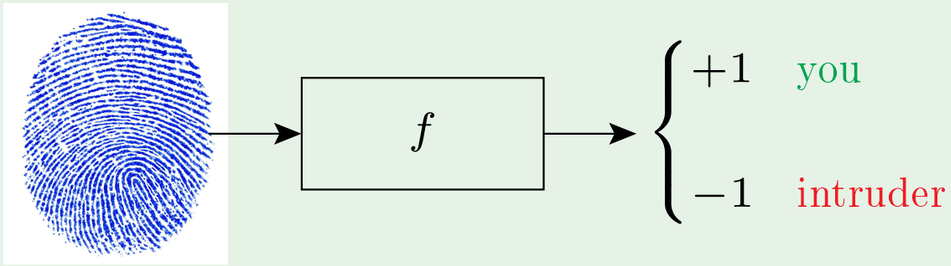
Review of Lecture 4

- Error measures
  - User-specified $\mathrm{e}\left(h(\mathbf{x}), f(\mathbf{x})\right)$



  - In-sample:

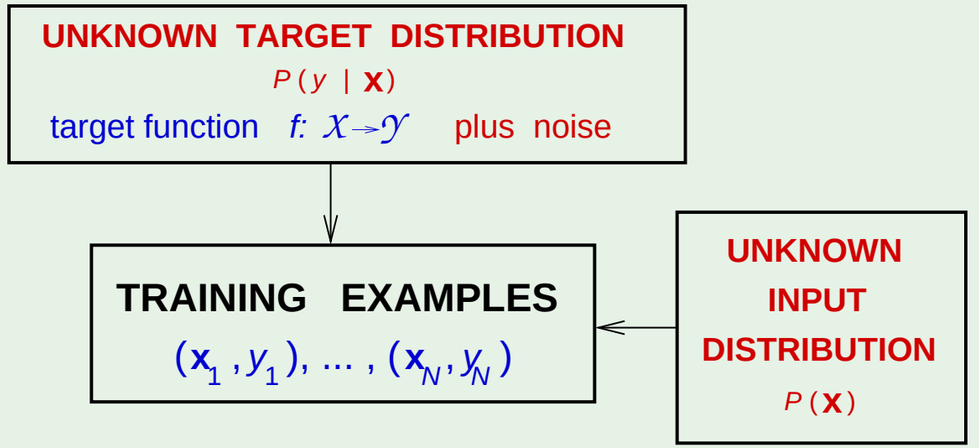
$$E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{e}\left(h(\mathbf{x}_n), f(\mathbf{x}_n)\right)$$

  - Out-of-sample

$$E_{\text{out}}(h) = \mathbb{E}_{\mathbf{x}}\left[\mathrm{e}\left(h(\mathbf{x}), f(\mathbf{x})\right)\right]$$

- Noisy targets

$$y = f(\mathbf{x}) \quad \longrightarrow \quad y \sim P(y \mid \mathbf{x})$$



  - $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$ generated by each independently

$$P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$$

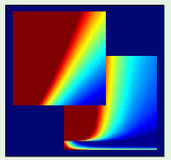
  - $E_{\text{out}}(h)$ is now $\mathbb{E}_{\mathbf{x},y}\left[\mathrm{e}\left(h(\mathbf{x}), y\right)\right]$

Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 5: **Training versus Testing**

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Tuesday, April 17, 2012



Outline

- From training to testing
- Illustrative examples
- Key notion: break point
- Puzzle

The final exam

Testing:

fixed hypothesis, hence M=1

$$\mathbb{P}\left[\left|E_{\text{in}} - E_{\text{out}}\right| > \epsilon\right] \leq \quad 2 \quad e^{-2\epsilon^2 N}$$

Training:

trialling hypotheses

$$\mathbb{P}\left[\left|E_{\text{in}} - E_{\text{out}}\right| > \epsilon\right] \leq 2M\, e^{-2\epsilon^2 N}$$

Where did the M come from?

The $\mathcal{B}$ad events $\mathcal{B}_m$ are

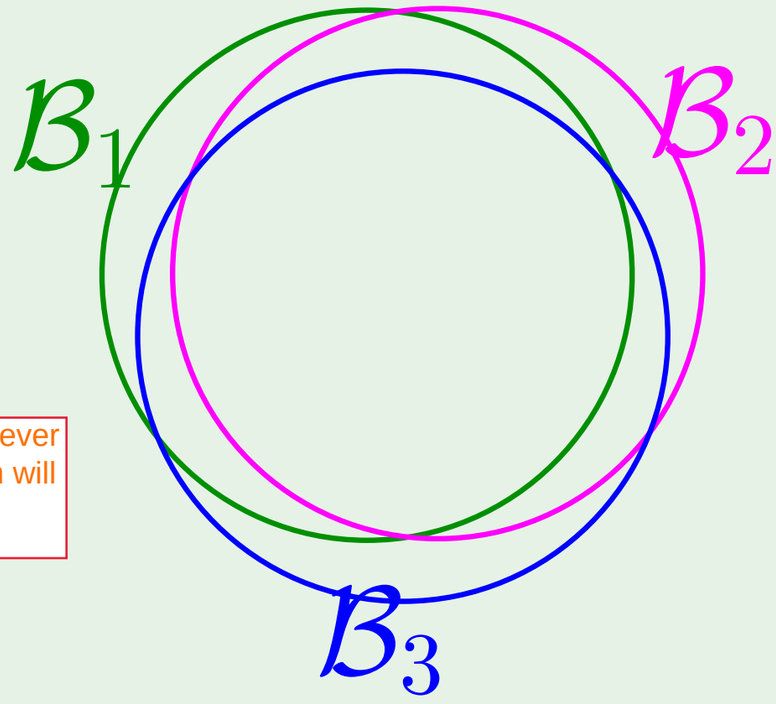
$$\text{“}|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\text{”}$$

The union bound:

$$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \mathcal{B}_2 \textbf{ or } \cdots \textbf{ or } \mathcal{B}_M]$$

We want P(B1 or B2 or ...) to be small so that whatever final hypothesis the learning algorithm picks, its Ein will track Eout well (which is something we require).

$$\leq \underbrace{\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots + \mathbb{P}[\mathcal{B}_M]}_{\text{no overlaps: } M \text{ terms}}$$



This not a very precise bound, since it assumes the 'worse case' or case of highest area where there is no overlap between the B's (they are disjoint) - we want to take into consideration the overlaps, since for two similar hypotheses h1 and h2, the events | Ein(h1)-Eout(h1) | >epsilon and | Ein(h2)-Eout(h2) | >epsilon are likely to coincide for most datasets

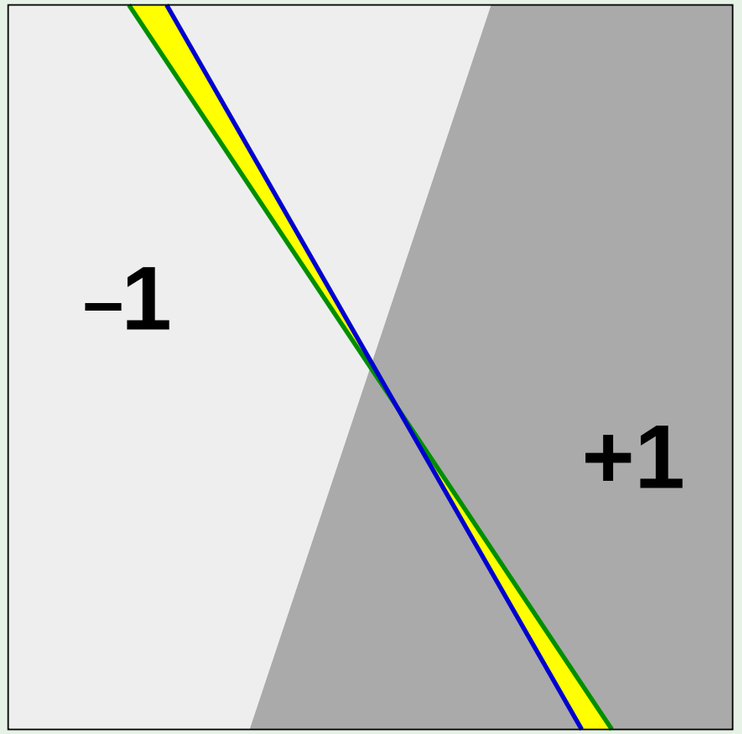
Can we improve on M?

We want to extract a quantity from the hypothesis set to characterize the overlap and get us a good bound without having to go through the details of how the events are correlated.

Yes, bad events are *very* overlapping!

ΔE_{out}: change in $+1$ and -1 areas

ΔE_{in}: change in labels of data points



$$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| \approx |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)|$$

We are interested in showing Ein tracks Eout for h1 is approximately equal to how they track for h2, or that if the LHS exceeds epsilon then RHS also exceeds epsilon most of the time - so there is a lot of overlap in the bad events B.

What can we replace M with?

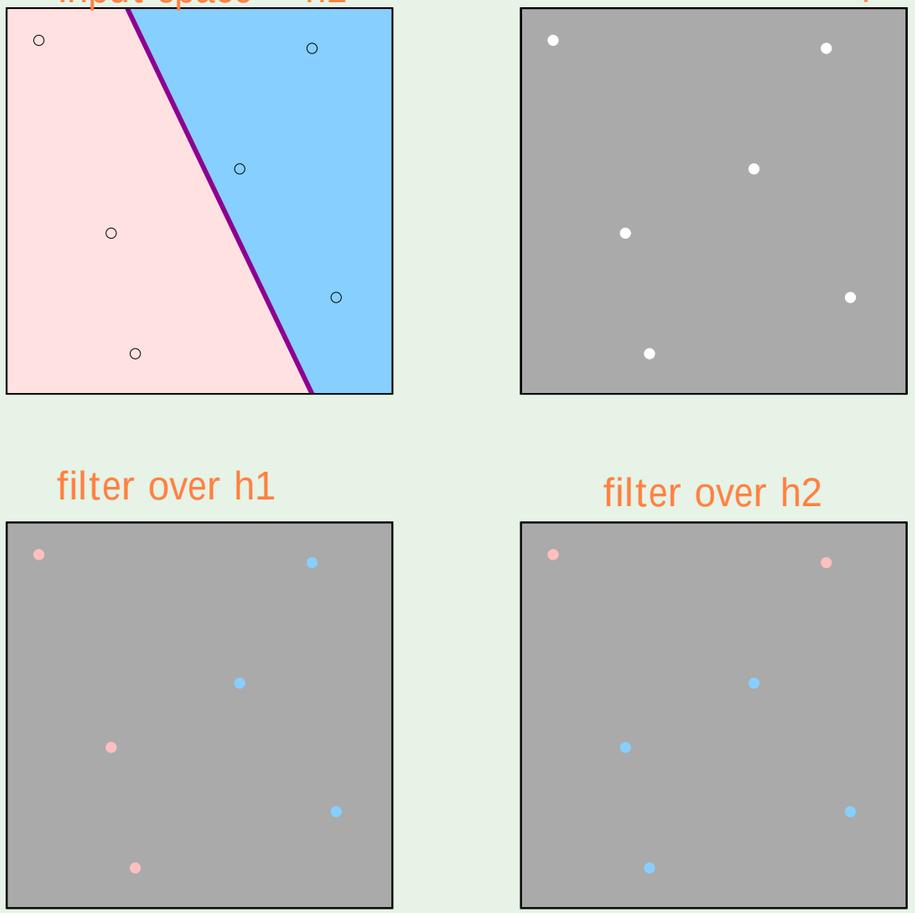
Instead of the whole input space,

we consider a finite set of input points,

and count the number of *dichotomies*

i.e. how many patterns of red/blue points can we get - it characterizes the strength/power of thr hypothesis set since if a hyp set can give all the combinations of red/blue points, it is a powerful set (and vice versa if it can only model a few combinations) - we tried to characterise the power of a hypothesis set with M.

How we can consider dichotomies



We reduce the problem to looking at it through a sheet with N holes over the data sample (so not the whole input space). So can change the hypothesis without the boundary crossing a point and the filtered picture doesnt cahnge. When we cross a point (h1->h2), we have another pattern.

Dichotomies: mini-hypotheses

A hypothesis $h : \mathcal{X} \to \{-1, +1\}$

A dichotomy $h : \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \to \{-1, +1\}$

Number of hypotheses $|\mathcal{H}|$ can be infinite

Number of dichotomies $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)|$ is at most 2^N

Candidate for replacing M

- where we apply each member h from the set H to all of the data points {x1, x2, ..., xN) and create a dichotomy: a vector of length N of +/- 1. Many of the h applied to {xi} will return the exact same dichotomy since it is quite restricted and only returns +/- 1 on each point, hence not infinite. A higher cardinality of a set of dichotomies suggests H is more 'diverse'.

The growth function

The growth function counts the most dichotomies on any N points

cardinality of set of dichotomies

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

max w.r.t. N points from X

The growth function satisfies:

$$m_{\mathcal{H}}(N) \leq 2^N$$

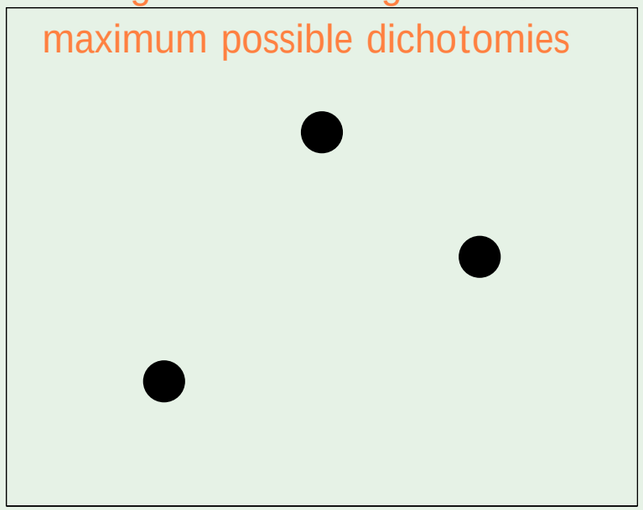
Let's apply the definition.

We may have an X and H where we can generate any pattern we want, so m can = 2^N. However in most cases, using h from H means we can miss out on some patterns, so < 2^N

- for N points, the growth function returns the most dichotomies possible (not N points from our data set, but it picks from the input space to give the max number of dichotomies). So we consider all the choices of N points from X with a view to maximizing the dichotomies such that the number we get is more than any number someones else can get with N points. It finds the most possible dichotomies - e.g. if all the points were on a line (colinear), it would have fewer dichotomies as fewer ways to separate, so this would not be the maximum value. m returns the most expressive facet of H on the N points.

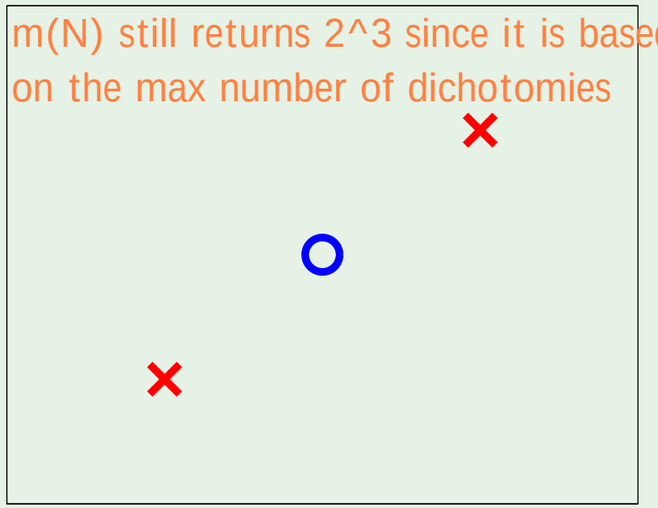
Applying $m_{\mathcal{H}}(N)$ definition - perceptrons

arrangement which gives the maximum possible dichotomies



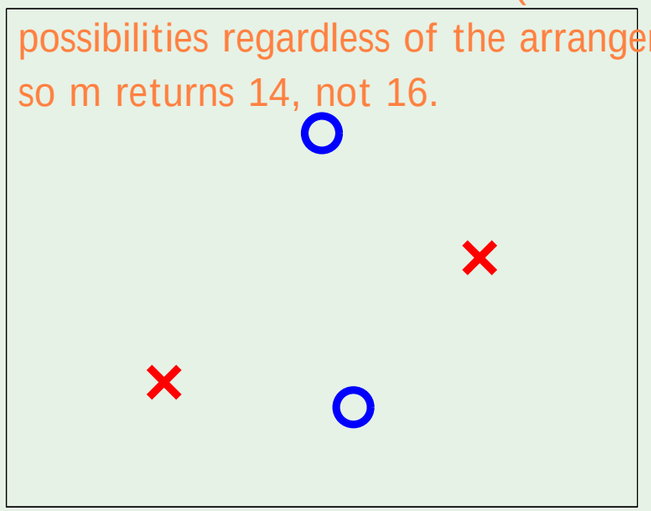
$N = 3$

due to collinearity, arrangement does not return 8 dichotomies but m(N) still returns 2^3 since it is based on the max number of dichotomies



$N = 3$

Here with N=4 we miss two (or more) of the possibilities regardless of the arrangement, so m returns 14, not 16.



$N = 4$

$$m_{\mathcal{H}}(3) = 8$$

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

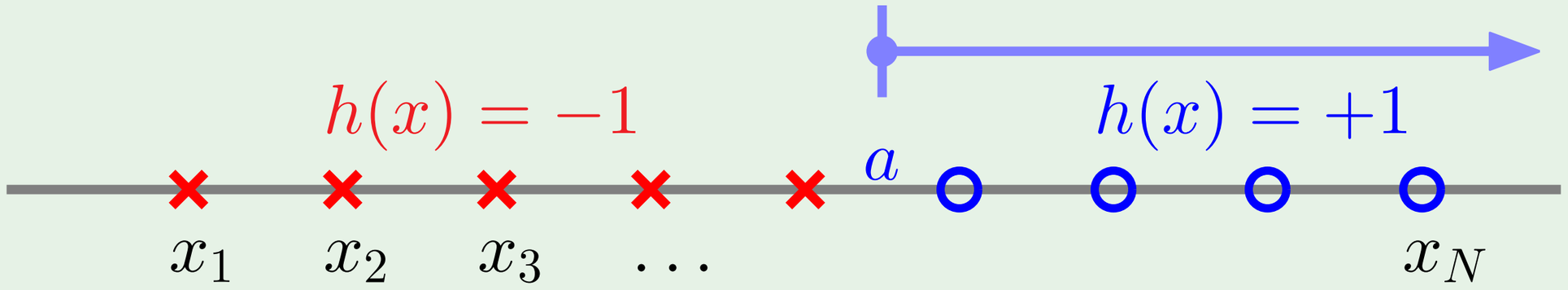
So for N=4, perceptrons start to become limited and cannot create two of the dichotomies

Outline

- From training to testing
- Illustrative examples
- Key notion: break point
- Puzzle

Let us now illustrate how to compute mH(N) for some simple hypothesis sets. These examples will confirm the intuition that mH(N) grows faster when the hypothesis set H becomes more complex. This is what we expect of a quantity that is meant to replace M in the generalization bound.

Example 1: positive rays



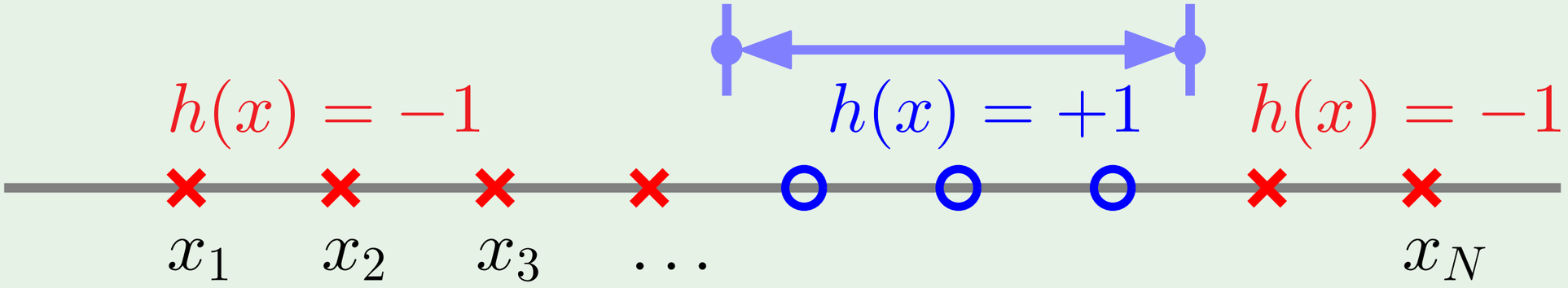
Must count number of possible choices of 'a' which give us different dichotomies, so since there are N-1 sandwiched line segments, +1 if a is to left of x1, +1 if a is to right of xN, so N+1 overall

$\mathcal{H}$ is set of $h : \mathbb{R} \to \{-1, +1\}$

$$h(x) = \text{sign}(x - a)$$

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

Example 2: positive intervals



$\mathcal{H}$ is set of $h : \mathbb{R} \to \{-1, +1\}$

Place interval ends in two of $N + 1$ spots

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

+1 to account for all red - when the two intervals are at the same point

Example 3: convex sets

h(x) = +1 if, at x, it is a convex region. A convex region is a region where, if you pick any two points within the region, the entity of the line segment connecting them lies within the region. N.B. a set is convex if the line segment connecting any two points in the set lies entirely within the set

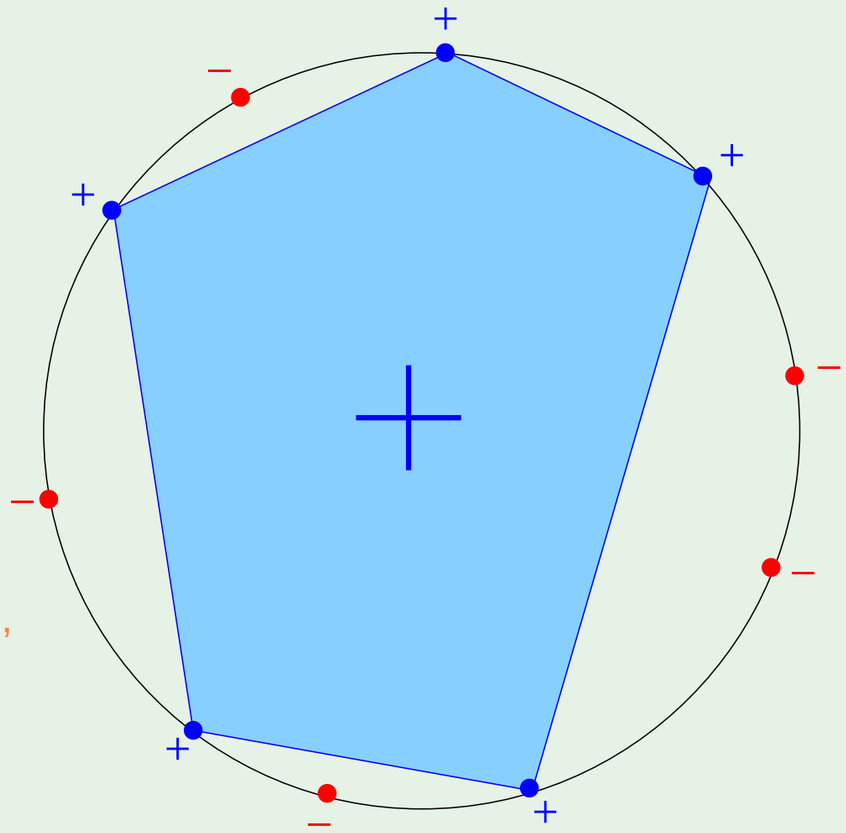
$\mathcal{H}$ is set of $h: \mathbb{R}^2 \to \{-1, +1\}$

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

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

The N points are 'shattered' by convex sets

occurs when H is capable of generating all possible dichotomies on {x1, ..., xN}, so H(x1, ..., xN) = {-1, +1}^N with cardinality 2^N. This signifies that H is as diverse as can be on this particular sample.



Putting the points on a circle gives the maximum possible dichotomies using convex regions since all the points that are convex (and therefore need to connect via a line segment) can connect if on the perimeter of a circle.

The 3 growth functions

- $\mathcal{H}$ is positive rays:

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

- $\mathcal{H}$ is positive intervals:

$$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

quadratic, so m grows faster than the linear m in the 'simpler' positive ray hypothesis set

- $\mathcal{H}$ is convex sets:

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

Back to the big picture

Remember this inequality?

$$\mathbb{P}\left[\left|E_{\text{in}} - E_{\text{out}}\right| > \epsilon\right] \leq 2M\, e^{-2\epsilon^2 N}$$

What happens if $m_{\mathcal{H}}(N)$ replaces M?

$m_{\mathcal{H}}(N)$ polynomial $\implies$ Good!

This is because polynomial learning is feasible given that, due to the exponential factor (with epsilon generally very small), the exponential will eventually (depending on the patience of the customer or size of the dataset) kill any polynomial it is multiplied by. Note that m will always either be polynomial or 2^N.

Just prove that $m_{\mathcal{H}}(N)$ is polynomial?

Outline

- From training to testing
- Illustrative examples
- Key notion: **break point**
- Puzzle

Break point of $\mathcal{H}$

Definition:

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

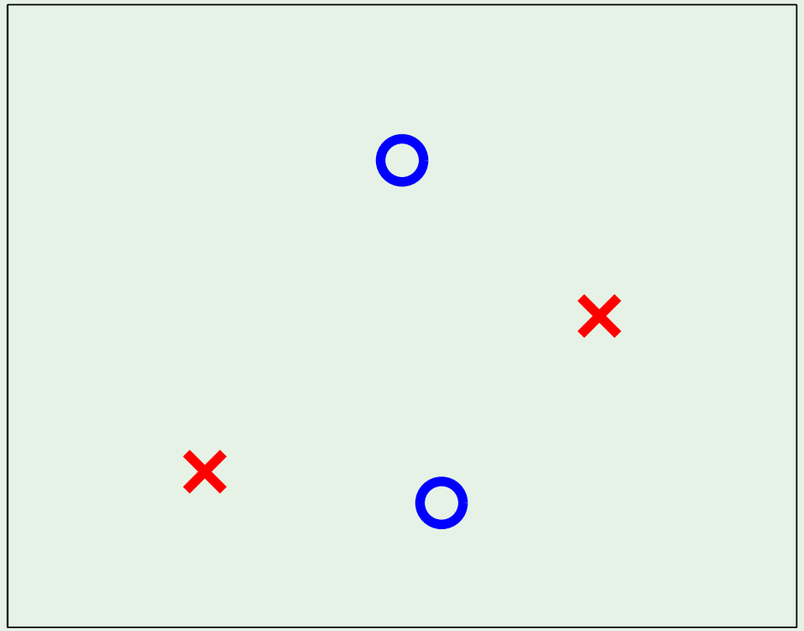
$$m_{\mathcal{H}}(k) < 2^k$$

In general it is easier to find a break point for H than to compute the full growth function for that H.

For 2D perceptrons, $k = 4$

N.B if k is a break point, all integers > k are also break points.

A bigger data set cannot be shattered either



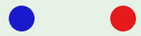
It is the data size at which there is no arrangement of points such that the hypothesis set can produce all possible dichotomies - e.g. k=4 for 2D perceptrons because at N=4, we could only get 14, not 2^4=16 dichotomies - see above diagram.

Break point - the 3 examples

i.e. for what N does m, the growth function, stop returning 2^N?

- 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 $k =$ '∞'

Main result

No break point $\implies$ $m_{\mathcal{H}}(N) = 2^N$

Any break point $\implies$ $m_{\mathcal{H}}(N)$ is **polynomial** in N

As a result of the above, in principle, the existence of a break point indicates that learning is feasible with the hypothesis set. Meanwhile the value of the break point tells us the resources needed (size of training data) for a certain performance.

We will now use the break point k to derive a significant bound on the growth function mH(N) for all values of N. It is not practical to try and compute the growth function for every H. So fortunately, as we are attempting to replace M with mH(N), we can use an upper bound on mH(N) instead of the exact value, and the Hoeffding Inequality will still hold.

Puzzle