Learning From Data
Lecture 14
Three Learning Principles

Occam’s Razor
Sampling Bias
Data Snooping

M. Magdon-Ismail
CSCI 4100/6100
RECAP: Validation and Cross Validation

**Validation**

\[ D_1 \rightarrow D_{\text{train}}(N-K) \rightarrow D_{\text{val}}(K) \rightarrow g \rightarrow E_{\text{val}}(g) \]

**Cross Validation**

\[ D_1 \rightarrow D_{\text{train}}(N-K) \rightarrow D_{\text{val}}(K) \rightarrow g \rightarrow E_{\text{val}}(g) \]

\[ D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_N \rightarrow g_1 \rightarrow g_2 \rightarrow \cdots \rightarrow g_N \rightarrow E_{\text{cv}} \]

Model Selection

\[ H_1 \rightarrow g_1 \rightarrow \cdots \rightarrow H_N \rightarrow g_N \]

Three Learning Principles: Occam, bias, snooping
We Will Discuss . . .

• **Occam’s Razor**: pick a model carefully

• **Sampling Bias**: generate the data carefully

• **Data Snooping**: handle the data carefully
Occam’s Razor
Occam’s Razor

use a ‘razor’ to ‘trim down’

“an explanation of the data to make it as simple as possible but no simpler.”

attributed to William of Occam (14th Century) and often mistakenly to Einstein
Simpler is Better

The simplest model that fits the data is also the most plausible.

...or, beware of using complex models to fit data
What is Simpler?

simple hypothesis $h$

$\Omega(h)$

low order polynomial

hypothesis with small weights

easily described hypothesis

The equivalence:

A hypothesis set with simple hypotheses must be small.

We had a glimpse of this:

soft order constraint (smaller $H$)

$\lambda \rightarrow \minimize E_{\text{aug}}$ (favors simpler $h$).

© AML Creator: Malik Magdon-Ismail
What is Simpler?

simple hypothesis $h$

$\Omega(h)$

low order polynomial
hypothesis with small weights
easily described hypothesis

simple hypothesis set $\mathcal{H}$

$\Omega(\mathcal{H})$

$\mathcal{H}$ with small $d_{vc}$
small number of hypotheses
low entropy set
What is Simpler?

simple hypothesis \( h \)

\( \Omega(h) \)

low order polynomial
hypothesis with small weights
easily described hypothesis

simple hypothesis set \( \mathcal{H} \)

\( \Omega(\mathcal{H}) \)

\( \mathcal{H} \) with small \( d_{VC} \)
small number of hypotheses
low entropy set

The equivalence:

A hypothesis set with \textbf{simple} hypotheses \textit{must be small}

We had a glimpse of this:

soft order constraint (smaller \( \mathcal{H} \)) \( \xrightarrow{\lambda} \) minimize \( E_{aug} \) (favors simpler \( h \)).
Why is Simpler Better

Mathematically: simple curtails ability to fit noise, VC-dimension is small, and blah and blah . . .

simpler is better because you will be more “surprised” when you fit the data.

If something unlikely happens, it is very significant when it happens.

Detective Gregory: “Is there any other point to which you wish to draw my attention?”
Sherlock Holmes: “To the curious incident of the dog in the night-time.”
Detective Gregory: “The dog did nothing in the night-time.”
Sherlock Holmes: “That was the curious incident.”

— Silver Blaze, Sir Arthur Conan Doyle
A Scientific Experiment

If an experiment has no chance of falsifying a hypothesis, then the result of that experiment provides no evidence one way or the other for the hypothesis.

Scientist 3

Who provides most evidence for the hypothesis "$\rho$ is linear in $T$"?
A Scientific Experiment

If an experiment has no chance of falsifying a hypothesis, then the result of that experiment provides no evidence one way or the other for the hypothesis.

Scientist 2

Scientist 3

Who provides most evidence for the hypothesis \( \rho \) is linear in \( T \)?
A Scientific Experiment

If an experiment has no chance of falsifying a hypothesis, then the result of that experiment provides no evidence one way or the other for the hypothesis.

Scientist 1

\[ \text{resistivity } \rho \quad \text{temperature } T \]

Scientist 2

\[ \text{resistivity } \rho \quad \text{temperature } T \]

Scientist 3

\[ \text{resistivity } \rho \quad \text{temperature } T \]

Who provides most evidence for the hypothesis "\( \rho \) is linear in \( T \)"?
A Scientific Experiment

Scientist 1

\[ \text{resistivity } \rho \]
\[ \text{temperature } T \]

Scientist 2

\[ \text{resistivity } \rho \]
\[ \text{temperature } T \]

Scientist 3

\[ \text{resistivity } \rho \]
\[ \text{temperature } T \]

Who provides most evidence for the hypothesis “\( \rho \) is linear in \( T \)”?
Scientist 2 Versus Scientist 3

Axiom.

If an experiment has no chance of falsifying a hypothesis, then the result of that experiment provides no evidence one way or the other for the hypothesis.

Scientist 1

Temperature \( T \)

Resistivity \( \rho \)

Scientist 2

Temperature \( T \)

Resistivity \( \rho \)

Scientist 3

Temperature \( T \)

Resistivity \( \rho \)

very convincing

some evidence?
Scientist 1 versus Scientist 3

**Scientist 1**

\[ \text{resistivity } \rho \text{ versus temperature } T \]

*no evidence*

**Scientist 2**

\[ \text{resistivity } \rho \text{ versus temperature } T \]

**Scientist 3**

\[ \text{resistivity } \rho \text{ versus temperature } T \]

*some evidence?*
Axiom of Non-Falsifiability

**Axiom.** If an experiment has no chance of falsifying a hypothesis, then the result of that experiment provides no evidence one way or the other for the hypothesis.

**Scientist 1**

\[ \text{resistivity } \rho \]

\[ \text{temperature } T \]

*no evidence*

**Scientist 2**

\[ \text{resistivity } \rho \]

\[ \text{temperature } T \]

*very convincing*
Falsification and $m_{\mathcal{H}}(N)$

If $\mathcal{H}$ shatters $x_1, \cdots, x_N$,

- Don’t be surprised if you fit the data.
- Can’t falsify “$\mathcal{H}$ is a good set of candidate hypotheses for $f$”.

The data must have a chance to win.
Falsification and $m_{\mathcal{H}}(N)$

If $\mathcal{H}$ shatters $\mathbf{x}_1, \cdots, \mathbf{x}_N$,

- Don’t be surprised if you fit the data.
- Can’t falsify “$\mathcal{H}$ is a good set of candidate hypotheses for $f$”.

If $\mathcal{H}$ doesn’t shatter $\mathbf{x}_1, \cdots, \mathbf{x}_N$, and the target values are uniformly distributed,

$$\mathbb{P}[\text{falsification}] \geq 1 - \frac{m_{\mathcal{H}}(N)}{2^N}.$$ 

A good fit is surprising with simple $\mathcal{H}$, hence significant. You can, but didn’t falsify “$\mathcal{H}$ is a good set of candidate hypotheses for $f$”
Falsification and $m_\mathcal{H}(N)$

If $\mathcal{H}$ shatters $x_1, \cdots, x_N$,

- Don’t be surprised if you fit the data.
- Can’t falsify “$\mathcal{H}$ is a good set of candidate hypotheses for $f$”.

If $\mathcal{H}$ doesn’t shatter $x_1, \cdots, x_N$, and the target values are uniformly distributed,

$$\mathbb{P}[\text{falsification}] \geq 1 - \frac{m_\mathcal{H}(N)}{2^N}.$$  

A good fit is surprising with simple $\mathcal{H}$, hence significant. You can, but didn’t falsify “$\mathcal{H}$ is a good set of candidate hypotheses for $f$”

The data **must** have a *chance* to win.
We may opt for ‘a simpler fit than possible’, namely an imperfect fit of the data using a simple model over a perfect fit using a more complex one. The reason is that the price we pay for a perfect fit in terms of the penalty for model complexity may be too much in comparison to the benefit of the better fit.

– *Learning From Data*, Abu-Mostafa, Magdon-Ismail, Lin
Postal Scam
Home team will win the Monday Night Football Game.
A Puzzle – The Football Oracle

Saturday, Oct 13, 2012

Home team will win the Monday Night Football Game. ✓
Home team will win the Monday Night Football Game. ✓

This happens for 5 weeks in a row.
A Puzzle – The Football Oracle … on the 6th week

Call 1-900-555-5555 for winner; $50 charge applied

\[ E_{in} = 0! \]
What did the Oracle Really Do?

Single hypothesis that worked?
What did the Oracle Really Do?

<table>
<thead>
<tr>
<th></th>
<th>day 1</th>
<th>day 2</th>
<th>day 3</th>
<th>day 4</th>
<th>day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>You</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Every possible hypothesis one of which worked?
What did the Oracle Really Do?

Every possible hypothesis one of which worked?

© Creator: Malik Magdon-Ismail

Three Learning Principles: 29 / 58

Oracle is every hypothesis
What did the Oracle Really Do?

You

day 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

day 2
1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0

day 3
1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0

day 4
1

day 5
0
What did the Oracle Really Do?

| day 1 | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| day 2 | 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 |
| day 3 | 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 |
| day 4 | 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 0 |
| day 5 | 0 |
### What did the Oracle Really Do?

<table>
<thead>
<tr>
<th>Day</th>
<th>Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Day 2</td>
<td>1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Day 3</td>
<td>1 1 1 1 0 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Day 4</td>
<td>1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0</td>
</tr>
<tr>
<td>Day 5</td>
<td>1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0</td>
</tr>
</tbody>
</table>

---

**Three Learning Principles:**

Oracle is every hypothesis
What did the Oracle Really Do?

Every possible hypothesis one of which worked?
A Puzzle – The Football Oracle ... on the 6th week

Call 1-900-555-5555 for winner; $50 charge applied

$E_{in} = 0!$

Meaningless without the ‘complexity’ of the process leading to that!
We Will Discuss . . .

- **Occam’s Razor**: pick a model carefully ✓

- **Sampling Bias**: generate the data carefully

- **Data Snooping**: handle the data carefully
Sampling Bias
Tribune wanted to show off its latest technology could go earlier to press.

Telephone poll on how people voted statisticians had done their thing and were confident.
Imagine Their Surprise When . . .
Sampling Bias in Learning

If the data is sampled in a biased way, learning will produce a similarly biased outcome.

...or, make sure the training and test distributions are the same.
You cannot draw a sample from one bin and make claims about another bin.
Examples

• Kids and social media – the highlight reel.

• Taller, Fatter, Older: How Humans Have Changed in 100 Years.

• The GRE: A test that fails.
Extrapolation is Hard
Extrapolation is Hard

Three Learning Principles: 42 / 58

Dealing with Mismatch
Dealing with the Training-Test Mismatch

Think more carefully about what $f$ should look like

- Need some additional help outside the data, by choosing a good $\mathcal{H}$
- In our ranking example, account for the fat tail $\rightarrow$ hyperbola

Account for the training-test mismatch during learning

- There are methods that reweight/resample data can help
- If test data have zero representation in training, you are in trouble
- Think carefully about $f$ 😊
Puzzle - Credit Analysis

- Determine credit given salary, debt, years in residence, ... .
- Banks have lots of data
  - customer information: salary, debt, etc.
  - whether or not they defaulted on their credit.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>32 years</td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
</tr>
<tr>
<td>salary</td>
<td>40,000</td>
</tr>
<tr>
<td>debt</td>
<td>26,000</td>
</tr>
<tr>
<td>years in job</td>
<td>1 year</td>
</tr>
<tr>
<td>years at home</td>
<td>3 years</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Approve for credit?

where is the sampling bias?
Puzzle - Credit Analysis

• Determine credit given salary, debt, years in residence, . . . .
• Banks have lots of data
  – customer information: salary, debt, etc.
  – whether or not who? defaulted on their credit.

<table>
<thead>
<tr>
<th>age</th>
<th>32 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>male</td>
</tr>
<tr>
<td>salary</td>
<td>40,000</td>
</tr>
<tr>
<td>debt</td>
<td>26,000</td>
</tr>
<tr>
<td>years in job</td>
<td>1 year</td>
</tr>
<tr>
<td>years at home</td>
<td>3 years</td>
</tr>
</tbody>
</table>

Approve for credit?

only data on approved customers
We Will Discuss . . .

- **Occam’s Razor**: pick a model carefully ✓

- **Sampling Bias**: generate the data carefully ✓

- **Data Snooping**: handle the data carefully
Data Snooping
Data Snooping

If a data set has affected any step in the learning process, it cannot be fully trusted in assessing the outcome.

...or, estimate performance with a *completely* uncontaminated test set

...and, choose \( \mathcal{H} \) *before* looking at the data
Puzzle: The Buy and Hold Strategy on S&P 500 Stocks

16.2% return

Sampling Bias: didn’t buy and hold a random sample of stocks.

Snooping: Choose which stocks to hold by ‘snooping’ into the test set (the future).

© Creator: Malik Magdon-Ismail
Puzzle: The Buy and Hold Strategy on S&P 500 Stocks

Sampling Bias: didn’t buy and hold a random sample of stocks.
Snooping: Choose which stocks to hold by ‘snooping’ into the test set (the future).
Data Snooping is a **Subtle** Happy Hell

- The data looks linear, so I will use a linear model, and it worked.
  
  If the data were different and didn’t look linear, would you do something different?
Data Snooping is a **Subtle** Happy Hell

- The data looks linear, so I will use a linear model, and it worked.
  
  If the data were different and didn’t look linear, would you do something different?

- Try linear, it fails; try circles it works.
  
  If *you torture the data enough, it will confess.*

- Try linear, it works; so I don’t need to try circles.
  
  Would you have tried circles if the data were different?
Data Snooping is a Subtle Happy Hell

- The data looks linear, so I will use a linear model, and it worked.
  
  If the data were different and didn’t look linear, would you do something different?

- Try linear, it fails; try circles it works.
  
  If you torture the data enough, it will confess.

- Try linear, it works; so I don’t need to try circles.
  
  Would you have tried circles if the data were different?

- Read papers, see what others did on the data. Modify and improve on that.
  
  If the data were different, would that modify what others did and hence what you did?
  
  the data snooping can happen all at once or sequentially by different people
Data Snooping is a **Subtle** Happy Hell

- The data looks linear, so I will use a linear model, and it worked.
  
  If the data were different and didn’t look linear, would you do something different?

- Try linear, it fails; try circles it works.
  
  If you torture the data enough, it will confess.

- Try linear, it works; so I don’t need to try circles.
  
  Would you have tried circles if the data were different?

- Read papers, see what others did on the data. Modify and improve on that.
  
  If the data were different, would that modify what others did and hence what you did? the data snooping can happen all at once or sequentially by different people

- Input normalization: normalize the data, now set aside the test set.
  
  Since the test set was involved in the normalization, wouldn’t your $g$ change if the test set changed?
Account for Data Snooping

Ask yourself: “If the data were different, could/would I have done something different?”

if yes, then there is data snooping.
Account for Data Snooping

Ask yourself: “If the data were different, could/would I have done something different?”

if yes, then there is data snooping.

You must account for every choice influenced by $\mathcal{D}$.

We know how to account for the choice of $g$ from $\mathcal{H}$.
Account for Data Snooping

Ask yourself: “If the data were different, could/would I have done something different?”
if yes, then there is data snooping.

You must account for every choice influenced by $\mathcal{D}$.

We know how to account for the choice of $g$ from $\mathcal{H}$. 
Three Learning Principles

- **Occam’s Razor**: pick a model carefully ✓
  Simpler $\mathcal{H}$ is better.

- **Sampling Bias**: generate the data carefully ✓
  Make sure you train and test from the same bin.

- **Data Snooping**: handle the data carefully ✓
  Account for all choices the data influenced. Choose $\mathcal{H}$ *before* you see the data.