilon^2/2}$$

$$= 2\, \mathbb{E}_X[N(F, X)]\, e^{-m\varepsilon^2/2}$$

Rearranging the result

to a more familiar form.

We have:

$$\mathbb{P}\left[\begin{matrix}\text{for any}\\ f\in F\end{matrix} \left| \begin{matrix}\text{train}\\ \text{error}\end{matrix} - \begin{matrix}\text{test}\\ \text{error}\end{matrix} \right| \geq \varepsilon\right] \leq \frac{2\,\mathbb{E}_X[N(F,X)]}{\exp\left(m\frac{\varepsilon^2}{2}\right)}$$

This may be rearranged to:

$$\begin{matrix}\text{for all}\\ f\in F\end{matrix} \left| \begin{matrix}\text{train}\\ \text{error}\end{matrix} - \begin{matrix}\text{test}\\ \text{error}\end{matrix} \right| < \sqrt{\frac{2}{m}\left[\ln \mathbb{E}_X N(F,X) + \ln\frac{2}{\delta}\right]}$$

with probability at least $1-\delta$.

VC dimension

A core quantity in the bounds was $\ln \mathbb{E}_X N(\mathcal{F}, X)$ — log of the average number of dichotomies, a.k.a. the "annealed entropy".

It's possible to upper bound the annealed entropy in terms of a new quantity, the "VC dimension":

$\mathrm{VC}(\mathcal{F})$ = maximum m for which there exists an X such that $\mathcal{F}$ realises all dichotomies

Thus the bounds can be re-expressed in terms of the VC dimension of $\mathcal{F}$.

Agenda for today

1. What is generalisation?
2. How does VC theory work?
3. Inequalities that we'll need
4. The main result
5. What does VC theory miss?

All dichotomies

What's the total number of dichotomies on m points?

Answer: 2^m, since there are 2 options for each point.

What happens to the VC-bound if the function class F can realise all 2^m dichotomies?

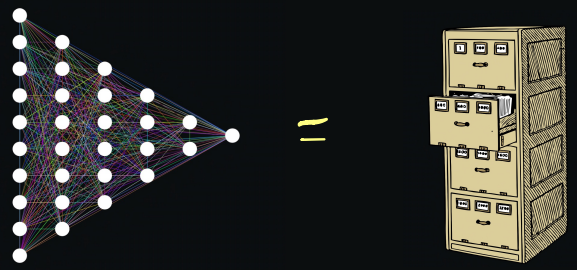
$$\mathbb{P}\left[\text{for any } f \in F \;\middle|\; \text{train error} - \text{test error}\right| \geq \varepsilon\right] \leq \frac{2\,\mathbb{E}_X[N(F,X)]}{\exp\left(m\frac{\varepsilon^2}{2}\right)} = \frac{2 \times 2^m}{\exp\left(m\frac{\varepsilon^2}{2}\right)}$$

For $0 \leq \varepsilon \leq 1$, the RHS grows rather than shrinks as $m \to \infty$

$\hookrightarrow$ the bound is vacuous, in this setting.

Neural networks

NNs can often realise all 2^m dichotomies of m training points, so from the point of view of VC theory,



yet when trained on the actual dichotomy...

... NNs still generalise.

Summary

VC theory seeks conditions for when no classifier in the function class F overfits the train set.

This happens when $\#\text{datapoints} \gtrsim \log(\#\text{dichotomies})$.

But for a neural net, $\#\text{dichotomies}$ is <u>huge</u>.

And do we really require that <u>no</u> classifier overfits?

Next lecture

PAC – Bayesian theory offers a solution to the problems that VC theory faces.