Learning From Data
Lecture 5
Training Versus Testing

The Two Questions of Learning
Theory of Generalization ($E_{in} \approx E_{out}$)
An Effective Number of Hypotheses
A Combinatorial Puzzle

M. Magdon-Ismail
CSCI 4100/6100
The Hoeffding generalization bound:

\[ E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2|\mathcal{H}|}{\delta}} \]

- \( E_{\text{in}} \): training (eg. the practice exam)
- \( E_{\text{out}} \): testing (eg. the real exam)

There is a tradeoff when picking \(|\mathcal{H}|\).
What Will The Theory of Generalization Achieve?

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2|\mathcal{H}|}{\delta}}$$

The new bound will be applicable to infinite $\mathcal{H}$. 
Why is $|\mathcal{H}|$ an Overkill

How did $|\mathcal{H}|$ come in?

Bad events

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

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

We do not know which $g$, so use a worst case union bound.

$$\mathbb{P}[\mathcal{B}_g] \leq \mathbb{P}[\text{any } \mathcal{B}_m] \leq \sum_{m=1}^{\mathcal{H}} \mathbb{P}[\mathcal{B}_m].$$

- $\mathcal{B}_m$ are events (sets of outcomes); they can overlap.
- If the $\mathcal{B}_m$ overlap, the union bound is loose.
- If many $h_m$ are similar, the $\mathcal{B}_m$ overlap.
- There are “effectively” fewer than $|\mathcal{H}|$ hypotheses.
- We can replace $|\mathcal{H}|$ by something smaller.

$|\mathcal{H}|$ fails to account for similarity between hypotheses.
Measuring the Diversity (Size) of $\mathcal{H}$

We need a way to measure the *diversity* of $\mathcal{H}$.

A simple idea:

Fix *any* set of $N$ data points.

If $\mathcal{H}$ is diverse it should be able to implement all functions

...on these $N$ points.
A Data Set Reveals the True Colors of an $\mathcal{H}$
A Data Set Reveals the True Colors of an $\mathcal{H}$

$\mathcal{H}$

$\mathcal{H}$ through the eyes of the $\mathcal{D}$
From the point of view of $\mathcal{D}$, the entire $\mathcal{H}$ is just one dichotomy.
An Effective Number of Hypotheses

If $\mathcal{H}$ is diverse it should be able to implement many dichotomys.

$|\mathcal{H}|$ only captures the maximum possible diversity of $\mathcal{H}$.

Consider an $h \in \mathcal{H}$, and a data set $x_1, \ldots, x_N$.

$h$ gives us an $N$-tuple of $\pm 1$'s:

$(h(x_1), \ldots, h(x_N))$.

A dichotomy of the inputs.

If $\mathcal{H}$ is diverse, we get many different dichotomies.
If $\mathcal{H}$ contains similar functions, we only get a few dichotomies.

The growth function quantifies this.
The Growth Function $m_{\mathcal{H}}(N)$

Define the the restriction of $\mathcal{H}$ to the inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$: 

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

(set of dichotomies induced by $\mathcal{H}$)

The Growth Function $m_{\mathcal{H}}(N)$

The largest set of dichotomies induced by $\mathcal{H}$:

$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \ldots, \mathbf{x}_N} |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)|$.

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

Can we replace $|\mathcal{H}|$ by $m_{\mathcal{H}}$, an effective number of hypotheses?

- Replacing $|\mathcal{H}|$ with $2^N$ is no help in the bound. (why?)
- We want $m_{\mathcal{H}}(N) \leq \text{poly}(N)$ to get a useful error bar.

\[ \text{(the error bar is } \sqrt{\frac{1}{2N} \ln \frac{2|\mathcal{H}|}{\delta}}) \]
Example: 2-D Perceptron Model

Cannot implement

Can implement all 8

Can implement at most 14

$m_{\mathcal{H}}(3) = 8 = 2^3$.

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

What is $m_{\mathcal{H}}(5)$?
Example: 1-D Positive Ray Model

- \( h(x) = \text{sign}(x - w_0) \)
- Consider \( N \) points.
- There are \( N + 1 \) dichotomies depending on where you put \( w_0 \).
- \( m_\mathcal{H}(N) = N + 1 \).
Example: Positive Rectangles in 2-D

$N = 4$

$N = 5$

$\mathcal{H}$ implements all dichotomies

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

some point will be inside a rectangle defined by others

$$m_\mathcal{H}(5) < 2^5$$

We have not computed $m_\mathcal{H}(5)$ – not impossible, but tricky.
Example Growth Functions

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>⋯</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-D perceptron</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>14</td>
<td>⋯</td>
<td></td>
</tr>
<tr>
<td>1-D pos. ray</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>⋯</td>
<td></td>
</tr>
<tr>
<td>2-D pos. rectangles</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>$&lt; 2^5$</td>
<td>⋯</td>
</tr>
</tbody>
</table>

- $m_H(N)$ drops below $2^N$ – there is hope for the generalization bound.
- A **break point** is any $n$ for which $m_H(n) < 2^n$. 
A Combinatorial Puzzle

A set of dichotomys

<table>
<thead>
<tr>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

Two points shattered →
Two points are *shattered*
A Combinatorial Puzzle

No pair of points is shattered
A Combinatorial Puzzle

4 dichotomies is max.

If $N = 4$ how many possible dichotomys with no 2 points shattered?

© Creator: Malik Magdon-Ismail