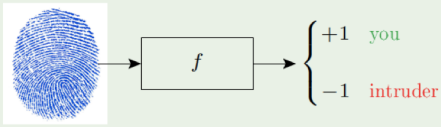
CSE4088 Introduction to Machine Learning

Training versus Testing

Slides are adopted from lecture notes of Yaser Abu-Mostafa

Review of last lecture

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



  - In-sample:

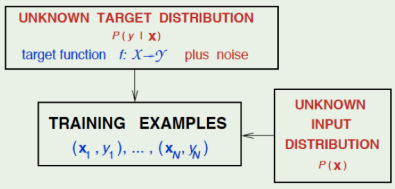
$$E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N} \text{e}\left(h(\mathbf{x}_n), f(\mathbf{x}_n)\right)$$

  - Out-of-sample

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

- 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

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

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

Outline

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

The final exam

Testing:

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

Training:

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

Where did *M* come from?

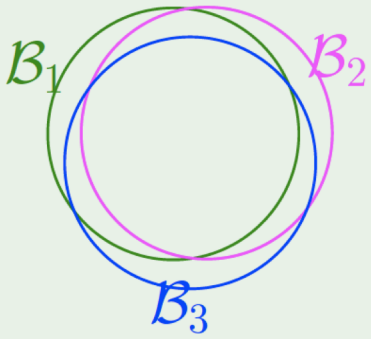
The $\mathcal{B}$ad events $\mathcal{B}_m$ are

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

The union bound:

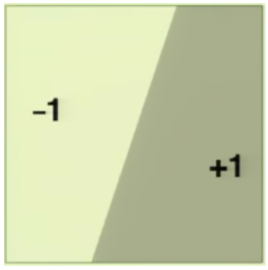
$$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \mathcal{B}_2 \textbf{ or } \cdots \textbf{ or } \mathcal{B}_M]$$

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

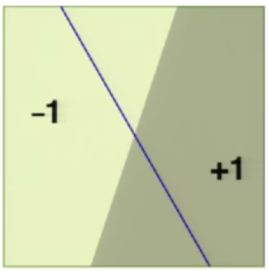


Can we improve on *M*?

- Yes, bad events are *very* overlapping!

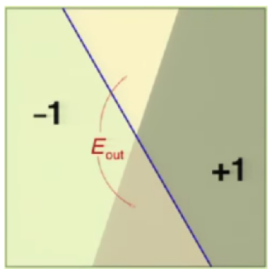


The target hypothesis for a perceptron.



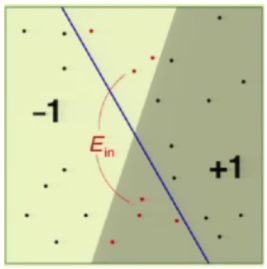
A hypothesis for a perceptron. Badly performing hypothesis

E_{out}:



E_{out} is the sum of the two areas

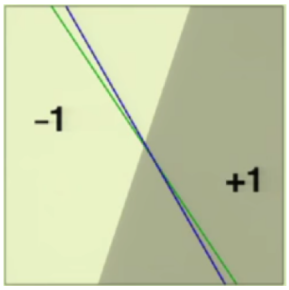
E_{in}:



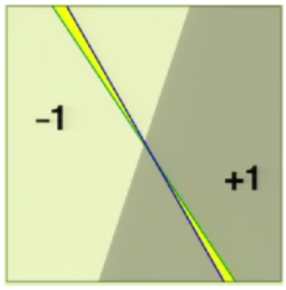
For E_{in} we need a sample. Some of them will fall into the bad region.

Can we improve on *M*?

- Yes, bad events are *very* overlapping!
- What is the change in E_{in} and E_{out} when you change the hypothesis?



Take another perceptron (hypothesis). They are very Close.



- Change in E_{out} is the yellow area. A very thin area.
- Change in E_{in} is small since the probability of a data point falling in that region is small.

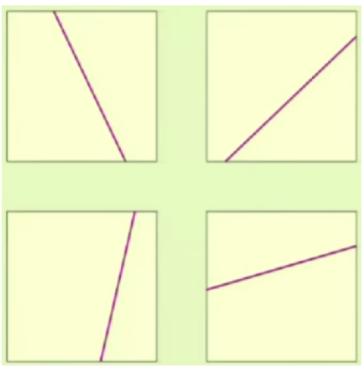
ΔE_{out}: change in $+1$ and -1 areas

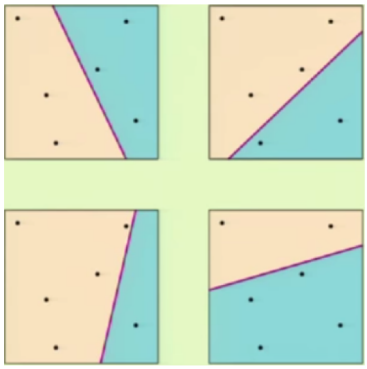
ΔE_{in}: change in labels of data points

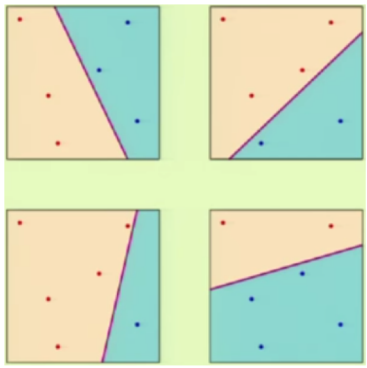
$$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| \approx |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)|$$

What can we replace *M* with?

- We will define the quantity that will replace M. Proof will come later.



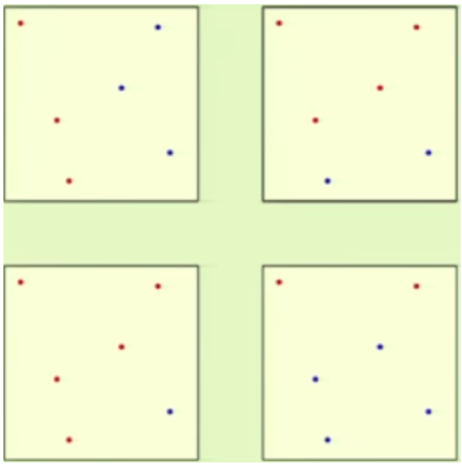




- Four different perceptrons (hypotheses).
- When we count the number of hypothesis we take into consideration the entire input space.
- Infinite number of hypothesis.

Instead of counting the number of hypothesis on the entire space we will restrict our attention **only** to the sample consisting of finite number of points.

What can we replace *M* with?



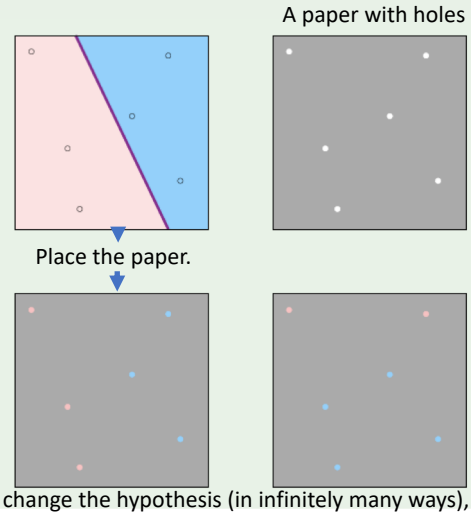
- Perceptron is somewhere and splitting the points.
- For these constellation of points, how many patterns of red and blue can I get? How many dichotomies can we get given N points?
- We are not counting all the hypothesis, but we are still counting on a restricted set.
  - A hypothesis that will give us all combinations of red and blue is a powerful hypothesis.
  - A hypothesis which will give us only a few combinations is not a powerful hypothesis.

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*



As you change the hypothesis (in infinitely many ways), you will not notice it until it changes the color of one of the points.

Dichotomies: mini-hypothesis

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

The growth function

The growth function counts the most dichotomies on any N points

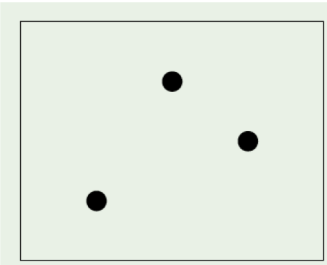
$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

The growth function satisfies:

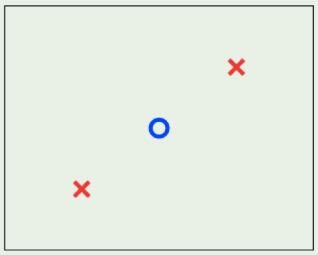
$$m_{\mathcal{H}}(N) \leq 2^N$$

Let's apply the definition.

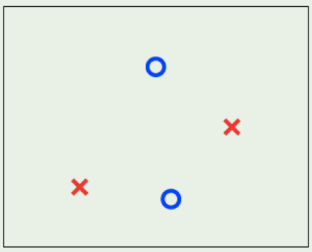
Applying $m_{\mathcal{H}}(N)$ definition - perceptrons



$N = 3$



$N = 3$



$N = 4$

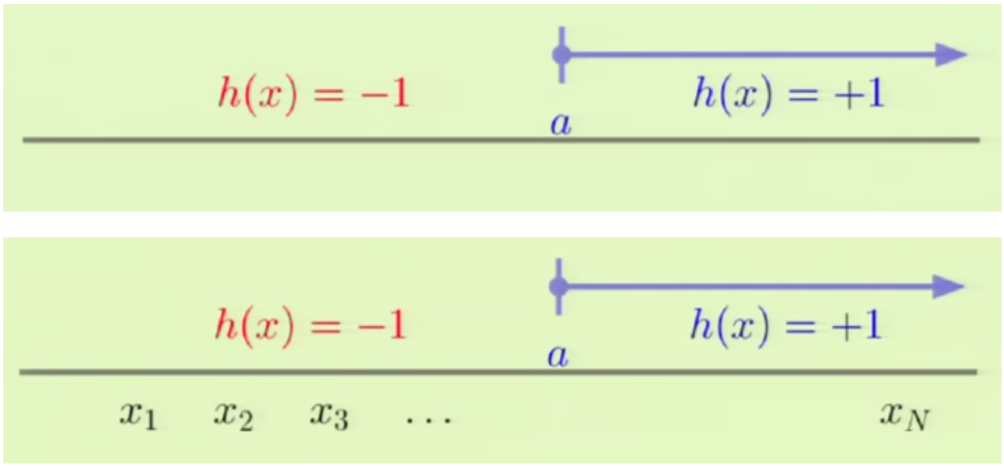
$m_{\mathcal{H}}(3) = 8$

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

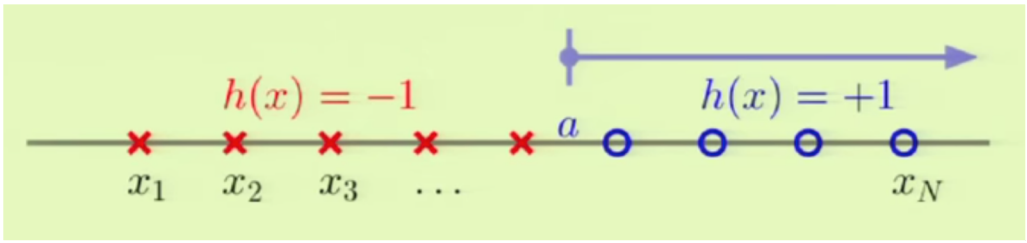
Outline

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

Example 1: positive rays



Example 1: positive rays

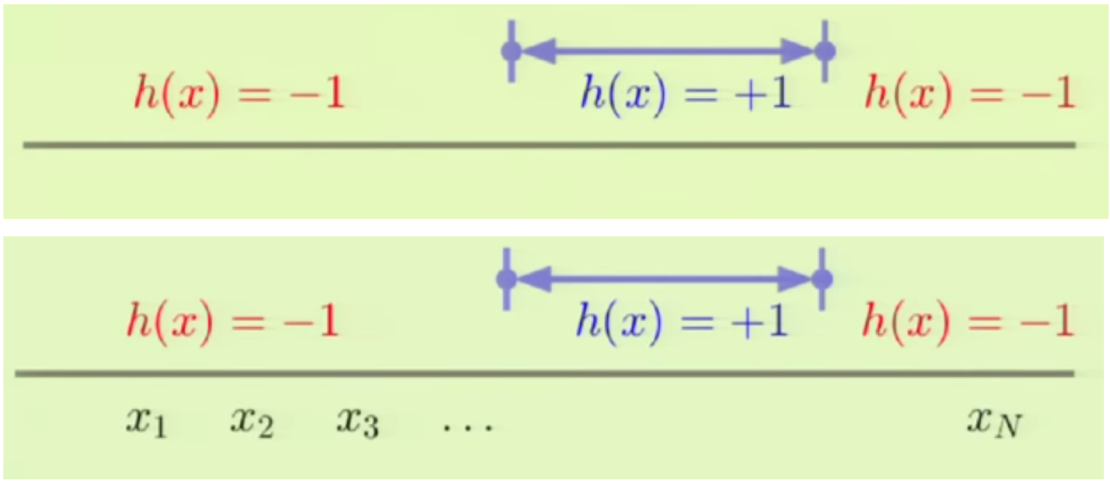


$\mathcal{H}$ is set of $h: \mathbb{R} \to \{-1, +1\}$

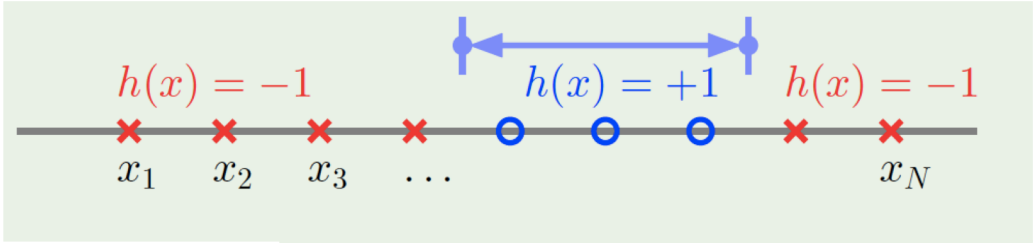
$h(x) = \text{sign}(x - a)$

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

Example 2: positive intervals



Example 2: positive intervals



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

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

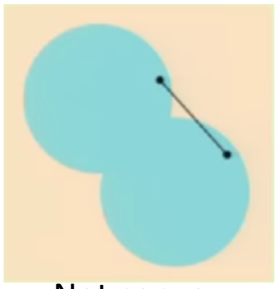
$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

Example 3: convex sets

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

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

convex



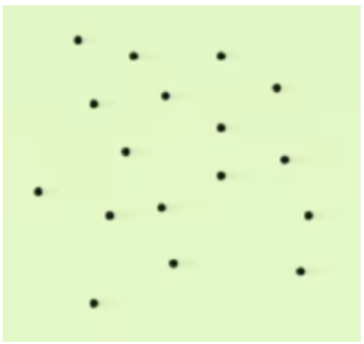
Not convex

Example 3: convex sets

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

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

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

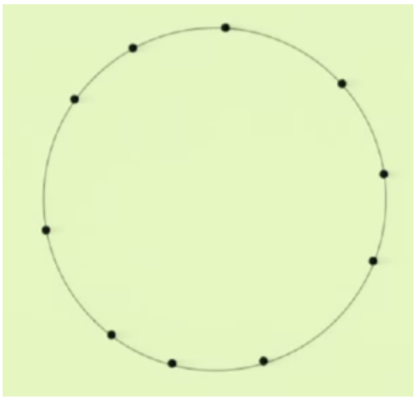


If we assign all the outermost points to +1, this will force all the inner points to be +1 as well since I need a convex region.

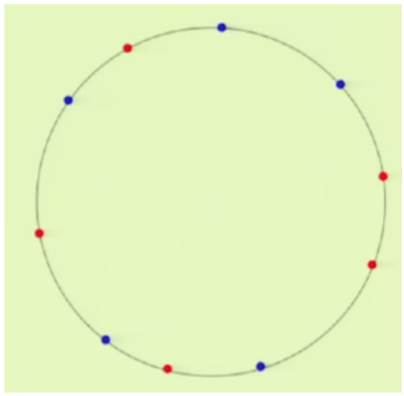
So I have to exclude many dichotomies. → counting is difficult

Can we position the points differently that will give all the dichotomies possible?

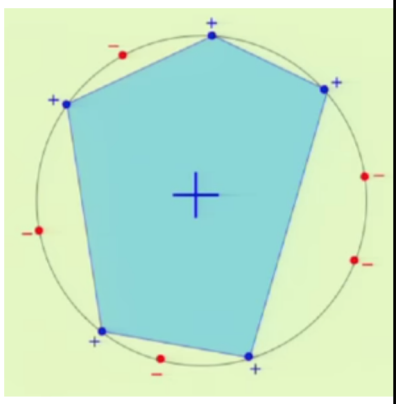
Example 3: convex sets



Take a circle and put your Points on the perimeter of the circle.



Can we realize this dichotomy using a convex region?



Connect the blue points with lines. This will give a convex region.

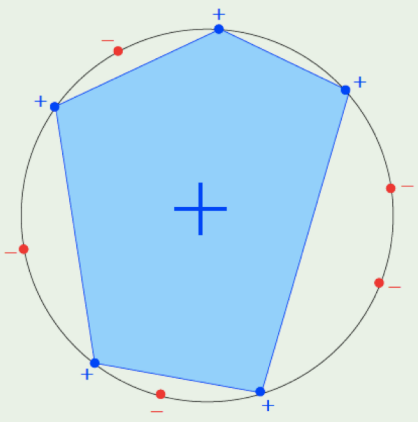
Example 3: convex sets

$\mathcal{H}$ 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



The 3 growth functions

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

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

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

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

- $\mathcal{H}$ 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!

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

Outline

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

Break point of $\mathcal{H}$

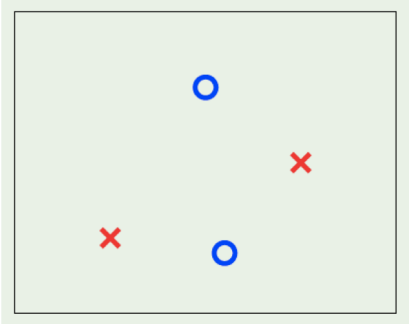
Definition:

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

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

For 2D perceptrons, $k = 4$

A bigger data set cannot be shattered either



Break point – the 3 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 $k =$ '∞'

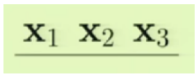
Main result

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

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

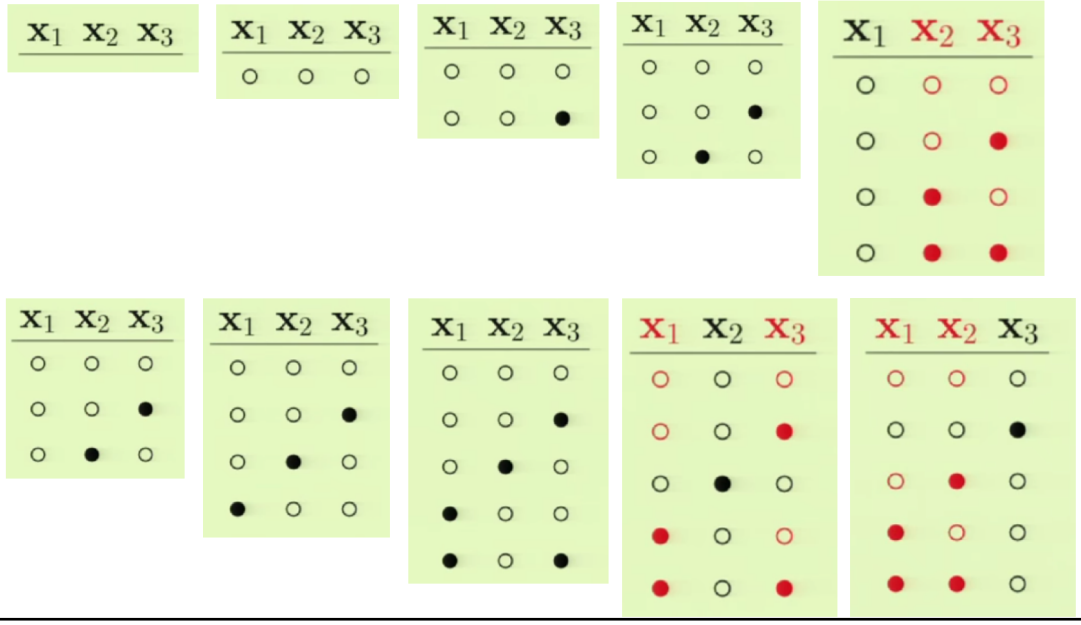
If **there is a break** point we can generalize and learning is feasible.

Puzzle



- We have three points.
- **Restriction**: The breakpoint is **two** for this hypothesis set. You can not get all possible 4 dichotomies on any two data points.
- How many dichotomies can you get on three points under this restriction?

Puzzle



Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
●	○	○

We lost half of the dichotomies