# ECE595 / STAT598: Machine Learning I
# Lecture 27 VC Dimension

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

# Outline

- Lecture 25 Generalization
- Lecture 26 Growth Function
- Lecture 27 VC Dimension

**Today's Lecture**:

- From Dichotomy to Shattering
  - Review of dichotomy
  - The Concept of Shattering
  - VC Dimension
- Example of VC Dimension
  - Rectangle Classifier
  - Perceptron Algorithm
  - Two Cases

# Probably Approximately Correct

- **Probably**: Quantify error using probability:

$$\mathbb{P}\left[\,|E_{\text{in}}(h) - E_{\text{out}}(h)| \le \epsilon\right] \ge 1 - \delta$$

- **Approximately Correct**: In-sample error is an approximation of the out-sample error:

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| \le \epsilon\right] \ge 1 - \delta$$

- If you can find an algorithm $\mathcal{A}$ such that for any $\epsilon$ and $\delta$, there exists an $N$ which can make the above inequality holds, then we say that the target function is **PAC-learnable**.

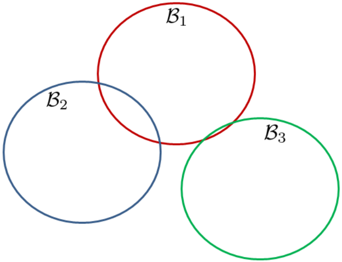
# Overcoming the $M$ Factor

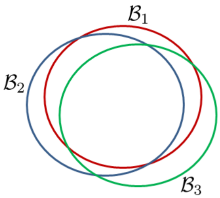
- The $\mathcal{B}$ad events $\mathcal{B}_m$ are

$$\mathcal{B}_m = \{|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\}$$

- The factor $M$ is here because of the Union bound:

$$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \ldots \textbf{ or } \mathcal{B}_M] \leq \mathbb{P}[\mathcal{B}_1] + \ldots + \mathbb{P}[\mathcal{B}_M].$$



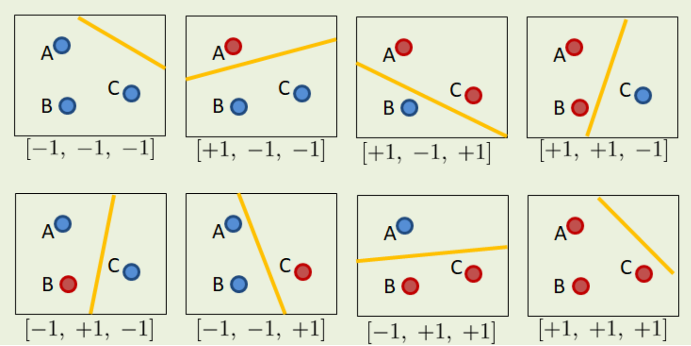



# Dichotomy

## Definition

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{X}$. The **dichotomies** generated by $\mathcal{H}$ on these points are

$$\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_N)) \mid h \in \mathcal{H}\}.$$

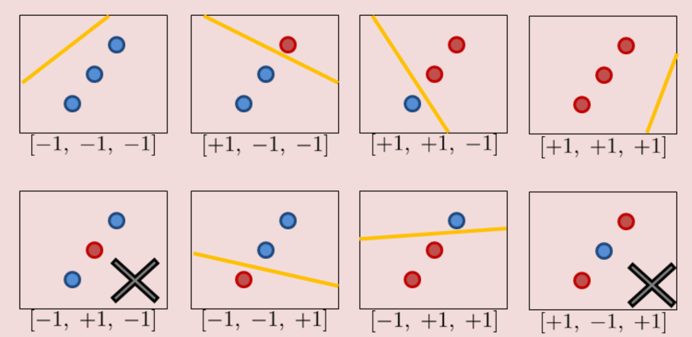


# Dichotomy

**Definition**

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{X}$. The **dichotomies** generated by $\mathcal{H}$ on these points are

$$\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_N)) \mid h \in \mathcal{H}\}.$$



# Candidate to Replace $M$

- So here is our candidate replacement for $M$.
- Define **Growth Function**

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

- You give me a hypothesis set $\mathcal{H}$
- You tell me there are $N$ training samples
- My job: Do whatever I can, by allocating $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, so that the number of dichotomies is maximized
- Maximum number of dichotomy $=$ the best I can do with your $\mathcal{H}$
- $m_{\mathcal{H}}(N)$: How expressive your hypothesis set $\mathcal{H}$ is
- Large $m_{\mathcal{H}}(N)$ $=$ more expressive $\mathcal{H}$ $=$ more complicated $\mathcal{H}$
- $m_{\mathcal{H}}(N)$ only depends on $\mathcal{H}$ and $N$
- Doesn't depend on the learning algorithm $\mathcal{A}$
- Doesn't depend on the distribution $p(\boldsymbol{x})$ (because I'm giving you the max.)

# Summary of the Examples

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

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

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

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

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

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

- So if we can replace $M$ by $m_{\mathcal{H}}(N)$
- And if $m_{\mathcal{H}}(N)$ is a polynomial
- Then we are good.

# Shatter

**Definition**

If a hypothesis set $\mathcal{H}$ is able to generate all $2^N$ dichotomies, then we say that $\mathcal{H}$ **shatter** $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

- $\mathcal{H}$ = hyperplane returned by a perceptron algorithm in 2D.
- If $N = 3$, then $\mathcal{H}$ can shatter
- Because we can achieve $2^3 = 8$ dichotomies
- If $N = 4$, then $\mathcal{H}$ cannot shatter
- Because we can only achieve 14 dichotomies

# VC Dimension

**Definition (VC Dimension)**

The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}$, is the largest value of $N$ for which $\mathcal{H}$ can shatter all $N$ training samples.

- You give me a hypothesis set $\mathcal{H}$, e.g., linear model
- You tell me the number of training samples $N$
- Start with a small $N$
- I will be able to shatter for a while, until I hit a bump
- E.g., linear in 2D: $N = 3$ is okay, but $N = 4$ is not okay
- So I find the **largest** $N$ such that $\mathcal{H}$ can shatter $N$ training samples
- E.g., linear in 2D: $d_{\mathrm{VC}} = 3$
- If $\mathcal{H}$ is complex, then expect large $d_{\mathrm{VC}}$
- Does not depend on $p(\boldsymbol{x})$, $\mathcal{A}$ and $f$

# Outline

- Lecture 25 Generalization
- Lecture 26 Growth Function
- Lecture 27 VC Dimension

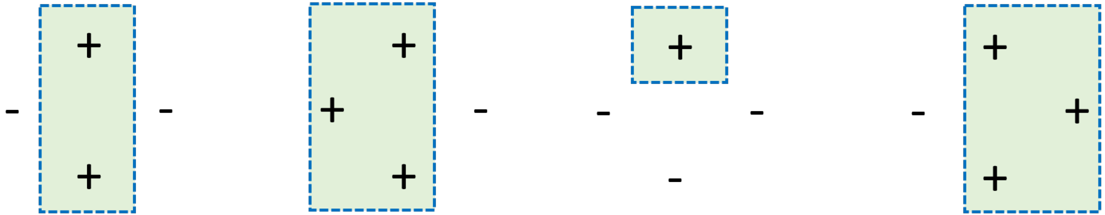
**Today's Lecture**:

- From Dichotomy to Shattering
  - Review of dichotomy
  - The Concept of Shattering
  - VC Dimension
- Example of VC Dimension
  - Rectangle Classifier
  - Perceptron Algorithm
  - Two Cases

# Example: Rectangle

What is the VC Dimension of a 2D classifier with a rectangle shape?

- You can try putting 4 data points in whatever way.
- There will be 16 possible configurations.
- You can show that the rectangle classifier can shatter all these 16 points
- If you do 5 data points, then not possible. (Put one negative in the interior, and four positive at the boundary.)
- So VC dimension is 4.



# VC Dimension of a Perceptron

**Theorem (VC Dimension of a Perceptron)**

*Consider the input space* $\mathcal{X} = \mathbb{R}^d \cup \{1\}$, *i.e.,* ($\boldsymbol{x} = [1,\ x_1,\ \ldots,\ x_d]^T$). *The VC dimension of a perceptron is*

$$d_{\mathrm{VC}} = d + 1.$$

- The "+1" comes from the bias term ($w_0$ if you recall)
- So a linear classifier is "no more complicated" than $d + 1$
- The best it can shatter is $d + 1$ in a $d$-dimensional space
- E.g., If $d = 2$, then $d_{\mathrm{VC}} = 3$

# Why?

- We claim $d_{\mathrm{VC}} \geq d+1$ and $d_{\mathrm{VC}} \leq d+1$
- $d_{\mathrm{VC}} \geq d+1$:

$$\mathcal{H} \text{ can shatter } \textbf{at least } d+1 \text{ points}$$

- It may shatter more, or it may not shatter more. We don't know by just looking at this statement
- $d_{\mathrm{VC}} \leq d+1$:

$$\mathcal{H} \text{ cannot shatter } \textbf{more than } d+1 \text{ points}$$

- So with $d_{\mathrm{VC}} \geq d+1$, we show that $d_{\mathrm{VC}} = d+1$

## $d_{\mathrm{VC}} \geq d + 1$

- Goal: Show that there is at least one configuration of $d+1$ points that can be shattered by $\mathcal{H}$
- Think about the 2D case: Put the three points anywhere not on the same line
- Choose

$$\boldsymbol{x}_n = [1,\ 0, \ldots,\ 1,\ \ldots,\ 0]^T.$$

- Linear classifier: $\text{sign}(\boldsymbol{w}^T \boldsymbol{x}_n) = y_n$.
- For all $d+1$ data points, we have

$$\text{sign}\left( \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & & 0 \\ & & & \ddots & 0 \\ 1 & 0 & \ldots & 0 & 1 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \right) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$$

## $d_{\mathrm{VC}} \geq d+1$

- We can remove the sign because we are trying to find **one** configuration of points that can be shattered.

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ & & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$$

- We are only interested in whether the problem **solvable**
- So we just need to see if we can ever find a $\boldsymbol{w}$ that shatters
- If there exists at least one $\boldsymbol{w}$ that makes all $\pm 1$ correct, then $\mathcal{H}$ can shatter (if you use that particular $\boldsymbol{w}$)
- So is this $(d+1) \times (d+1)$ system invertible?
- Yes. It is. So $\mathcal{H}$ can shatter at least $d+1$ points

# $d_{\mathrm{VC}} \leq d + 1$

- Can we shatter more than $d + 1$ points?
- No.
- You only have $d + 1$ variables
- If you have $d + 2$ equations, then one equation will be either redundant or contradictory
- If redundant, you can ignore it because it is not the worst case
- If contradictory, then you cannot solve the system of linear equation
- So we cannot shatter more than $d + 1$ points
- You can always construct a nasty $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}$ to cause contradiction

# $d_{\mathrm{VC}} \leq d + 1$

- You give me $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}, \boldsymbol{x}_{d+2}$
- I can always write $\boldsymbol{x}_{d+2}$ as

$$\boldsymbol{x}_{d+2} = \sum_{i=1}^{d+1} a_i \boldsymbol{x}_i$$

- Not all $a_i$'s are zero. Otherwise it will be trivial.
- My job: Construct a dichotomy which cannot be shattered by any $h$.
- Here is a dichotomy.
- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}$ get $y_i = \mathrm{sign}(a_i)$.
- $\boldsymbol{x}_{d+2}$ gets $y_{d+2} = -1$.

# $d_{\mathrm{VC}} \leq d+1$

- Then

$$\mathbf{w}^T\mathbf{x}_{d+2} = \sum_{i=1}^{d+1} a_i \mathbf{w}^T\mathbf{x}_i.$$

- Perceptron: $y_i = \mathrm{sign}(\mathbf{w}^T\mathbf{x}_i)$.
- By our design, $y_i = \mathrm{sign}(a_i)$.
- So $a_i\mathbf{w}^T\mathbf{x}_i > 0$
- This forces

$$\sum_{i=1}^{d+1} a_i \mathbf{w}^T\mathbf{x}_i > 0.$$

- So $y_{d+2} = \mathrm{sign}(\mathbf{w}^T\mathbf{x}_{d+2}) = +1$, contradiction.
- So we found a dichotomy which cannot be shattered by any $h$.

# Summary of the Examples

- $\mathcal{H}$ is positive ray: $m_{\mathcal{H}}(N) = N + 1$.
  - If $N = 1$, then $m_{\mathcal{H}}(1) = 2$
  - If $N = 2$, then $m_{\mathcal{H}}(2) = 3$
  - So $d_{\mathrm{VC}} = 1$
- $\mathcal{H}$ is positive interval: $m_{\mathcal{H}}(N) = \frac{N^2}{2} + \frac{N}{2} + 1$.
  - If $N = 2$, then $m_{\mathcal{H}}(2) = 4$
  - If $N = 4$, then $m_{\mathcal{H}}(4) = 5$
  - So $d_{\mathrm{VC}} = 2$
- $\mathcal{H}$ is perceptron in $d$-dimensional space
  - Just showed
  - $d_{\mathrm{VC}} = d + 1$
- $\mathcal{H}$ is convex set: $m_{\mathcal{H}}(N) = 2^N$
  - No matter which $N$ we choose, we always have $m_{\mathcal{H}}(N) = 2^N$
  - So $d_{\mathrm{VC}} = \infty$
  - The model is as complex as it can be

## Reading List

- Yasar Abu-Mostafa, Learning from Data, chapter 2.1
- Mehrya Mohri, Foundations of Machine Learning, Chapter 3.2
- Stanford Note http://cs229.stanford.edu/notes/cs229-notes4.pdf

# Appendix

# Radon Theorem

The perceptron example we showed in this lecture can be proved using Radon's theorem.

**Theorem (Radon's Theorem)**

*Any set $\mathcal{X}$ of $d+2$ data points in $\mathbb{R}^d$ can be partitioned into two subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ such that the convex hulls of $\mathcal{X}_1$ and $\mathcal{X}_2$ intersect.*

Proof: See Mehryar Mohri, Foundations of Machine Learning, Theorem 3.13.

- If two sets are separated by a hyperplane, then their convex hulls are separated.
- So if you have $d+2$ points, Radon says the convex hulls intersect.
- So you cannot shatter the $d+2$ points.
- $d+1$ is okay as we have proved. So the VC dimension is $d+1$.