Three Learning Principles
Occam’s Razor
Sampling Bias
Data Snooping
Occam’s Razor

“An explanation of the data should be made as simple as possible, but no simpler.”

-- Einstein?

“entia non sunt multiplicanda praeter necessitatem”
(entities must not be multiplied beyond necessity)

-- William of Occam

“trimming down”
unnecessary explanation
The *simplest* model that fits the data is also the most *plausible*.

What does it mean to be simple?

Why is simple better?
Simple Model?

• For a hypothesis set \( \mathcal{H} \) to be simple
  • \# dichotomies it can generate is small
  • VC Dimension is small

• For a hypothesis \( h \) to be simple
  • lower order polynomial
  • smaller weights (think about the regularization)
  • easy to describe?
  • fewer number of parameters (fewer bits to describe)
Simple Model?

Connection:

A hypothesis set with *simple* hypotheses should be *simple*

Consider a hypothesis $h$ can be specified by $\ell$ bits

$\Rightarrow \mathcal{H}$ contains all such $h$

$\Rightarrow$ The size of $\mathcal{H}$ is $2^\ell$

Simple: small model complexity / VC dimension / size of hypothesis set
Why is Simple Better?

simple -> small VC dimension -> good generalization, less overfitting, ...

Simple $\mathcal{H}$
⇒ small growth function $m_{\mathcal{H}}(N)$
⇒ if data labels are generated randomly, the probability of fitting perfectly is?

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

⇒ more significant when fit really happens

Falsifiability is important!
Falsifiability

Say you want to examine whether resistivity is linear in temperature (assume no measure error)
A Classical Puzzle

Imagine you got an email before each Cardinals game for the first 5 games.

Before Game 1: “Cardinals will win” -> Cardinals wins Game 1
Before Game 2: ”Cardinals will lose” -> Cardinals loses Game 2
....

Before Game 6:
If you pay me $50 dollars, I’ll tell you whether Cardinals will win or not

It’s not falsifiable:
Imagine if this person contacts $2^{10}$ persons, split them into two groups each game $2^5$ persons will receive perfect prediction for the first 5 games
Occam’s Razor

Sampling Bias

Data Snooping
1948 US Presidential Election

• Truman vs. Dewey
• Chicago Daily Tribune decided to run a phone poll of how people voted
Truman
What happened?

One explanation: we cannot claim anything for certain.

However, there are bigger issues here...

- Phones are expensive in 1948...
- Dewey was more favored in rich populations

- Imagine you are polling from people in DC/Texas/NY to predict who will win the presidential election...
Sampling Bias

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

Make sure the training and test distributions are as close as possible...
- Example: importance weighting

Not always possible....
- If you just don’t have some region of points in training, but they appear in the testing distribution
Credit card example

- Determine whether to approve credit cards given applicants’ financial information
- Banks have lots of data:
  - Customer information
  - Whether they are good customers or not
- Are there any issues here?

<table>
<thead>
<tr>
<th></th>
<th>32 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td></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>
</tbody>
</table>

... ... 

Approve for credit?
Occam’s Razor
Sampling Bias
Data Snooping
If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.
Looking at the data
A Subtle Example

- Predict US Dollar vs. British Pound
  - $x$: the change for the previous 20 days
  - $y$: the change in the 21th day
- Normalize data
- Randomly split $D_{\text{train}}$ and $D_{\text{test}}$
- Where does snooping happen?
Reuse of a data set

• Try one model after another on the same data set, you will eventually succeed.

“If you torture the data long enough, it will confess”

• VC dimension of the total learning models
• May even include what others tried (e.g., if you read their paper...)
• p-hacking...
JELLY BEANS CAUSE ACNE!

SCIENTISTS! INVESTIGATE!

BUT WE'RE PLAYING MINECRAFT!

... Fine.

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($p > 0.05$).

THAT SETTLES THAT.

I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.

SCIENTISTS!

BUT MINECRAFT!

From xkcd, by Randall Munroe: http://xkcd.com/882
From xkcd, by Randall Munroe: http://xkcd.com/882
What should we do...

Avoid data snooping
- Strict discipline
- E.g., be **honest** and lock the test data

Account for data snooping
- Measure how much data is contaminated
- E.g., what we discussed in validation
Acknowledgements:
The slides are based on the course slides by Yaser S. Abu-Mostafa, Malik Magdon-Ismail and Hsuan-Tien Lin.