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**: Quantify error using probability:
  \[
  \mathbb{P}
  \left[
  \left|E_{in}(h) - E_{out}(h)\right| \leq \epsilon
  \right] \geq 1 - \delta
  \]

- **Approximately Correct**: In-sample error is an approximation of the out-sample error:
  \[
  \mathbb{P}
  \left[
  \left|E_{in}(h) - E_{out}(h)\right| \leq \epsilon
  \right] \geq 1 - \delta
  \]

- If you can find an algorithm \(A\) 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 bad events $B_m$ are
  \[ B_m = \{|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\} \]

- The factor $M$ is here because of the Union bound:
  \[ P[B_1 \text{ or } \ldots \text{ or } B_M] \leq P[B_1] + \ldots + P[B_M]. \]
**Dichotomy**

**Definition**

Let $x_1, \ldots, x_N \in \mathcal{X}$. The **dichotomies** generated by $\mathcal{H}$ on these points are

$$\mathcal{H}(x_1, \ldots, x_N) = \{(h(x_1), \ldots, h(x_N)) \mid h \in \mathcal{H}\}.$$
Dichotomy

Definition

Let $x_1, \ldots, x_N \in \mathcal{X}$. The **dichotomies** generated by $\mathcal{H}$ on these points are

$$\mathcal{H}(x_1, \ldots, x_N) = \{(h(x_1), \ldots, h(x_N)) \mid h \in \mathcal{H}\}.$$
Candidate to Replace $M$

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

$$m_{\mathcal{H}}(N) = \max_{x_1, \ldots, x_N \in \mathcal{X}} |\mathcal{H}(x_1, \ldots, x_N)|$$

- You give me a hypothesis set $\mathcal{H}$
- You tell me there are $N$ training samples
- My job: Do whatever I can, by allocating $x_1, \ldots, x_N$, 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 $m_{\mathcal{H}}(N) = \text{more expressive}$ $\mathcal{H} = \text{more complicated}$ $\mathcal{H}$
- $m_{\mathcal{H}}(N)$ only depends on $\mathcal{H}$ and $N$
- Doesn’t depend on the learning algorithm $\mathcal{A}$
- Doesn’t depend on the distribution $p(x)$ (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** \( x_1, \ldots, x_N \).

- \( \mathcal{H} = \) 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
Definition (VC Dimension)

The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\text{VC}}$, 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: $d_{\text{VC}} = 3$
- If $\mathcal{H}$ is complex, then expect large $d_{\text{VC}}$
- Does not depend on $p(x)$, $A$ 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., $(x = [1, x_1, \ldots, x_d]^T)$. The VC dimension of a perceptron is

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

- 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 $d_{VC} = 3$
Why?

- We claim $d_{VC} \geq d + 1$ and $d_{VC} \leq d + 1$
- $d_{VC} \geq d + 1$:
  \[ \mathcal{H} \text{ can shatter at least } d + 1 \text{ points} \]
- It may shatter more, or it may not shatter more. We don’t know by just looking at this statement
- $d_{VC} \leq d + 1$:
  \[ \mathcal{H} \text{ cannot shatter more than } d + 1 \text{ points} \]
- So with $d_{VC} \geq d + 1$, we show that $d_{VC} = d + 1$
\( d_{VC} \geq d + 1 \)

- 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
  \[
x_n = [1, 0, \ldots, 1, \ldots, 0]^T.
  \]
- Linear classifier: \( \text{sign}(w^T x_n) = y_n \).
- For all \( d + 1 \) data points, we have
  \[
  \text{sign} \begin{pmatrix}
  1 & 0 & 0 & \cdots & 0 \\
  1 & 1 & 0 & \cdots & 0 \\
  1 & 0 & 1 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  1 & 0 & \cdots & 0 & 1 \\
  \end{pmatrix}
  \begin{pmatrix}
  w_0 \\
  w_1 \\
  \vdots \\
  w_d \\
  \end{pmatrix}
  = \begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_{d+1} \\
  \end{pmatrix}
  = \begin{pmatrix}
  \pm 1 \\
  \pm 1 \\
  \vdots \\
  \pm 1 \\
  \end{pmatrix}
  \]
\[ d_{VC} \geq d + 1 \]

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

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

- We are only interested in whether the problem solvable
- So we just need to see if we can ever find a \( w \) that shatters
- If there exists at least one \( w \) that makes all \( \pm 1 \) correct, then \( H \) can shatter (if you use that particular \( w \))
- So is this \((d + 1) \times (d + 1)\) system invertible?
- Yes. It is. So \( H \) can shatter at least \( d + 1 \) points
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 $x_1, \ldots, x_{d+1}$ to cause contradiction
You give me \( x_1, \ldots, x_d+1, x_{d+2} \)

I can always write \( x_{d+2} \) as

\[
x_{d+2} = \sum_{i=1}^{d+1} a_i x_i
\]

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.

\( x_1, \ldots, x_{d+1} \) get \( y_i = \text{sign}(a_i) \).

\( x_{d+2} \) gets \( y_{d+2} = -1 \).
$d_{VC} \leq d + 1$

- Then

$$w^T x_{d+2} = \sum_{i=1}^{d+1} a_i w^T x_i.$$ 

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

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

- So $y_{d+2} = \text{sign}(w^T x_{d+2}) = +1$, 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 $d_{VC} = 1$

- $\mathcal{H}$ 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 $d_{VC} = 2$

- $\mathcal{H}$ is perceptron in $d$-dimensional space
  - Just showed
  - $d_{VC} = d + 1$

- $\mathcal{H}$ 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 $d_{VC} = \infty$
  - 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
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.


- 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 $d + 1$. 