Preamble to the theory
Training versus testing

- Out-of-sample error (generalization error):

\[ E_{\text{out}} = E_x[e(h(x), f(x))] \]

What we want: small \( E_{\text{out}} \)
Training versus testing

- Out-of-sample error (generalization error):

\[ E_{\text{out}} = E_x[e(h(x), f(x))] \]

What we want: small \( E_{\text{out}} \)

- In-sample error (training error):

\[ E_{\text{in}} = \frac{1}{N} \sum_{n=1}^{N} e(h(x_n), f(x_n)) \]

This is what we can minimize
The 2 questions of learning

- $E_{out}(g) \approx 0$ is achieved through:

$$E_{out}(g) \approx E_{in}(g) \quad \text{and} \quad E_{in}(g) \approx 0$$
The 2 questions of learning

- $E_{\text{out}}(g) \approx 0$ is achieved through:

  \[ E_{\text{out}}(g) \approx E_{\text{in}}(g) \quad \text{and} \quad E_{\text{in}}(g) \approx 0 \]

- Learning is thus split into 2 questions:
  - Can we make sure that $E_{\text{out}}(g) \approx E_{\text{in}}(g)$?
    - Hoeffding’s inequality (?)
  - Can we make $E_{\text{in}}(g)$ small?
    - Optimization (done)
What the theory will achieve

- Currently we only know

\[ P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2N} \]
What the theory will achieve

- Currently we only know

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

- What if \( M = \infty \)?
  (e.g., perceptron)
What the theory will achieve

- Currently we only know

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

- What if \( M = \infty \)?
  (e.g., perceptron)

- **Todo:**
  We will establish a finite quantity to replace \( M \)

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

- Study \( m_\mathcal{H}(N) \) to understand the trade-off for model complexity
Reducing $M$ to finite number
Where did the $M$ come from?

- The Bad events $B_m$:
  
  
  $|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon$ with probability $\leq 2e^{-2\epsilon^2 N}$
Where did the \( M \) come from?

- The \( \text{Bad} \) events \( \mathcal{B}_m \): \n  \[ \left| E_{\text{in}}(h_m) - E_{\text{out}}(h_m) \right| > \epsilon \]  with probability \( \leq 2e^{-2\epsilon^2 N} \)

- The union bound:
  \[
P[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \cdots \text{ or } \mathcal{B}_M] \leq P[\mathcal{B}_1] + P[\mathcal{B}_2] + \cdots + P[\mathcal{B}_M] \leq 2Me^{-2\epsilon^2 N}
\]

\[\text{consider worst case: no overlaps}\]

---

![Diagram showing overlap]

No overlap: bound is tight

Large overlap
Can we improve on $M$?
Can we improve on $M$?
Can we improve on $M$?
Can we improve on $M$?

- $\Delta E_{\text{out}}$: change in $+1$ and $-1$ areas
- $\Delta E_{\text{in}}$: change in labels of data points
Can we improve on $M$?

- $\Delta E_{\text{out}}$: change in $+1$ and $-1$ areas
- $\Delta E_{\text{in}}$: change in labels of data points

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

Overlapped events!
What can we replace $M$ with?

Instead of the whole input space
What can we replace $M$ with?

Instead of the whole input space
Let’s consider a finite set of input points
What can we replace $M$ with?

Instead of the whole input space
Let’s consider a finite set of input points
How many patterns of red and blue can you get?
A hypothesis: $h : \mathcal{X} \rightarrow \{-1, +1\}$
Dichotomies: mini-hypotheses

- A hypothesis: \( h : \mathcal{X} \rightarrow \{-1, +1\} \)
- A dichotomy: \( h : \{x_1, x_2, \ldots, x_N\} \rightarrow \{-1, +1\} \)

Number of hypotheses \(|\mathcal{H}|\) can be infinite
Number of dichotomies \(|\mathcal{H}(x_1, x_2, \ldots, x_N)|\): at most \(2^N\)

Candidate for replacing \(M\)
Dichotomies: mini-hypotheses

- A hypothesis: \( h : \mathcal{X} \rightarrow \{-1, +1\} \)
- A dichotomy: \( h : \{x_1, x_2, \cdots, x_N\} \rightarrow \{-1, +1\} \)
- Number of hypotheses \(|\mathcal{H}|\) can be infinite
Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \rightarrow \{-1, +1\}$
- A dichotomy: $h : \{x_1, x_2, \cdots, x_N\} \rightarrow \{-1, +1\}$
- Number of hypotheses $|\mathcal{H}|$ can be infinite
- Number of dichotomies $|\mathcal{H}(x_1, x_2, \cdots, x_N)|$: at most $2^N$
Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \rightarrow \{-1, +1\}$
- A dichotomy: $h : \{x_1, x_2, \cdots, x_N\} \rightarrow \{-1, +1\}$
- Number of hypotheses $|\mathcal{H}|$ can be infinite
- Number of dichotomies $|\mathcal{H}(x_1, x_2, \cdots, x_N)|$:
  at most $2^N$
Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \rightarrow \{-1, +1\}$
- A dichotomy: $h : \{x_1, x_2, \ldots, x_N\} \rightarrow \{-1, +1\}$
- Number of hypotheses $|\mathcal{H}|$ can be infinite
- Number of dichotomies $|\mathcal{H}(x_1, x_2, \ldots, x_N)|$: at most $2^N$
- Candidate for replacing $M$
The growth function

The growth function counts the most dichotomies on any $N$ points:

$$m_{\mathcal{H}}(N) = \max_{x_1, \cdots, x_N \in \mathcal{X}} |\mathcal{H}(x_1, \cdots, x_N)|$$
The growth function

- The growth function counts the most dichotomies on any $N$ points:

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

- The growth function satisfies:

$$m_{\mathcal{H}}(N) \leq 2^N$$
Growth function for perceptrons

Compute $m_\mathcal{H}(3)$ in 2-D space

What's $|\mathcal{H}(x_1, x_2, x_3)|$?
Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)

$$m_{\mathcal{H}}(3) = 8$$
Growth function for perceptrons

Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)
Growth function for perceptrons

Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)

Doesn’t matter because we only counts the most dichotomies
Growth function for perceptrons

- What’s $m_{\mathcal{H}}(4)$?
Growth function for perceptrons

- What’s $m_\mathcal{H}(4)$?
- (At least) missing two dichotomies:
Growth function for perceptrons

- What’s $m_{\mathcal{H}}(4)$?
- (At least) **missing** two dichotomies:

  ![Diagram showing two examples](image)

- $m_{\mathcal{H}}(4) = 14 < 2^4$
Example I: positive rays

\[ h(x) = -1 \]

\[ x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_N \]

\[ h(x) = +1 \]

\[ a \]

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

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

\[ m_\mathcal{H}(N) = N + 1 \]
Example II: positive intervals

\[ h(x) = -1 \]

\( x_1 \quad x_2 \quad x_3 \quad \ldots \)

\[ h(x) = +1 \]

\[ h(x) = -1 \]

\( x_N \)

\( \mathcal{H} \) 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 III: convex sets

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

- How many dichotomies can we generate?

[Diagram of points within a rectangle]
Example III: convex sets

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

- How many dichotomies can we generate?
Example III: convex sets

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

- How many dichotomies can we generate?
Example III: convex sets

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

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

- We say 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:
  \[ m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 \]

- $\mathcal{H}$ is convex sets:
  \[ m_{\mathcal{H}}(N) = 2^N \]
What’s next?

- Remember the inequality

\[ \mathbb{P}[|E_{\text{in}} - E_{\text{out}}| > \epsilon] \leq 2M e^{-2\epsilon^2 N} \]
What’s next?

- Remember the inequality
  \[ P[|E_{in} - E_{out}| > \epsilon] \leq 2Me^{-2\epsilon^2N} \]

- What happens if we replace \( M \) by \( m_H(N) \)?
  \( m_H(N) \) polynomial \( \Rightarrow \) Good!
What’s next?

- Remember the inequality

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

- What happens if we replace \( M \) by \( m_{\mathcal{H}}(N) \)?

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

- How to show \( m_{\mathcal{H}}(N) \) is polynomial?
Conclusions

- Next class: LFD 2.1, 2.2

Questions?