FSAN/ELEG815: Statistical Learning

Gonzalo R. Arce

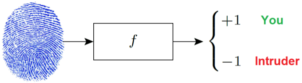
Department of Electrical and Computer Engineering
University of Delaware

4: Training vs Testing

Review

- **Error measures:**

User specified $\mathrm{e}(h(\mathbf{x}), f(\mathbf{x}))$



In-sample:

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

Out-of-sample:

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

Outline

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

Example - The Final Exam

Before the final exam, a professor may hand out practice problems and solutions to the class (training set).

Why not to give out the exam problems?

The goal is for the students to learn the course material (small E_{out}), not to memorize the practice problems (small E_{in}).

Having memorized all the practice problems (small E_{in}) does not guarantee to learn the course material (small E_{out}).

The Final Exam

Testing:

- The hypothesis is fixed (you already prepared for the test).
- The hypothesis is tested over unseen data (the test does not include the same practice problems) i.e. E_{in} is computed using the hypothesis set.

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

- For a large N (number of questions), E_{in} tracks E_{out} (your performance gauges how well you learned).

The Final Exam

Training: Performance on practice problems.

- The hypothesis is adjusted (since you know the answers, you repeat a problem until getting it right).

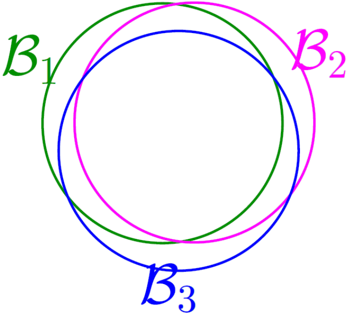
$$\mathbb{P}[|E_{in} - E_{out}| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

- E_{in} is computed using the practice set.
- Small $E_{in} \rightarrow$ not necessarily small E_{out}.
  You may have not learned and have memorized the problems solutions.
- M is the number of hypotheses to explore.
  Depending on the times you repeat a problem, your performance may no longer accurately gauge how well you learned.

Goal: We want to replace M by another quantity that is not infinity.

Where did the M Come from?

Venn Diagram of $\mathcal{B}$ad events



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

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

The union bound consider $\mathcal{B}_m$ as disjoint events:

$$\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \cdots \text{ or } \mathcal{B}_M] \leq \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots \mathbb{P}[\mathcal{B}_M]$$

It is a poor bound when there is overlap.

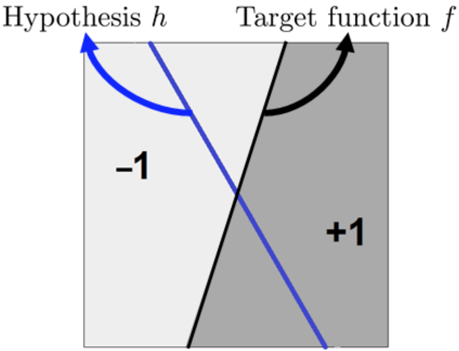
Can we Improve on M ?

Yes, bad events are very overlapping

Remember the perceptron:

$$h(\mathbf{x}) = \begin{cases} 1 & \text{if 'approved'} \\ -1 & \text{if 'deny credit'} \end{cases}$$

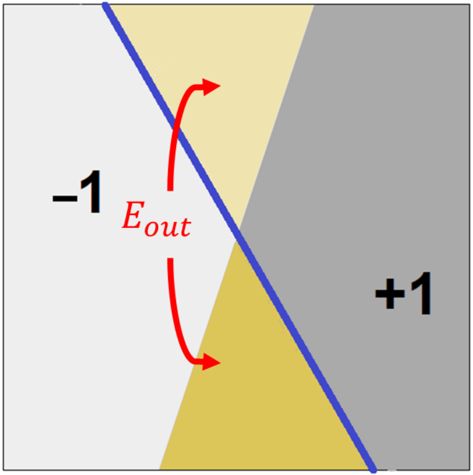
$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x})$$

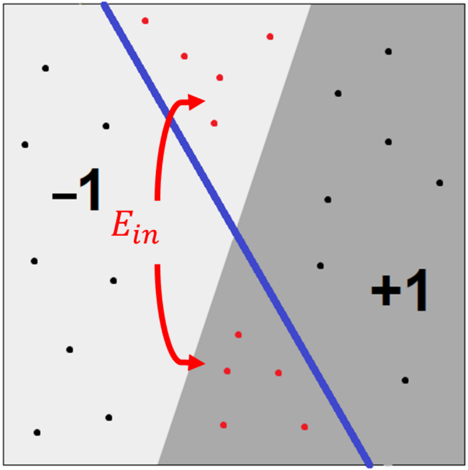


For any perceptron ($\mathbf{w}$): The line $w_0 + w_1x_1 + w_2x_2 = 0$ splits the plane into $+1$ and -1

Can we Improve on M ?

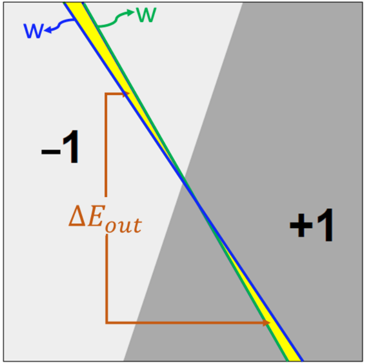
For the given perceptron (**w**) , consider the out-of-sample error E_{out} and the in-sample error E_{in}:

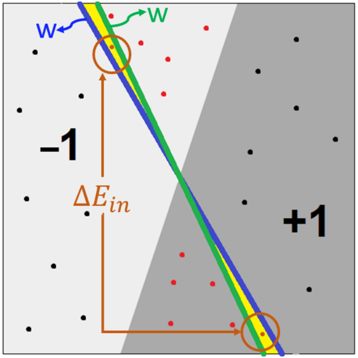




Can we Improve on M ?

Consider a different perceptron **w**:

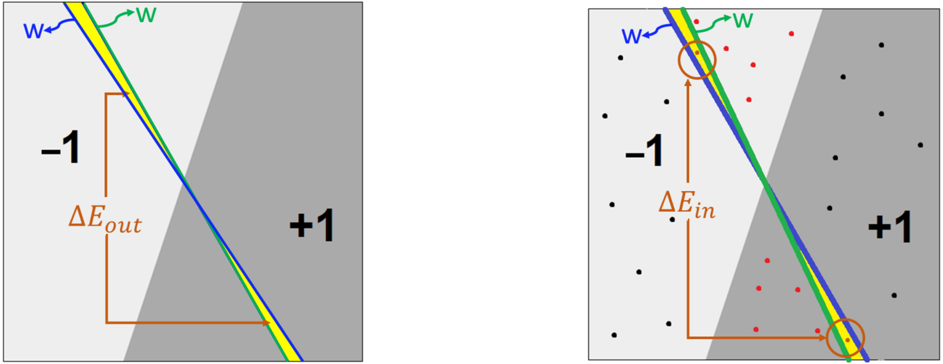




$\triangle E_{out}$ and $\triangle E_{in}$ move in the same direction

Area of yellow part increases $\rightarrow$ probability of data points falling in yellow part increases.

Can we Improve on M ?



$$|E_{in}(h_1) - E_{out}(h_1)| \approx |E_{in}(h_2) - E_{out}(h_2)| \qquad \text{(Both exceed } \epsilon\text{)}$$

Many hypotheses are similar. In PLA, if we slowly vary **w**, we get infinitely many hypotheses that differ from each other infinitesimally.

What can we Replace M with?

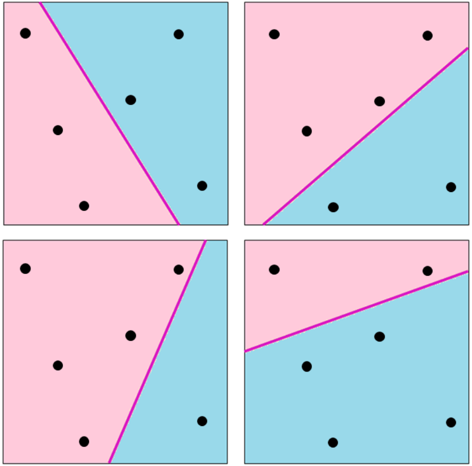
Since the input space $\mathcal{X}$ is infinity, the possible hypotheses are infinity.

Instead of counting the hypotheses over the whole input space, consider a finite set of input points.

On a finite set of input points, how many different 'hypotheses' can I get?

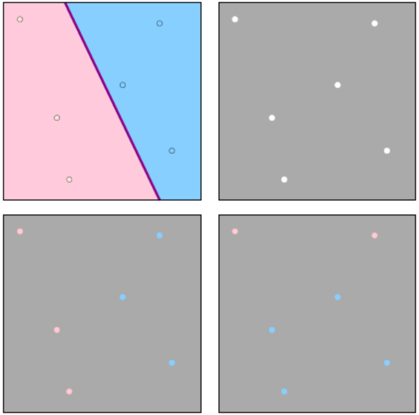
Classification by the four perceptrons is different in at least one data point, so we have four different 'hypotheses'.

Four different perceptrons:



What can we Replace M with?

Hypotheses are seen through the eyes of N points only



Define *dichotomy* as different 'hypotheses' over the finite set of N input points.

Definition: Let $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}$. The *dichotomies* generated by $\mathcal{H}$ are

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) | h \in \mathcal{H}\}$$

Vary perceptron until the line crosses one of the points $\rightarrow$ different *dichotomy*.

Dichotomies: Mini-Hypotheses

A hypotheses $h : \mathcal{X} \rightarrow \{-1, +1\}$

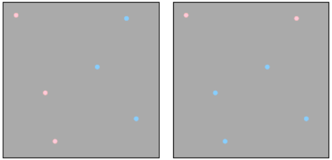
A dichotomy $h : \{\mathbf{x}_1, \mathbf{x}_2, , \cdots, \mathbf{x}_N, \} \rightarrow \{-1, +1\}$

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.

Ex: The two *dichotomies* in the picture could be:
$[-1, -1, -1, +1, +1, +1]$,
$[-1, -1, +1, +1, +1, +1]$.



The Growth Function

The growth function counts the most dichotomies on any N points

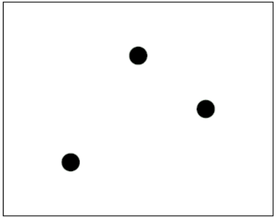
$$m_{\mathcal{H}}(N) \;=\; \max_{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\in\mathcal{X}} |\mathcal{H}(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N)|$$

The value of $m_{\mathcal{H}}(N)$ is at most $|\{-1,+1\}^N|$. Hence, the growth function satisfies:

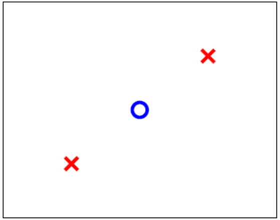
$$m_{\mathcal{H}}(N) \leq 2^N$$

Let's apply the definition.

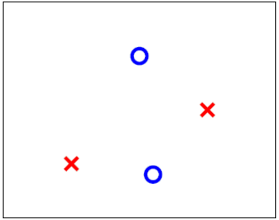
Applying $m_{\mathcal{H}}(N)$ Definition - 2D Perceptrons



Maximum 8 dichotomies with three points.



Dichotomy on 3 colinear points cannot be generated $(N = 4)$



Dichotomy here cannot be generated

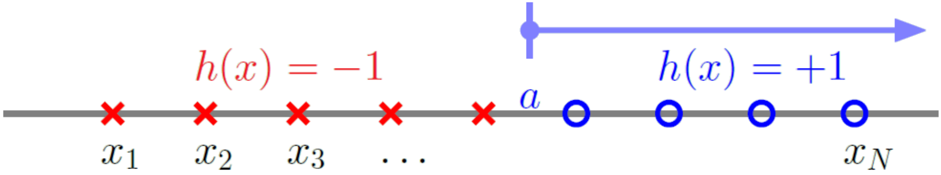
$$m_{\mathcal{H}}(3) = 8 \qquad m_{\mathcal{H}}(4) = 14$$

Note: At most 14 out of the possible 16 dichotomies on any 4 points can be generated.

Outline

- From training to testing
- Illustrative examples
  These examples confirm the intuition that $m_{\mathcal{H}}(N)$ grows faster when $\mathcal{H}$ becomes more complex.
- Key notion: break point
- Puzzle

Example 1: Positive Rays

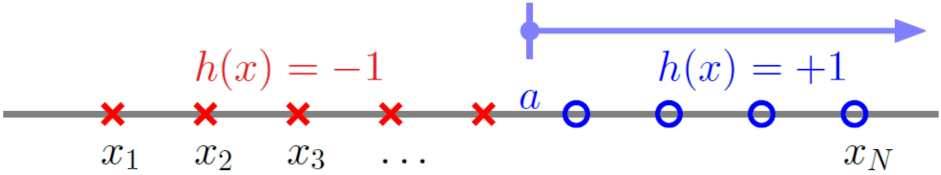


$\mathcal{H}$ is set of $h : \mathbb{R} \rightarrow \{-1, +1\}$

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

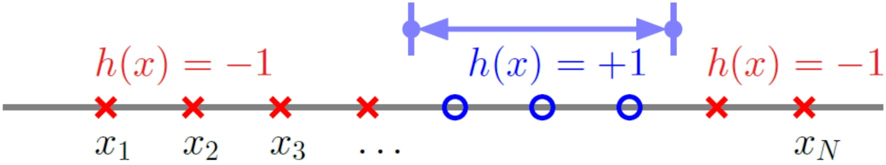
Hypotheses are defined on a one-dimensional input space, and they return -1 to the left of a and $+1$ to the right of a.

Example 1: Positive Rays



N points, split line into $N+1$ regions. As we vary a we get different dichotomies.
The growth function: $m_{\mathcal{H}}(N) = N + 1$
At most $N+1$ dichotomies given any N points.

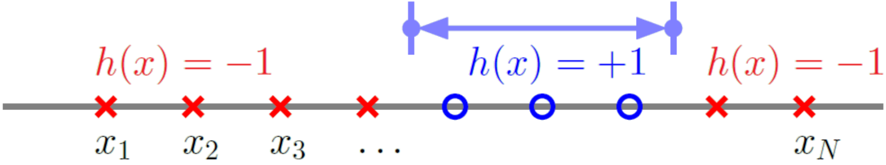
Example 2: Positive Intervals



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

Hypotheses defined on a one-dimensional input space, and they return $+1$ over some interval and -1 otherwise.

Example 2: Positive Intervals



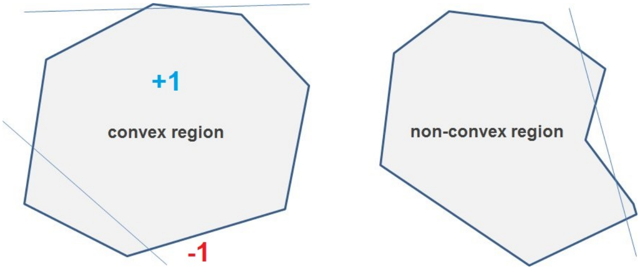
N points, split line into $N+1$ regions.

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

Dichotomies are decided by end values of interval, we have $\binom{N+1}{2}$ possibilities. Add the case in which both end values fall in the same region.

Example 3: Convex Sets

A set is **convex** if a line segment connecting any two points in the set lies entirely within the set

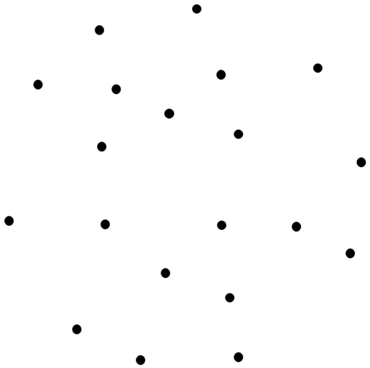


$\mathcal{H}$ consists of all hypotheses in two dimensions that are positive inside some convex set and negative elsewhere

$\mathcal{H}$ is set of $h : \mathbb{R}^2 \to \{-1, +1\}$ $\qquad h(\mathbf{x}) = +1$ is convex

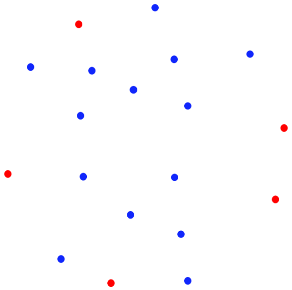
Example 3: Convex Sets

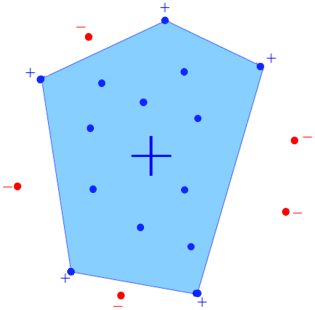
How many patterns can I get out of these data points using convex regions?



Example 3: Convex Sets

How many patterns can I get out of these data points using convex regions?

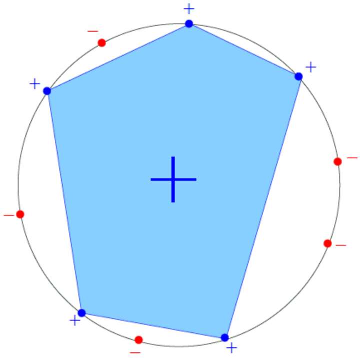




If we consider some outer points to be $+1$, then all interior points are $+1$ (not many dichotomies).

Example 3: Convex Sets

Find another distribution of points to get all possible dichotomies using convex regions?



Place N points over the perimeter of the circle. We get all possible combinations (maximum number of dichotomies).

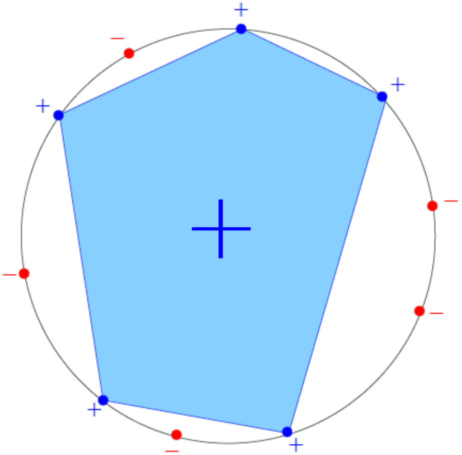
Example 3: Convex Sets

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

Any dichotomy on these N points can be realized using a convex hypothesis.

The N points are 'shattered' by convex sets.

Note: $m_{\mathcal{H}}(N)$ is an upper bound. The number of possible dichotomies for given data points may be less than 2^N because of interior points.



The hypothesis shatters all points

The 3 Growth Functions

- $\mathcal{H}$ is positive rays:
$$m_{\mathcal{H}}(N) = N + 1$$
- $\mathcal{H}$ is positive intervals:
$$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$
- $\mathcal{H}$ is convex sets:
$$m_{\mathcal{H}}(N) = 2^N$$

$m_{\mathcal{H}}(N)$ grows faster when $\mathcal{H}$ becomes more complex.

Back to the Big Picture

Remember this inequality?

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

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

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

If $m_{\mathcal{H}}(N)$ can be bounded by any polynomial, the generalization error will go to zero as $N \to \infty$ $\implies$ Learning is feasible.

Just prove that $m_{\mathcal{H}}(N)$ can be bounded by a polynomial?

Outline

- From training to testing
- Illustrative examples
- Key notion: break point
  It would enable us to proof that $m_{\mathcal{H}}(N)$ can be bounded by a polynomial
- Puzzle

Break Point of $\mathcal{H}$

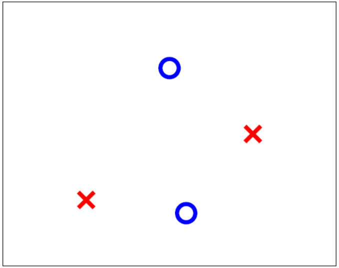
Definition:
If data set of size k cannot be shattered by $\mathcal{H}$, then k is a break point for $\mathcal{H}$

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

The break point k is the number of data points at which we fail to get all possible dichotomies.

A bigger data set cannot be shattered either.

Remember the 2D perceptrons



At most 14 out of 16 dichotomies on any 4 points can be generated.

$$k = 4$$

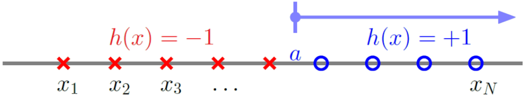
Break Point - the 3 Examples

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

- Positive rays $\qquad m_{\mathcal{H}}(N) = N + 1$

$$\begin{array}{lll} k = 1 & m_{\mathcal{H}}(1) = 2 \not< 2^1 & \\ k = 2 & m_{\mathcal{H}}(2) = 3 < 2^2 & \rightarrow \quad \textbf{break point} \end{array}$$

Intuitively, remember the positive rays:



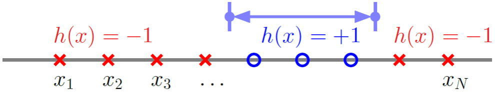
There is no way for the positive ray to generate: 

Break Point - the 3 Examples

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

$$\begin{array}{ll} k=1 & m_{\mathcal{H}}(1) = 2 \nless 2^1 \\ k=2 & m_{\mathcal{H}}(2) = 4 \nless 2^2 \\ k=3 & m_{\mathcal{H}}(3) = 7 < 2^3 \end{array} \quad \rightarrow \quad \textbf{break point}$$

Intuitively, remember the positive intervals:



There is no way to generate:

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

break point $k = `\infty'$

Main Result

We observe how the break point increases with the complexity of the model.

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

Any break point $\rightarrow$ Use k to bound $m_{\mathcal{H}}(N)$ by a polynomial in N

Remember: If $m_{\mathcal{H}}(N)$ can be bounded by any polynomial, the generalization error will go to zero as $N \rightarrow \infty \implies$ Learning is feasible.

To consider learning feasible, all that we need to know now is that there exist a break point.

What we Want

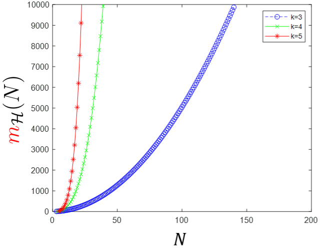
Instead of:

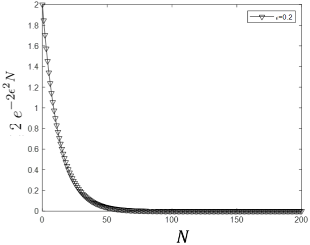
$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad M \quad e^{-2\epsilon^2 N}$$

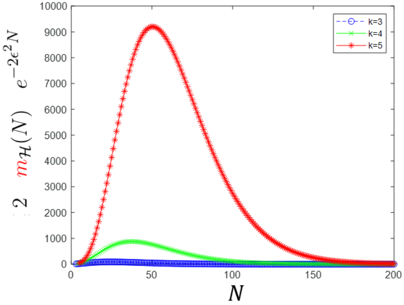
We want:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad m_{\mathcal{H}}(N) \quad e^{-2\epsilon^2 N}$$

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad m_{\mathcal{H}}(N) \quad e^{-2\epsilon^2 N}$$

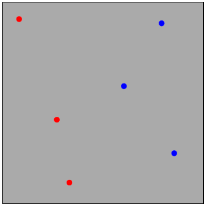






Review

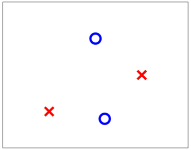
- **Dichotomies:**



- **Growth Function:**

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

- **Break Point** k **:**



At most 14 out of the possible 16 dichotomies on any 4 points can be generated. $k = 4$

- **Maximum # of dichotomies**

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

Bounding the Growth Function

For a given $\mathcal{H}$, if the break point k is fixed, $m_{\mathcal{H}}(N)$ can be bounded by a polynomial$^{(*)}$:

Theorem:
If $m_{\mathcal{H}}(k) < 2^k$ for some value k, then

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

for all N. The RHS is polynomial of degree $k-1$.

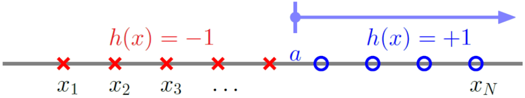
Note: This ensures good generalization on the Hoeffding's Inequality.

$^{(*)}$ Proof can be found on the book: Learning from Data, Yaser S. Abu-Mostafa, Malik Magdon-Ismail and Hsuan-Tien Lin, AMLbook 2012.

Three examples

Let's take the hypothesis sets for which we compute the growth function:

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



We compute before:

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

No need to know anything about the hypothesis set just that break point $k = 2$

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{1} \binom{N}{i} = N + 1$$

Three examples

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

- $\mathcal{H}$ is positive intervals: (break point $k=3$)

$$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 \quad \leq \quad \sum_{i=0}^{2} \binom{N}{i} = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

- $\mathcal{H}$ is 2D perceptrons: (break point $k=4$)

$$m_{\mathcal{H}}(N) = \quad ? \quad \leq \quad \sum_{i=0}^{3} \binom{N}{i} = \frac{1}{6}N^3 + \frac{5}{6}N + 1$$

What we Want

Instead of:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad M \quad e^{-2\epsilon^2 N}$$

We want:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad m_{\mathcal{H}}(N) \quad e^{-2\epsilon^2 N}$$

Let's consider a pictorial proof:

How does $m_{\mathcal{H}}(N)$ relate to overlaps?

Instead of:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad M \quad e^{-2\epsilon^2 N}$$

We wanted:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 \quad m_{\mathcal{H}}(N) \quad e^{-2\epsilon^2 N}$$

but rather, we get:

$$\boxed{\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4 \quad m_{\mathcal{H}}(2N) \quad e^{-\frac{1}{8}\epsilon^2 N}}$$

The Vapnik-Chervonenkis Inequality

Definition of VC Dimension

The Vapnik-Chervonenkis (VC) dimension of a hypothesis set $\mathcal{H}$ denoted by $d_{\mathsf{VC}}(\mathcal{H})$, is

Largest value of N for which $m_{\mathcal{H}}(N) = 2^N$

" the maximum number of points $\mathcal{H}$ can shatter"

$$k > d_{\mathsf{VC}}(\mathcal{H}) \implies k \text{ is a break point for } \mathcal{H}$$

$$d_{\mathsf{VC}}(\mathcal{H}) = k - 1$$

The Growth Function

In terms of a break point k:

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

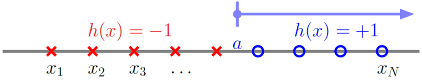
In terms of the d_{VC}:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{\mathsf{VC}}} \binom{N}{i}$$

Maximum power is $N^{d_{\mathsf{VC}}}$

Examples

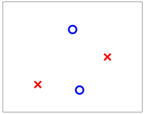
- $\mathcal{H}$ is positive rays: $d_{\text{VC}} = 1$



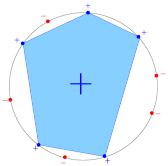
if $N = 2$, we cannot have

- $\mathcal{H}$ is 2D perceptrons: $d_{\text{VC}} = 3$

if $N = 4$, we cannot have



- $\mathcal{H}$ is convex sets: $d_{\text{VC}} = \infty$

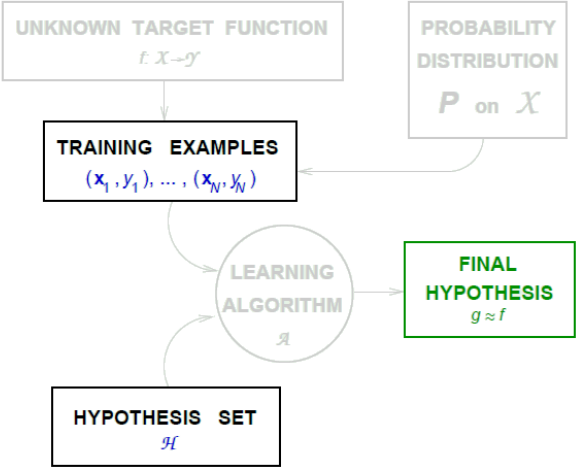


VC Dimension and Learning

Result: If $d_{\mathrm{VC}}(\mathcal{H})$ is finite, $g \in \mathcal{H}$ will generalize.

This statement is true independently of:

- **Learning algorithm**
- **Input distribution**
- **Target function**



VC Dimension and Learning

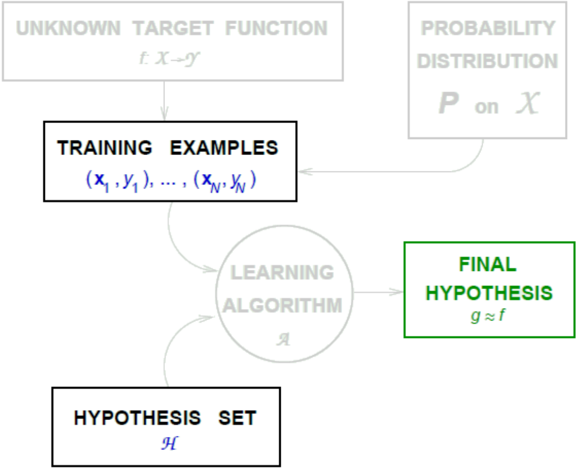
Result: If $d_{\mathrm{VC}}(\mathcal{H})$ is finite, $g \in \mathcal{H}$ will generalize.

This statement depends on:

- **Final hypothesis**
- **Hypothesis set**

  VC dimension depends only on the hypothesis set.
- **Training samples**

  Exist a small chance of having a data set that won't allow generalization.



VC Dimension of Perceptrons

Consider the 2D perceptron:

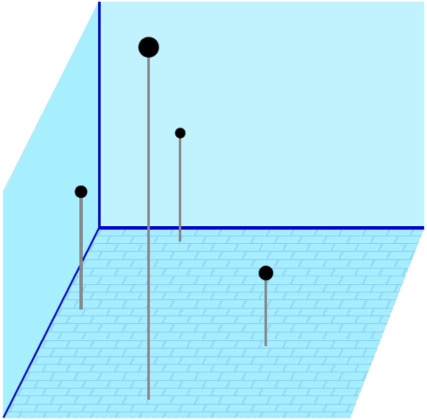
$d = 2$, $d_{\text{VC}} = 3$

In general, for a d-dimensional perceptron:

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

To prove this, we are going to show that:

$$\begin{aligned} d_{\text{VC}} &\leq d+1 \\ d_{\text{VC}} &\geq d+1 \end{aligned}$$



Putting it Together

VC dimension of a d-dimensional perceptron is:

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

What is $d+1$ in the perceptron?

It is the number of parameters $w_0, w_1, \ldots, w_d$,

Note: The more parameters a model has, the more diverse its hypothesis set is, which is reflected in a larger value of the growth function.

Degrees of Freedom

Parameters create degrees of freedom

# of parameters: **analog** degrees of freedom

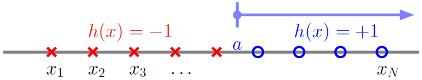
d_{VC}: translates to degrees of freedom.

Parameters are consider as knobs

The Usual Suspects

Let's see if the correspondence between degrees of freedom and VC dimension holds.

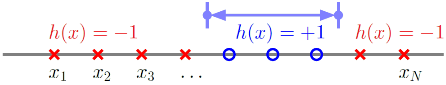
- Positive rays ($d_{\mathsf{VC}} = 1$):



we cannot have

Each hypothesis is specified by the parameter a (one degree of freedom).

- Positive Intervals ($d_{\mathsf{VC}} = 2$)



we cannot have

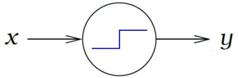
Each hypothesis is specified by the two end values of the interval (two degrees of freedom).

Not Just Parameters

Parameters may not contribute degrees of freedom:

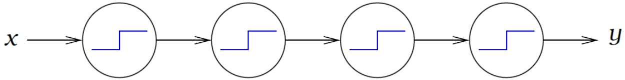
Example: consider a one-dimensional perceptron $h(x) = \text{sign}(w_0 + w_1 x)$ where w_0 is a threshold.

$$y = h(x) = \begin{cases} 1 & \text{if } w_1 x > -w_0 \\ -1 & \text{if } w_1 x < -w_0 \end{cases}$$



2 parameters and 2 degrees of freedom.

Creating a cascade of perceptrons:



Eight parameters in this model and still two degrees of freedom.

d_{VC} measures the **effective** number of parameters.

Number of Data Points Needed

Two small quantities in the VC inequality:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

If we want certain ϵ and δ, how does N depend on d_{VC}

Let us look at $N^{d}e^{-N}$

Fix $N^{d}e^{-N}$ = small value

How does N change with d?
It is basically proportional.

Rule of thumb:

$$N \geq 10 d_{\text{VC}}$$