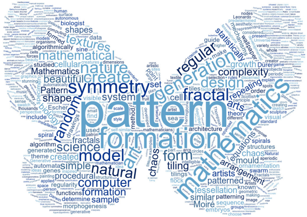
Advanced Mathematical Perspectives 1

Lecture 19: Measuring Success



Matthew Roughan
<matthew.roughan@adelaide.edu.au>
www.maths.adelaide.edu.au/matthew.roughan/notes/AMP1/

School of Mathematical Sciences,
University of Adelaide

THE UNIVERSITY of ADELAIDE

SUB CRUCE LUMEN

ACEMS

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Measure what is measurable,
and make measurable what is not so."
Galilei, Galileo (1564 - 1642)

To measure is to know.
Lord Kelvin

Section 1

Measuring Success

How do you tell if you have a good model?

Does your model fit the data?

- Find a *metric* relevant to *your* problem, to measure how close your model is to reality
  - map a big complex set of “stuff” to a single *number*
  - the bigger the number, the further our data is from the model
- Find a *test* to decide if your metric is good or bad
  - assign a meaningful “scale” to the metric
  - provide an interpretation of the result

Metrics

- Model error, *e.g.*,
  - RMS (Root-Mean-Squared) error between model and data
- Application metrics: *e.g.*,
  - how well does the model work in my application?
- Qualitative measures: *e.g.*,
  - image compression quality measured by people making qualitative assessments

RMS error

$$RMS(\mathbf{x}, \hat{\mathbf{x}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}$$

- Seems simple but:
  - you have to choose the right feature to measure
  - doesn't have to be error in obvious terms, *e.g.*, could do Fourier transform and measure errors in frequency domain
- Sometimes big RMS errors are not very obvious (to people), and sometimes small RMS errors are very obvious
- Stochastic data won't ever fit deterministic expectations precisely

Tests

Assume we have a good metric for measuring our model, how do we judge the size of that error as good or bad?

Statistical tests!

- I will present only 1 example, and I won't go into detail
- See our statistical courses for a proper explanation

Simple hypothesis test

A hypothesis test tests between two hypotheses: the *null*-hypothesis $\mathcal{H}_0$ and an *alternate*-hypothesis $\mathcal{H}_a$

For example

- $\mathcal{H}_0$ the model is valid
- $\mathcal{H}_a$ the model is invalid

Plan:

1. form a metric or *test-statistic*
2. choose a *statistical significance* α, e.g., 5% (or 0.05)
3. using the distribution predicted by the null-hypothesis, work out a *p-value* for your statistic *the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed*.
4. if $p \leq \alpha$ reject the null-hypothesis

Example hypothesis test: χ^2-test

- $\mathcal{H}_0$: data matches the distribution predicted by the model
- $\mathcal{H}_a$: data doesn't fit the model

Test statistic: break distributions into n bins, and count number of values in each bin

$$t = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where

$$\begin{aligned} O_i &= \text{observed number in bin } i \\ E_i &= \text{expected number in bin } i \end{aligned}$$

p-value comes from χ^2 distribution, hence the name

Example χ^2-test

Test whether a 6-side dice is *fair*, *i.e.*, the probability of each possible result is $1/6$

1. roll the dice M times, and record

$$O_i = \text{ number of times we see } i$$

2. $E_i = M/6$
   - The value of M needs to be large enough that E_i isn't too small
3. Calculate t as above
4. Calculate p from a χ^2 distribution (with $5 = n - 1$ degrees of freedom because there are $n = 6$ bins)

Tricks and Tips

- Sometimes we can calculate E_i from analysis; sometimes we have to simulate the model to get it
- Interpretation of null
  - we can't say the null-hypothesis is "true", can only "retain" it
  - we could have $p > \alpha$ for two reasons
    1. the null is true
    2. you don't have enough data
  - like "not guilty" in a trial (insufficient evidence to prove guilt)
- There are well-known variants of these tests for most simple situations you will come across, so it's often just a matter of finding the right software package to solve your problem (most statisticians seem to prefer "R")
- More generally, we often come up with *null-models* against which we can compare our data to see if a more complex model is needed

Other important issues

- A simple model might not fit a particular dataset as well, but may be more "universal" model
  - a model with more parameters should be able to fit a dataset better, but might over-fit

    > I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.
    >
    > *Enrico Fermi*

  - complex models
    - more difficult to estimate parameters
    - harder to interpret results
  - the map is not the territory – Bonini's paradox

Bonini's paradox

> The map is not the territory
>
> *Alfred Korzybski*

- Start with a rough map of Adelaide
  - it's pretty bad at representing routes, for instance
- So you make it better, more detailed
  - of course it has to be bigger to show the new detail
  - but it is still missing some details
- So you make it more detailed again, and so on
- The ultimate map would be a 1:1 representation of Adelaide
  - perfect detail
  - perfectly useless

Bonini's paradox expresses the fact that as a model becomes more detailed and accurate it becomes less useful.

Summary

- You must test models
- All of statistics is at your disposal
- It's not trivial to get right

The Map is Not the Territory

> .. In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.
>
> *Jorge Luis Borges [Bor75]*

The Map is Not the Territory

> "What a useful thing a pocket-map is!" I remarked.
>
> "That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"
>
> "About six inches to the mile."
>
> "Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"
>
> "Have you used it much?" I enquired.
>
> "It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."
>
> *Lewis Carroll's Sylvie and Bruno Concluded, 1895*

Further reading I

- Jorge Luis Borges, *A universal history of infamy*, ch. On Rigor in Science, Penguin Books, 1975, (translated by Norman Thomas de Giovanni).

- Umberto Eco, *How to travel with a salmon & other essays*, ch. On the Impossibility of Drawing a Map of the Empire on a Scale of 1 to 1, Houghton Mifflin Harcourt, 1995.