Recall

Unknown Target Function

\[ f: \mathcal{X} \rightarrow \mathcal{Y} \]

Training Data

\[ \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \]

Hypothesis Set

\[ \mathcal{H} \]

Learned Hypothesis

\[ g \approx f \]

Probability Distribution

\[ \mathcal{P} \text{ on } \mathcal{X} \]

Learning Algorithm

\[ \mathcal{A} \]
$E(h, f)$ is a function that measures how close a hypothesis $h$ is to $f$: the smaller the better

$g = \arg \min_{h \in \mathcal{H}} E(h, f)$

$E$ is usually defined in terms of a pointwise error function $e(h, f, \vec{x})$

- Binary error (classification): $e(h, f, \vec{x}) = \|f(\vec{x}) \neq h(\vec{x})\|$
- Squared error (regression): $e(h, f, \vec{x}) = (f(\vec{x}) - h(\vec{x}))^2$
Error: is $h \approx f$?

- In-sample error: $E_{in}(h) = \frac{1}{n} \sum_{i=1}^{n} e(h, f, \bar{x}_i)$

- Out-of-sample error: $E_{out}(h) = \mathbb{E}_{x \sim \mathcal{P}}[e(h, f, \bar{x})]$
<table>
<thead>
<tr>
<th>$f(x)$</th>
<th>+1</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(x)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>No error</td>
<td>False positive</td>
</tr>
<tr>
<td>-1</td>
<td>False negative</td>
<td>No error</td>
</tr>
</tbody>
</table>
Error: Classification

• Fingerprint recognition:
  • Inputs are fingerprints
  • Outputs: +1 means “you”, -1 means “not you”

• For personalized coupons:

\[
\begin{array}{c|cc}
 f(x) & +1 & -1 \\ \hline
 h(x) & +1 & 0 & 1 \\ & -1 & 100 & 0 \\
\end{array}
\]

For unlocking phones:

\[
\begin{array}{c|cc}
 f(x) & +1 & -1 \\ \hline
 h(x) & +1 & 0 & 1000 \\ & -1 & 1 & 0 \\
\end{array}
\]
Unknown Target Function
\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

Probability Distribution
\( \mathcal{P} \text{ on } \mathcal{X} \)

Training Data
\[ \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \]

Error Measure
\( E \)

Learning Algorithm
\( \mathcal{A} \)

Hypothesis Set
\( \mathcal{H} \)

Learned Hypothesis
\( g \approx f \)
Noise

- The target function is not always deterministic, it is sometimes **stochastic**
- Instead of a target function, a target **distribution**
- Instead of \( y = f(\vec{x}) \), \( P\{y|\vec{x}\} \)
  - \( y = f(\vec{x}) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \)
Unknown Target Distribution

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \text{ plus noise} \]

Training Data

\[ \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \]

Probability Distribution

\[ \mathcal{P} \text{ on } \mathcal{X} \]

Error Measure

\[ E \]

Learning Algorithm

\[ \mathcal{A} \]

Hypothesis Set

\[ \mathcal{H} \]

Learned Hypothesis

\[ g \approx f \]
The general supervised learning setup

Unknown Target Distribution

\( f: \mathcal{X} \rightarrow \mathcal{Y} \) plus noise

Training Data

\( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \)

Probability Distribution

\( \mathcal{P} \) on \( \mathcal{X} \)

Error Measure

\( E \)

Learning Algorithm

\( \mathcal{A} \)

Hypothesis Set

\( \mathcal{H} \)

Learned Hypothesis

\( g \approx f \)
• Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, \ldots, h_m\}$

• $E_{in}(g) = \text{in-sample error of best hypothesis in } \mathcal{H}$

• $E_{out}(g) = \text{out-of-sample error of best hypothesis in } \mathcal{H}$

• Generalization error of $g = |E_{in}(g) - E_{out}(g)|$

• $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2(m)e^{-2\epsilon^2n}$

Or...

• $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2n}\log\left(\frac{2m}{\delta}\right)}$ with probability $\geq 1 - \delta$
Recall

• Suppose \( \mathcal{H} \) is finite i.e. \( \mathcal{H} = \{h_1, \ldots, h_m\} \)

\[
P\{|E_{in}(g) - E_{out}(g)| > \epsilon\}
\]

\[
\leq P \left\{ \bigcup_{j=1}^{m} |E_{in}(h_j) - E_{out}(h_j)| > \epsilon \right\}
\]

\[
\leq \sum_{j=1}^{m} P\{|E_{in}(h_j) - E_{out}(h_j)| > \epsilon\}
\]

\[
\leq \sum_{j=1}^{m} 2e^{-2\epsilon^2n} = 2(m)e^{-2\epsilon^2n}
\]
The Union Bound...

\[ P\{A \cup B\} \leq P\{A\} + P\{B\} \]
The Union Bound is bad

\[ P\{A \cup B\} \leq P\{A\} + P\{B\} \]
\[ P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\} \]
If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

\[ |E_{in}(h_1) - E_{out}(h_1)| > \epsilon \text{ and } |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \]

are very likely to overlap

\[ P\{|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \ \cap \ |E_{in}(h_2) - E_{out}(h_2)| > \epsilon\} \text{ is big} \]
Given some finite sample of points $(x_1, ..., x_n)$ from the input space and a single hypothesis $h \in \mathcal{H}$, applying $h$ to each point in $(x_1, ..., x_n)$ results in a **dichotomy**

- $(h(x_1^1), ..., h(x_n^1))$ is a vector of $n$ +1’s and -1’s

Given $(x_1, ..., x_n)$, each hypothesis in $\mathcal{H}$ generates a dichotomy but not necessarily a unique dichotomy!

- The set of dichotomies induced by $\mathcal{H}$ on $(x_1, ..., x_n)$ is $\mathcal{H}(x_1, ..., x_n) = \{(h(x_1^1), ..., h(x_n^1)) \mid h \in \mathcal{H}\}$
Dichotomy Example

\[ \mathcal{H} = \{h_1, h_2, h_3\} \]
Dichotomy Example

\[ \mathcal{H} = \{h_1, h_2, h_3\} \]

\[
(h_1(x_1), h_1(x_2), h_1(x_3), h_1(x_4)) = (-1, +1, -1, +1)
\]
Dichotomy Example

\[ \mathcal{H} = \{h_1, h_2, h_3\} \]

\[
(h_2(\overrightarrow{x_1}), h_2(\overrightarrow{x_2}), h_2(\overrightarrow{x_3}), h_2(\overrightarrow{x_4}))
= (-1, +1, -1, +1)
\]
Dichotomy Example

\[ \mathcal{H} = \{ h_1, h_2, h_3 \} \]

\[
(h_3(x_1), h_3(x_2), h_3(x_3), h_3(x_4))
= (+1, +1, -1, -1)\

Dichotomy Example

\[ \mathcal{H} = \{h_1, h_2, h_3\} \]

\[ \mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4) = \{(-1, +1, -1, +1), (+1, +1, -1, -1)\} \]

\[ |\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4)| = 2 \]
Dichotomy Example

\[ \mathcal{H} = \{ h_1, h_2, h_3 \} \]

\[ \mathcal{H}(x_1, x_2, x_3, x_4) = \{ (+1, +1, -1, -1) \} \]

|\mathcal{H}(x_1, x_2, x_3, x_4)| = 1
• The growth function of $\mathcal{H}$ is the largest number of dichotomies $\mathcal{H}$ can induce across all possible data sets of size $n$

$$m_{\mathcal{H}}(n) = \max_{(x_1, \ldots, x_n) \in X} |\mathcal{H}(x_1, \ldots, x_n)|$$
• Observe that $m_{\mathcal{H}}(n) \leq 2^n \forall \mathcal{H}$ and $n$

• Given $\mathcal{H}$, if $\exists \ (x_1, \ldots, x_n)$ s.t. $|\mathcal{H}(x_1, \ldots, x_n)| = 2^n$, then $\mathcal{H}$ shatters $(x_1, \ldots, x_n)$

• If $\exists \ (x_1, \ldots, x_n)$ that is shattered by $\mathcal{H}$, then $m_{\mathcal{H}}(n) = 2^n$
• \( \mathcal{X} = \mathbb{R}^2 \) and \( \mathcal{H} \) = Linear classifiers (perceptrons)

• What is \( m_\mathcal{H}(3) \)?
• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} = \text{Linear classifiers (perceptrons)}$

• What is $m_\mathcal{H}(3)$?
Growth Function: Example

• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} = \text{Linear classifiers (perceptrons)}$

• What is $m_\mathcal{H}(3)$?
• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H}$ = Linear classifiers (perceptrons)

• What is $m_\mathcal{H}(3)$?

| $\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ | = 6 |
| $\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ | = 8 |
• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H}$ = Linear classifiers (perceptrons)

• $m_H(3) = 8 = 2^3$

| $|\mathcal{H}(x_1, x_2, x_3)| = 6$ | $|\mathcal{H}(x_1, x_2, x_3)| = 8$ |
Break: What is $m_{\mathcal{H}}(4)$?
• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} =$ Linear classifiers (perceptrons)

• What is $m_\mathcal{H}(4)$?
Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} =$ Linear classifiers (perceptrons)

- What is $m_\mathcal{H}(4)$?
• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} =$ Linear classifiers (perceptrons)

• What is $m_\mathcal{H}(4)$?

$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)| = 14$
Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H}$ = Linear classifiers (perceptrons)

- $m_\mathcal{H}(4) = 14 < 2^4$

$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)| = 14$$
• $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} =$ Positive rays: $h(x) = \text{sign}(x - a)$

• What is $m_{\mathcal{H}}(n)$?
Growth Function: Example

- $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \text{Positive rays}: h(x) = \text{sign}(x - a)$

- $m_{\mathcal{H}}(n) = n + 1$
Growth Function: Example

- $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} =$ Positive intervals

- What is $m_{\mathcal{H}}(n)$?
Growth Function: Example

- $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} =$ Positive intervals

- $m_\mathcal{H}(n) = \binom{n+1}{2} + 1$
Growth Function: Example

- \( \mathcal{X} = \mathbb{R}^2 \) and \( \mathcal{H} = \) Convex sets

- What is \( m_{\mathcal{H}}(n) \)?
Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} = \text{Convex sets}$
- What is $m_{\mathcal{H}}(n)$?
• $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} =$ Convex sets

• What is $m_{\mathcal{H}}(n)$?
Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} =$ Convex sets

- $m_\mathcal{H}(n) = 2^n$
• If $m_H(k) < 2^k$, then $k$ is a break point for $H$
  • For 2D linear separators, $k = 4$ is a break point
  • For 1D positive rays, $k = 2$ is a break point
  • For 1D positive intervals, $k = 3$ is a break point
  • For 2D convex sets, there are no break points

• If there is at least one break point for $H$, then $m_H(n)$ is polynomial in $n$