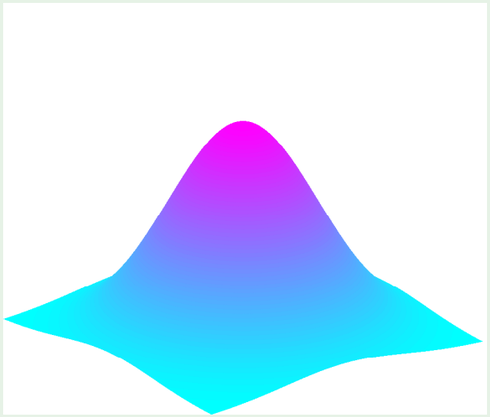
Review of Lecture 16

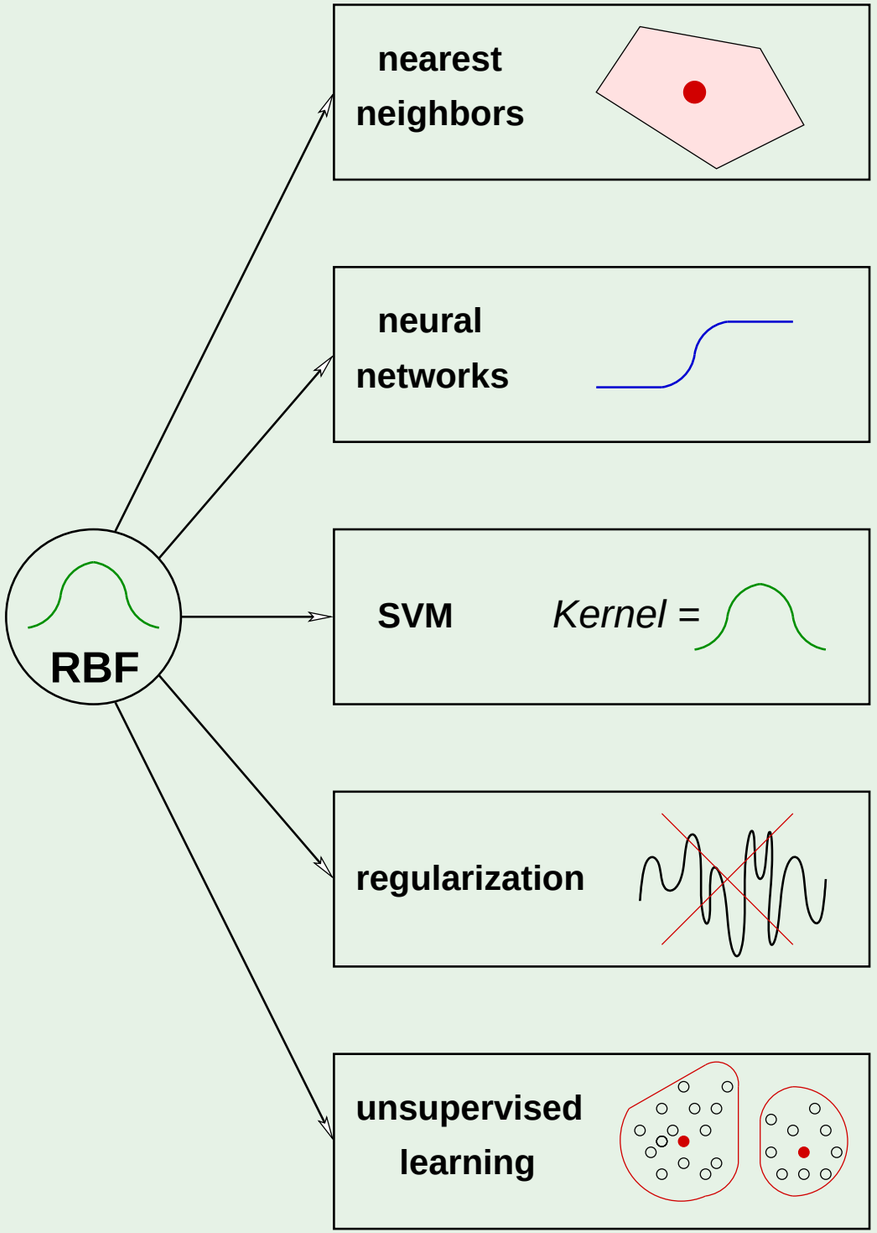
- Radial Basis Functions

$$h(\mathbf{x}) = \sum_{k=1}^{K} w_k \exp\left(-\gamma \|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right)$$



Choose $\boldsymbol{\mu}_k$'s: Lloyd's algorithm

Choose w_k's: Pseudo-inverse

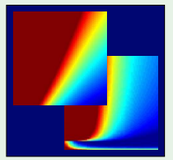


Learning From Data

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 (relates to the model - beware of fitting the data with complex models as they perform well in sample but generalize poorly)
- Sampling Bias (relates to collecting the data)
- Data Snooping (relates to handling the data)

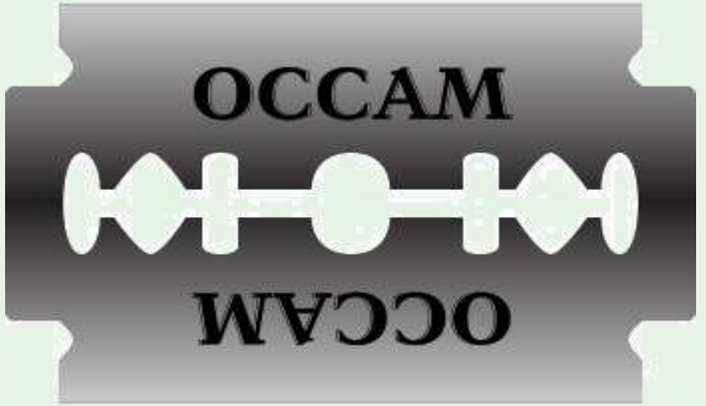
Recurring theme - simple hypotheses

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

(repeatedly trim the explanation of the data to the bare minimum which is still consistent with the data)



Occam's Razor

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

(plausible = the most likely to be true)

(i.e. when you use a simpler model that still fits the data, on average you will be getting a better performance)

Two questions:

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

2. How do we know that simpler is better?

In general, 'simpler model' refers to having 'fewer assumptions' (about the data and real behavior of unseen data)

First question: 'simple' means?

Measures of complexity - two types: complexity of h (complexity of an object) and complexity of $\mathcal{H}$ (complexity of a set of objects)

Complexity of h: MDL (minimum description length), 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 ...

(in the example of MDL)

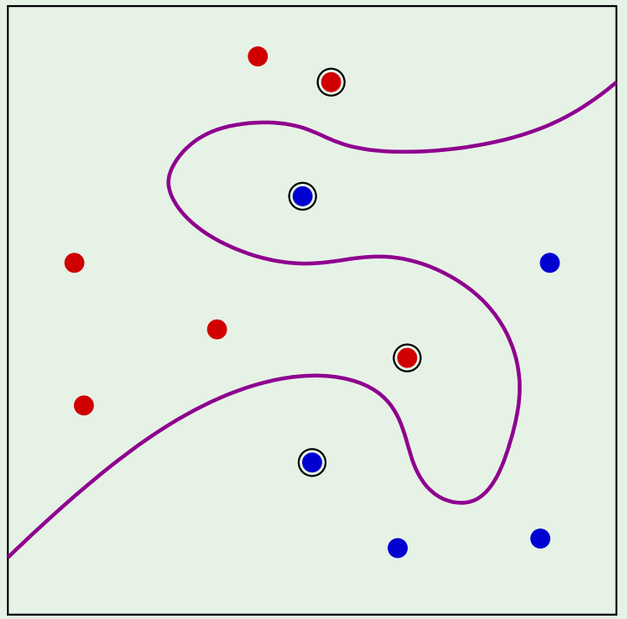
counting: ℓ bits specify h $\Longrightarrow$ h is one of 2^ℓ elements of a set $\mathcal{H}$

Something is being complex in its own right when it is one of many elements, while it is simple in its own right when it is one of few.

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

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

(Looks complex but it is defined by a small number of SVs so
it is one of few and we do not have to pay the full price for it being complex)



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?**

The person is sending letters to 32 people, half the home team will win (1) and half they will lose (0). He repeats this with the people who he sent the right result to over a period of 5 weeks. This leaves him with one person who got the correct results each time.

The connection to learning is that you thought the prediction ability was great as you only saw your letters. So one hypothesis got it right perfectly, but the hypothesis set is very complex and so the prediction value is meaningless.

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)

There are Fewer simple hypotheses than complex ones

$m_{\mathcal{H}}(N)$ (maximum number of dichotomies H could generate)

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

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

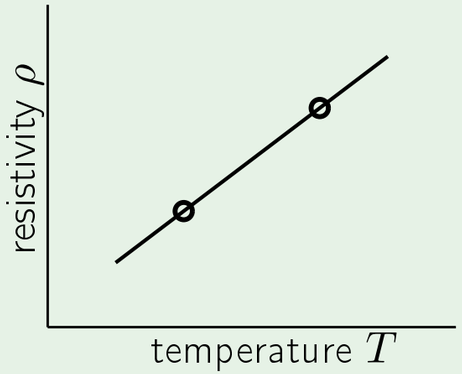
i.e. in entropy, you get more information when an unlikely event (with small p) occurs since log(1/p)

$\Rightarrow$ more significant when it happens

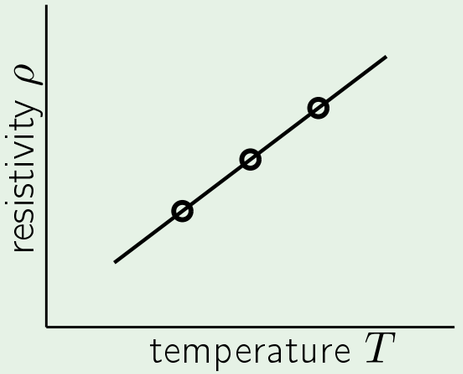
The postal scam: $m_{\mathcal{H}}(N) = 1$ versus 2^N

The growth function from your point of view receiving the letters is 1 since we have the perfect hypothesis. There is only one possible dichotomy for any N samples (extremely simple model) and its the one that predicts the results of the game (we assigned a lot of value to that because it was unlikely to happen). In reality the hypothesis is completely random, so it could generate any dichotomy by chance, so the growth function is equal to the number of possible dichotomies 2^N. As a result, the event that they predicted it right was certain to happen, so when it happens it was meaningless.

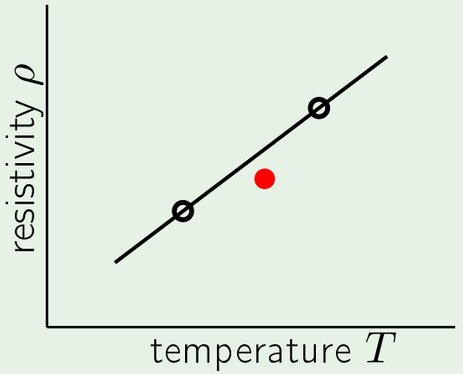
A fit that means nothing



Scientist A



Scientist B



"falsifiable"

Conductivity linear in temperature?

Two scientists conduct experiments

What evidence do A and B provide?

If your data has no chance of falsifying the assertion, then by the same token it does not provide any evidence for that assertion (axiom of non falsifiability). Linear model is way to complex for A for the size of the dataset to be able to generalize at all, so no evidence there.

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:

(decisively meaning that he won above the error bar
- the probability of the opposite being true is vanishingly small)

Chicago Daily Tribune

THE WORLD'S GREATEST NEWSPAPER

64 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

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.**

(data needs to be representative of what we want to model - sampling bias makes you vulnerable to the part of the population/input space you did not touch)

Example: normal period in the market

Testing: live trading in real market

Matching the distributions

One of the assumptions in Hoeffding is that the testing and training samples are picked from the same distribution - sampling bias violates this. Matching the probability distributions of the training and testing data is a way of countering this bias, done either by giving different weights to/scaling the training data or re-sampling the training data, such that we get another training set which is as if it were pulled from the other (correct) distribution.

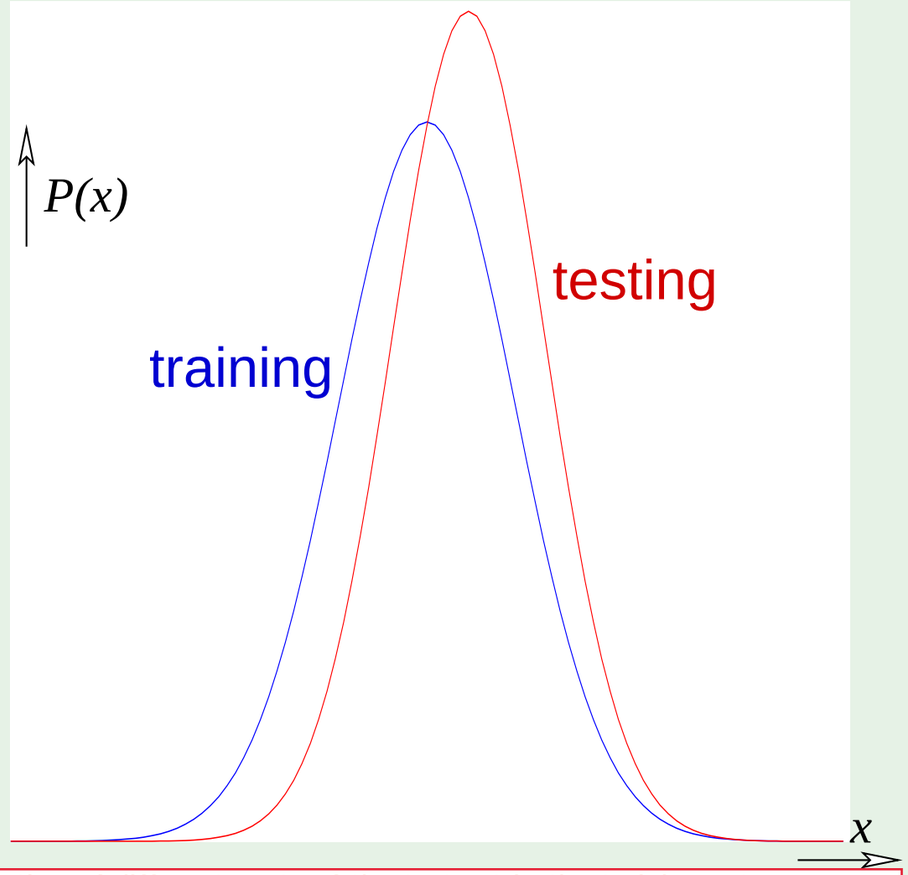
Methods to match training and testing distributions

Doesn't work if:

There is a region in the input space where

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

Like in the phone example, P=0 that people did not have a phone in the sample, but P>0 in the general population; however, this cannot obviously be fixed in terms of matching. In these cases where we cannot cure the bias we must admit we do not know how the system will perform in the parts of the input space not sampled (so warn against using the system in that particular sub-domain).



We see that there is under and over representation of different parts of the test set in the training set, so we scale/change the emphasis of the examples such that we compensate for the discrepancy and make it as if we are coming from the test set (there are also re-sample methods to achieve the same effect). In the Netflix example, where we do not necessarily have these distributions, we look at the input space in terms of coordinates. For example, looking at the balance between heavy users (rated lots of movies) and light users, in the test and training set we try to have equivalent distributions as far as the number of rating are concerned. We then this for other coordinates and try to match those coordinates.

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
...	...

Where is the sampling bias? Well, we are using the historical data of customers that were approved, as these are the ones we have credit behavior on. We have no data on those who were rejected from getting credit. A new applicant may belong to the population that was never part of the training sample. However, banks like to be aggressive in handing out credit, since they can make a lot of profit from borderline cases. As a result, the boundary will be represented by the customers that were already accepted since the bank will have made some mistakes in accepting those people, so it is likely that the section of population missing from the training data will be deep on one side of the boundary (interior points) - with making some mistakes, we essentially find the support vectors of the boundary, so the interior points matter much less. This system, even with sampling bias, does pretty well modeling future customers. Note it would be possible to correct the bias: if your bank rejects someone, then they go to another bank, get accepted and lose them money, you know your decision was correct you can determine how much of an impact the sampling bias made.

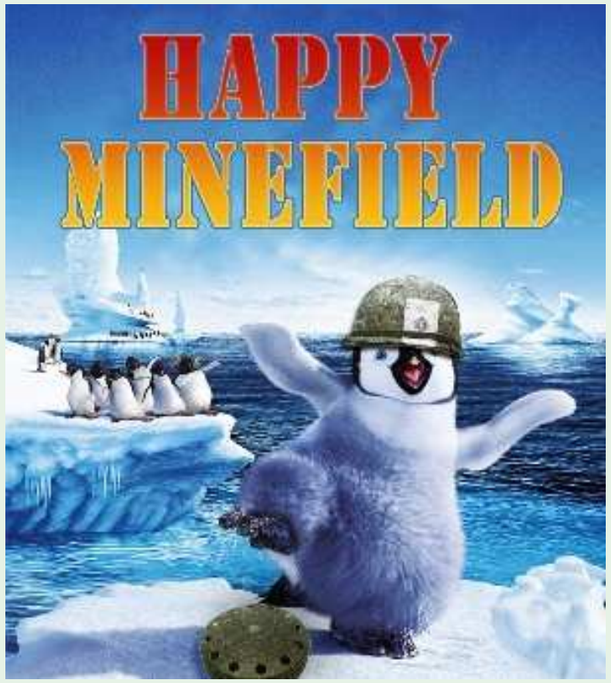
Outline

- 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 ☹



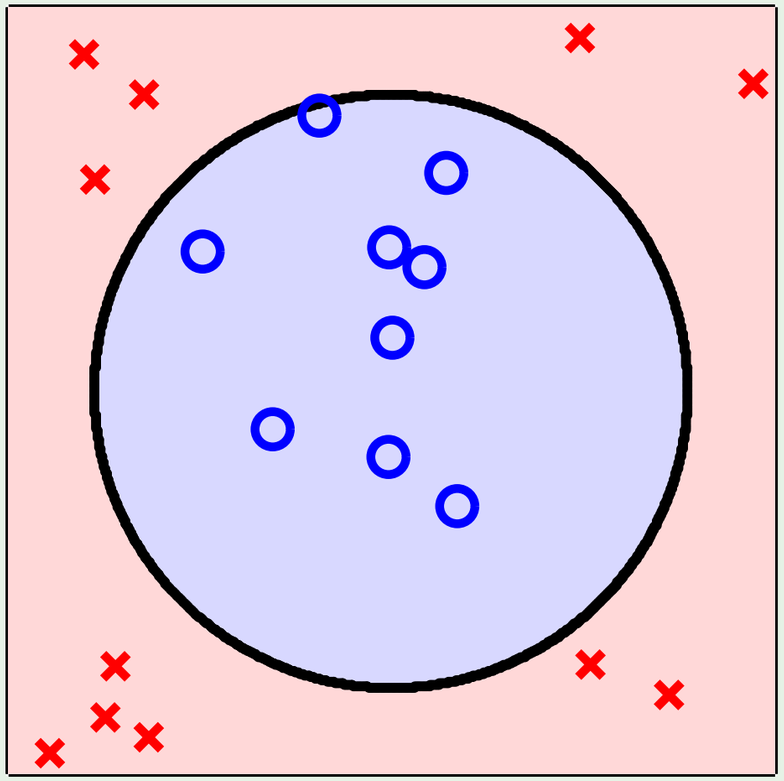
Looking at the data

Remember nonlinear transforms?

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

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



You are acting as a learning algorithm in your own right and narrowing down H free of charge - the problem is you are only charging for a VC dimension of 2 (which is only what the last part of the learning cost you). Looking at the dataset makes you vulnerable to designing your model or your choices in the learning based on the idiosyncrasies of that dataset. Therefore we may perform well on this dataset but we would not perform well on another independently generated dataset from the same distribution (our out-of-sample). On the other hand, it is important to look at all other information related to the target function and input space except for the realization of the dataset that we are going to use for training, unless we are going to charge accordingly.

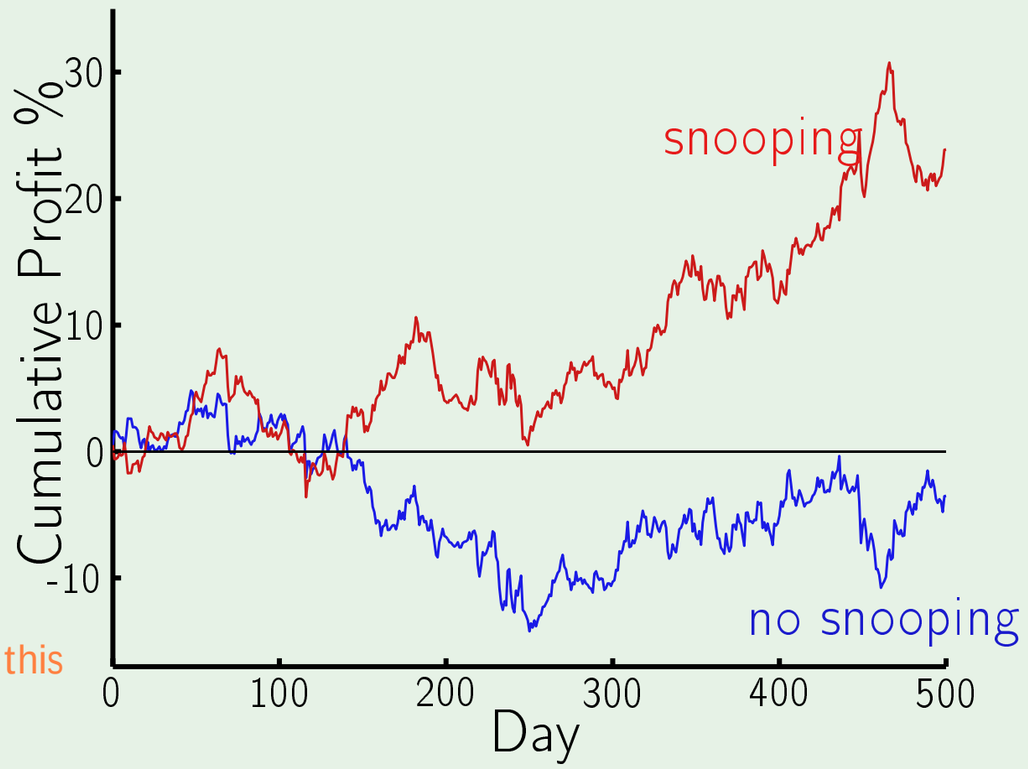
For example, we can ask for the number of inputs, the range of the inputs, how the inputs were measured, if they are physically correlated, if the customer knows of any properties we can apply, is it monotonic in some sense etc. This is all completely valid and vital in order to zoom in correctly, we are just using the properties of the target function and input space proper, so improving the chances of selecting the correct model.

Puzzle 4: Financial forecasting

Predict US Dollar versus British Pound

Normalize data, split randomly: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$
(zero mean, unit var)

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



$$\underbrace{\Delta r_{-20}, \Delta r_{-19}, \cdots, \Delta r_{-1}}_{\text{input}} \longrightarrow \underbrace{\Delta r_0}_{\text{output}}$$

delta(r) is the change in rate each day

The snooping occurred when we normalized the data before splitting into Dtrain and Dtest - thus we took into consideration the mean and variance of the test set. Instead, you should split the data, take the training set and take the normalization. The mu and sigma^2 which achieved this normalization for Dtrain should then be frozen and applied to Dtest so that they live in the same range of values. This slight snooping made an enormous difference to the cumulative profit. The parameters for normalization should be extracted exclusively from the training set. To avoid this, consider what information is available to you if you have no access to the test set.

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

Trying multiple models on the same dataset increases the VC dimension without realizing - the final model we use to learn is the union of all of the above, it just happens that some of them were rejected by the learning algorithm. The data snooping is the use of the failure of the previous model to direct you to the choice of the new model (without accounting for the VC dimension of having done that). It is like the previous model looked at the data and made a decision, and we did not charge for it. In accounting for it, we mean that you consider the effective VC dimension of the entire model and you have to get/use more data for the entire model in order to generalize.

VC dimension of the **total** learning model

May include what **others** tried!

You did not look at the data, but you used something which was affected by the data through the work of others

Key problem: matching a *particular* data set

You are trying to match a particular dataset too well: you keep trying models and after a while you know exactly what to do for this particular data set and its particular idiosyncrasies (i.e. you start fitting the noise). If another dataset is independently generated from the same distribution you will perform poorly on it.

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

The bias is from the 'currently traded stocks' which excludes all the companies which were there and took a dive in the stock market, so we are at an unfair advantage as we base our models only on the stocks which turned out to be successful. Some people also consider this as 'snooping' as we are effectively looking at the future (the current day) from the past and we are told which stocks will still be traded, which is not allowed.

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

Sampling bias caused by 'snooping'