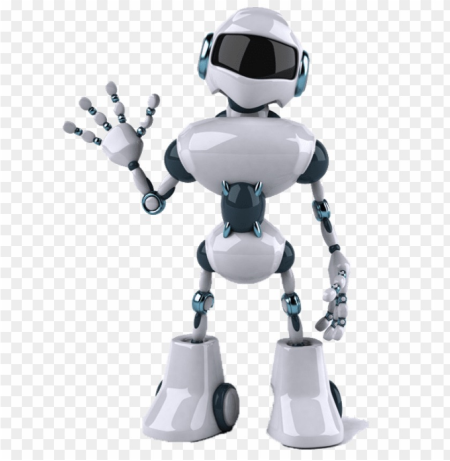
Machine Learning

CS60050

Core Learning Principles



Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 17: Three Learning Principles

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Tuesday, May 29, 2012

Outline

- Occam's Razor

- Sampling Bias

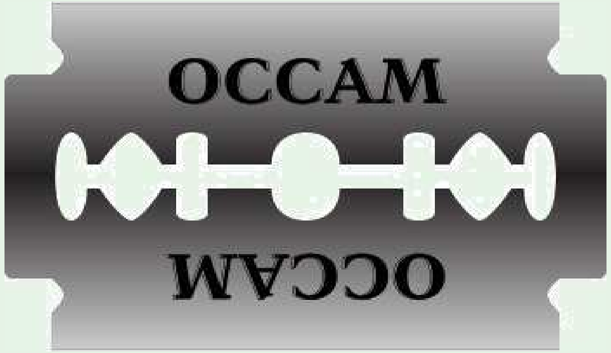
- Data Snooping

Recurring theme - simple hypotheses

A "quote" by Einstein:

> An explanation of the data should be made *as simple as possible, but no simpler*

The razor: symbolic of a principle set by William of Occam



Occam's Razor

$$\boxed{\text{The simplest model that fits the data is also the most plausible.}}$$

Two questions:

1. What does it mean for a model to be simple?

2. How do we know that simpler is better?

First question: 'simple' means?

Measures of complexity - two types: **complexity of h** and **complexity of $\mathcal{H}$**

Complexity of h: MDL, order of a polynomial

Complexity of $\mathcal{H}$: Entropy, VC dimension

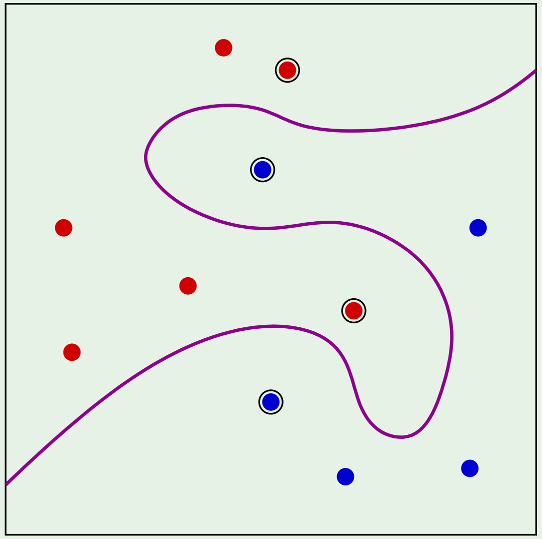
- When we think of simple, it's in terms of h
- Proofs use simple in terms of $\mathcal{H}$

and the link is . . .

counting: $\underline{\ell \text{ bits}}$ specify h $\Longrightarrow$ h is $\underline{\text{one of } 2^{\ell}}$ elements of a set $\mathcal{H}$

Real-valued parameters? **Example:** 17th order polynomial - complex and one of "many"

Exceptions? Looks complex but is one of few - **SVM**



Puzzle 1: Football oracle

```
00000000000000001111111111111111    0
00000000111111110000000011111111    1
00001111000011110000111100001111    0
00110011001100110011001100110011    1
01010101010101010101010101010101    1
           ↑
```

- Letter predicting game outcome
- Good call!
- More letters - for 5 weeks
- Perfect record!
- Want more? $50 charge ☺
- **Should you pay?**

Second question: Why is simpler better?

Better doesn't mean more elegant! It means better out-of-sample performance

The basic argument: (formal proof under different idealized conditions)

Fewer simple hypotheses than complex ones

$$m_{\mathcal{H}}(N)$$

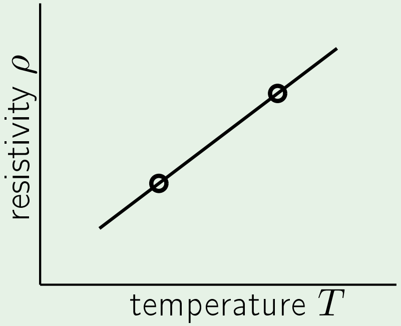
$\Rightarrow$ less likely to fit a given data set

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

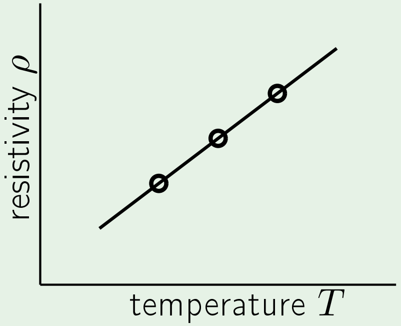
$\Rightarrow$ more significant when it happens

The postal scam: $m_{\mathcal{H}}(N) = 1$ versus 2^N

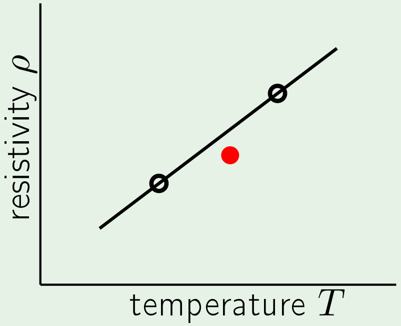
A fit that means nothing



Scientist A



Scientist B



“falsifiable”

Conductivity linear in temperature?

Two scientists conduct experiments

What evidence do A and B provide?

Outline

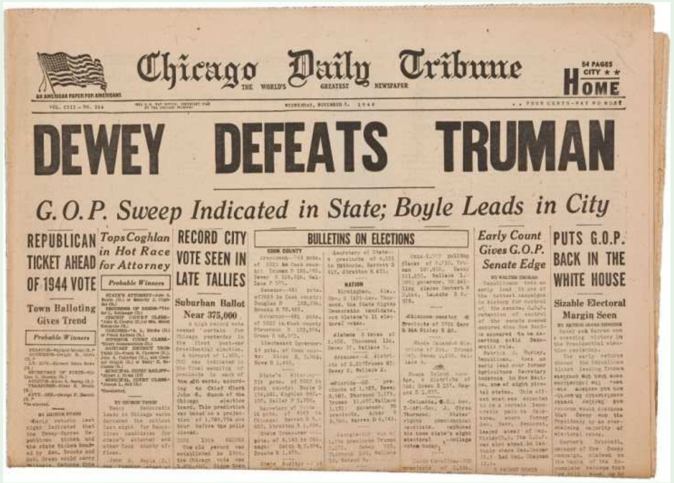
- Occam's Razor
- Sampling Bias
- Data Snooping

Puzzle 2: Presidential election

In 1948, **Truman** ran against **Dewey** in close elections

A newspaper ran a phone poll of how people voted

Dewey won the poll decisively - newspaper declared:



Chicago Daily Tribune

THE WORLD'S GREATEST NEWSPAPER

AN AMERICAN PAPER FOR AMERICANS

54 PAGES CITY ★★ HOME

DEWEY DEFEATS TRUMAN

G. O. P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Town Balloting Gives Trend

Probable Winners

Tops Coghlan in Hot Race for Attorney

Probable Winners

RECORD CITY VOTE SEEN IN LATE TALLIES

Suburban Ballot Near 375,000

BULLETINS ON ELECTIONS

COOK COUNTY

NATION

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

Sizable Electoral Margin Seen

On to the victory rally ...

... of Truman ☺

It's not δ's fault:

$$\mathbb{P}\left[\, |E_{\text{in}} - E_{\text{out}}| > \epsilon \,\right] \;\leq\; \delta$$

The bias

In 1948, phones were expensive.

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

Example: normal period in the market

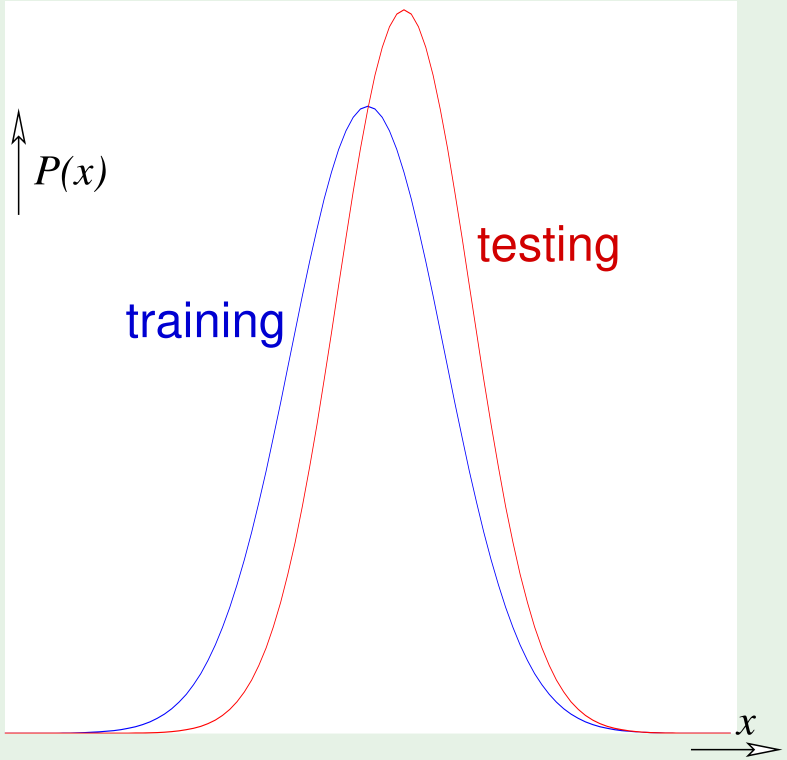
Testing: live trading in real market

Matching the distributions

Methods to match training and testing distributions

Doesn't work if:

Region has $P = 0$ in training, but $P > 0$ in testing



Puzzle 3: Credit approval

Historical records of customers

Input: information on credit application:

Target: profitable for the bank

age	23 years
gender	male
annual salary	$30,000
years in residence	1 year
years in job	1 year
current debt	$15,000
...	...

Outline

- Occam's Razor

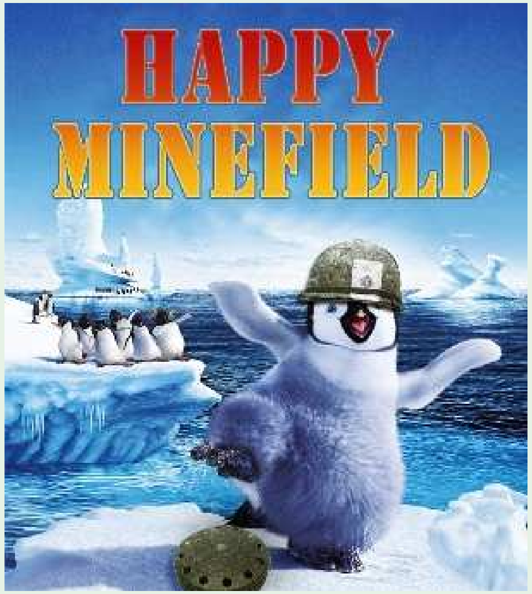
- Sampling Bias

- Data Snooping

The principle

> If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.

Most common trap for practitioners - many ways to slip ☹



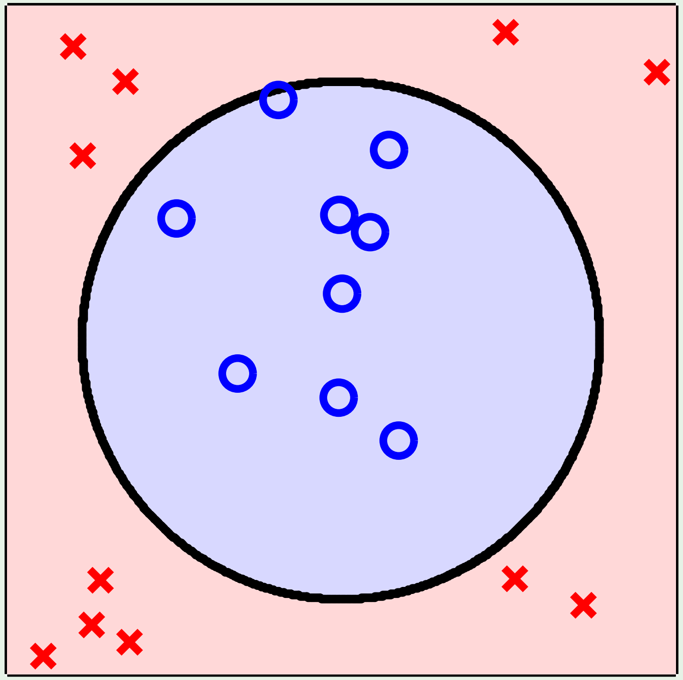
Looking at the data

Remember nonlinear transforms?

$\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$

or $\mathbf{z} = (1, x_1^2, x_2^2)$ or $\mathbf{z} = (1, x_1^2 + x_2^2)$

Snooping involves $\mathcal{D}$, not other information

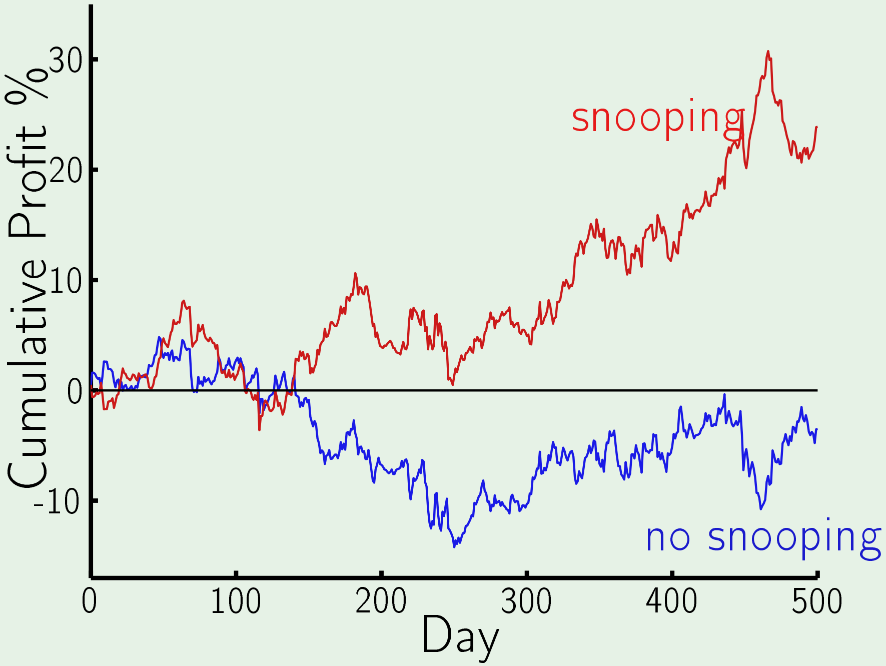


Puzzle 4: Financial forecasting

Predict US Dollar versus British Pound

Normalize data, split randomly: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$

Train on $\mathcal{D}_{\text{train}}$ only, test g on $\mathcal{D}_{\text{test}}$



$$\Delta r_{-20}, \Delta r_{-19}, \cdots, \Delta r_{-1} \longrightarrow \Delta r_0$$

Reuse of a data set

Trying one model after the other **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 model

May include what **others** tried!

Key problem: matching a *particular* data set

Two remedies

1. **Avoid** data snooping

   strict discipline

2. **Account for** data snooping

   how much data contamination

Puzzle 5: Bias via snooping

Testing long-term performance of "buy and hold" in stocks. Use **50 years** worth of data

- All currently traded companies in S&P500
- Assume you strictly followed buy and hold
- Would have made great profit!

Sampling bias caused by 'snooping'

Thank You!