In the data: interdisciplinary modes of machine learning

Adrian Mackenzie

Sociology Department, Lancaster University

a.mackenzie@lancaster.ac.uk

Abstract

This paper explores ways of thinking about digital data that lie somewhere between blithe faith and critical dismissal. It focuses on the machine learning, an increasingly prevalent bundle of techniques and approaches that lies at the centre of contemporary data processing. Machine learning is used to program computers to find patterns, associations, and correlations, to classify events and make predictions on a large scale. As a set of techniques for classifying and predicting, machine learning lies close to centre of calculation in social network media, finance markets, robotics, and contemporary sciences such as genomics and epidemiology. This paper will discuss who is doing machine learning, who could do machine learning, and how they might do it differently.

The problem

- if you had all the data, what then?
  - this is the horizon of the contemporary lifeworld. Someone, perhaps not us, not yet, has all the data. Probably they are a business like Walmart, amazon, google or Facebook. They might be a State, like the United States or the United Kingdom. Perhaps we will have the data. There is, after all, a lot of data already available, and more all the time. It may be that civil society, non-governmental groups like Wikileaks, or various forms of massive data leak will render valuable data available. Or maybe just the increasingly powerful techniques of record linkage will allow comprehensive data streams to appear.
  - if we had it, what would we do?
- pattern finding vs models?
  - Abbott - inaccurate opposition between linear and pattern
- finding the function that generated the data

The data

if you had all the data, what then?

I want to present several vignettes that illustrate what people do when they have all the data. They nearly always do what is variously called data mining, knowledge discovery or currently, *machine learning*

infrastructural re-dimensioning

- cancer genomes + google compute
- begin with someone/something that does have all the data – Google and certain genomic scientists work together on cancer genomes;
- Let me illustrate what I mean by infrastructural re-dimensioning:

  Google Compute Engine + Institute of Systems Biology, June 2012

  The world's 3rd largest supercomputer *learns associations* between genomic features' (Anthony 2012)

  The Google Compute Engine, a globally-distributed ensemble of computers, was briefly turned over to exploration of cancer genomics during 2012, and publicly demonstrated during the annual Google I/O conference. Midway through the demonstration, visualized as a circular genome, the speaker, Urs Hölzle, Senior Vice President of Infrastructure at Google 'then went even further and scaled the application to run on 600,000 cores across Google's global data centers' (GoogleInc. 2012). The world's '3rd largest supercomputer', as it was called by TechCrunch, a prominent technology blog, 'learns associations between genomic features' (Anthony 2012).

- the scientists are from Institute for Systems Biology (ISB), University of Washington. Its a fairly well-established and long-standing icon of big data science, with some Nobel Prizes, etc, associated with it. [SHOW Image of building]
- Urs Hölzle, the Vice-President of Infrastructure at Google shows the fruits of their collaboration to an audience of developers at Google I/O in San Francisco, 2012. His title is significant – VP of *Infrastructure*. [SHOW the video]

- On the one hand, the scientists at ISB have the data. They are somewhat like us – they are academics. They are different to us in that they have the best data possible on cancer genomes. Their data reaches down into clinics, it comes from the best scientific instruments money can buy, and they benefit from PhDs, programmers, engineers, disk drives and an assortment of other services at their beck and call.

- Yet, they need or want google. Or maybe Google wants them for something. What the Google I/O conference shows is the availability of flexible instrastructure that can be quickly, rapidly made available on demand. It can shift orders of magnitude in size in seconds. Not many infrastructures can do that. But even if it can, why are such infrastructures needed? It is as if Oxford St in London could suddenly switch between a pedestrian mall and an 8 lane motorway. What form of life or sociality would require that mutability?

- in this case, what the scientists want to do is to sift through massive numbers of associations between different parts of a genome to identify patterns and processes associated with mutagenesis. The algorithm they use is called RF-ACE, and an implementation in the R statistical programming language is freely available. While it is state-of-the-art, this algorithm is a typical machine learning algorithm. It is an example of unsupervised learning algorithm that tries to *find structure* in a data-set without any prior knowledge about the data.

data is wide, dirty and mixed: 3 million features

- In the Google Compute case, it is hard to see what machine learning is doing. We hear it is discovering associations, but associations between what? For present purposes, the issue here is the shape of the date.

- The shape of datasets is widely discussed in machine learning, and shape perhaps more than the size (as in 'big') matters to who algorithms go about finding structures or patterns. It is shape too that occasions much of the infrastructural re-dimensioning we have just seen displayed.

- The shape of contemporary datasets is sometimes described as 'wide, dirty and mixed'. Each of these terms carries interesting connotations, and a figurative analysis of data talk would be quite useful. I'm not going to offer that here, but just point to some examples of what they mean.

- Statistics textbooks and statistical practice are full of datasets like this one: classic dataset: 'iris' R.A. Fisher (1936)

```
## [1] 5
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrl}
##   \hline
##  & Sepal.Length & Sepal.Width & Petal.Length & Petal.Width & Species \\
##   \hline
## 1 & 5.10 & 3.50 & 1.40 & 0.20 & setosa \\
##   2 & 4.90 & 3.00 & 1.40 & 0.20 & setosa \\
##   3 & 4.70 & 3.20 & 1.30 & 0.20 & setosa \\
##   4 & 4.60 & 3.10 & 1.50 & 0.20 & setosa \\
##   5 & 5.00 & 3.60 & 1.40 & 0.20 & setosa \\
##    \hline
## \end{tabular}
## \end{table}
```

- Fisher's widely cited 'iris' data is often used to demonstrate various statistical techniques.
- In machine learning textbooks such as (Hastie, Tibshirani, and Friedman 2009), iris is mentioned, but datasets are commonly much wider. That is, they have many more columns. For instance, in the late 1990s, when biomedical data from genechips or microarrays started to become widely available, machine learning techniques were applied to their analysis. A typical 'wide dataset' would be Golub's c.1999 'blood pressure' dataset

```
## columns in Golub dataset: 102
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrrrl}
##   \hline
##  & \begin{sideways} RS2495368 \end{sideways} & \begin{sideways} RS2292857 \end{sideways}
##   \hline
## 1 &   0 &   0 &   0 &   0 &   0 & 171.19 & normal \\
##   2 &   0 &   0 &   0 &  &   0 & 179.25 & normal \\
##   3 &   0 &   0 &   0 &   1 &   0 & 188.23 & normal \\
##   4 &   0 &   0 &   0 &   0 &   0 & 175.44 & high \\
##   5 &   0 &   0 &   0 &   0 &   0 & 177.98 & normal \\
##    \hline
## \end{tabular}
## \end{table}
```

- the Golub blood pressure data from 1999 has over 100 columns rather than 5. This is not especially wide, but radically changes the structures that might be found in the data.

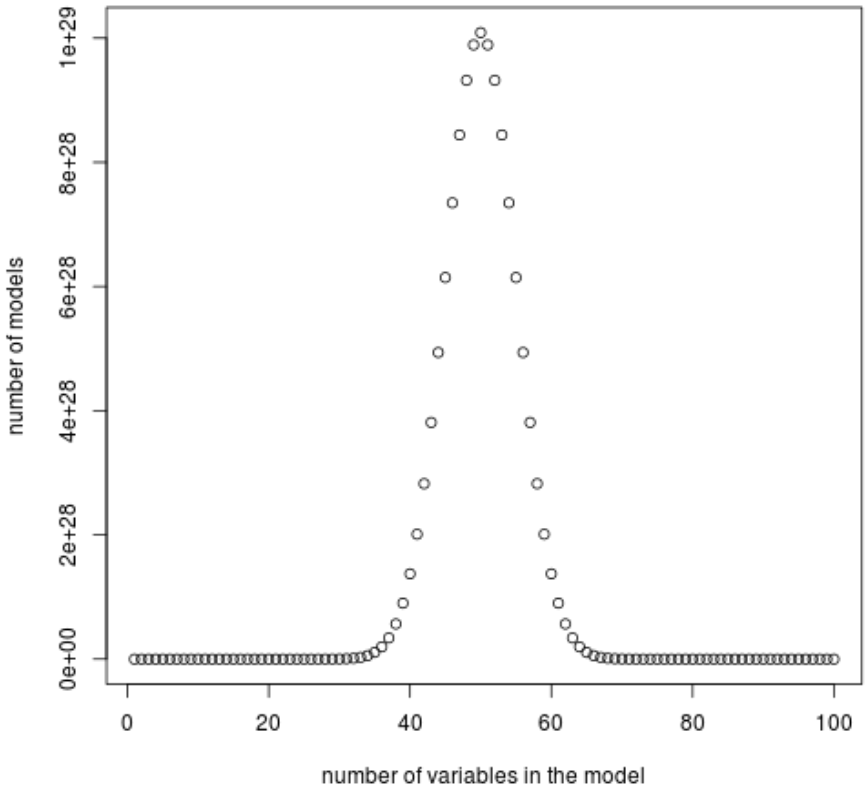

Figure 1: plot of chunk name

- as we can see, the number of different models that you might build to look at patterns between blood pressure, height and this genetic data very quickly becomes huge when there are numerous variables or 'features' in the data.
- but this kind of measurement data is quickly dwarfed by text or image-derived data. Even relatively tightly defined data forms such as URLS. Machine learning researchers spend quite a lot of time trying to build predictive models for mundane forms of data such as URLs. They can involve millions of variables, that could be interacting in countless combinations.
- Suspicious URLS dataset (Ma et al. 2009)

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{lr}
##   \hline
## feature & value \\
##   \hline
## suspicious & -1.00 \\
##   4 & 0.08 \\
##   5 & 0.12 \\
##   6 & 0.12 \\
##   11 & 0.43 \\
##   155211 & 1.00 \\
##   155212 & 1.00 \\
##   155213 & 1.00 \\
##   945789 & 1.00 \\
##   1988571 & 1.00 \\
##   2139257 & 1.00 \\
##   2987739 & 1.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

- The point here is that finding structure in data means working out which bits of data are related to which other bit. Typically a row of data in a dataset relates to one thing. The thing might be a person, an event, a sample, a place, but machine learning is rather indifferent to that. What matters is finding, amidst the many possible combinations of features or variables those combinations that carry some weight, that pattern the other data, or that stand out.
- While the data we have seen is wide, it is not necessarily very mixed or dirty. Mixed means that the types of data are varied. A mixed dataset includes numbers, categories, text, and so forth. Dirty data, as you can

imagine, describes data lacks consistency, that might have holes in it or is somehow noisy.

- text data and shaping the world
  - James Lin, Twitter
  - e-discovery
  - leximancer
- signals – image and sound
  - driverless car - Thrum - winner of the Darpa challenge
  - google cat
  - Eulerian video

The people

- I haven't yet talked about machine learning actually seeks structure in data, but I keep talking about people, companies and scientists doing machine learning without saying much about who they are. Who does machine learning?
- the field lies somewhere between computer science and statistics. For instance, one of the main textbooks in the field (Hastie, Tibshirani, and Friedman 2009), written by three very well-known statisticians from Stanford University. On the other hand, the Stanford computer scientist Andrew Ng is one of the most well-known figures in the field. His Youtube lectures on machine learning CS229, a postgraduate course, show viewing figures peaking at 500,000.
- So the quintessential machine learning expert is a Stanford PhD, who goes on to work in Silicon Valley, Wall Street or for some US government agency.
  - Ng: stanford Phds
  - programming collective intelligence
- the wonderful people
  - Hilary Mason
  - Cath O'Neill & Rachel whats her name
  - Heather Arthur
- the social scientists
  - Gary King
  - machine learning for hackers
  - manovich – cultural analytics
  - Savage - descriptive assemblage

What to do

- Machine learning is a hard field to get into in some ways. On the one hand, it is becoming enormously available and increasingly as Thrift would say, part of the contemporary time-space signature. Myriad mundane examples could be given. On the other hand, to get into it, to occupy the fluxing dimensionality of the data is technically difficult, and conceptually challenging. It takes very significant investments in time and attention (for instance, going through 27 hours of Youtube lectures with lots of equations written on blackboards is not trivial).
- I am suggesting that getting into it analytically and practically might be worthwhile, at least for some people for several reasons
  - its ubiquity in science, business and government renders it a key contemporary control practice, that differs in powerful yet subtle ways from existing ways of knowing, predicting, anticipating or controlling
  - its mobility as a cluster of methods has strongly performative effects – parts of the world are heavily re-configured through it
  - the mobility of its methods has no clear boundaries. The spread of machine learning into social sciences and humanities has started. And much of our research already depends on it anyway (google search, etc). The question is how we move in relation to machine learning.
  - like all sciences and technologies, machine learning contains zones of slippage, inconsistency or friction where things can happen. While it might not be us who occupy those zones most easily, in making sense of machine learning all the way down, and pointing to the structures, processes or relations at play, we help free up the possibilities of gaming the models.
- Jaron Lanier
- occupydata
- animation
- text/coding/reproducibility

References

Anthony, Sebastian. 2012. "Google Compute Engine: For $2 million/day, your company can run the third fastest supercomputer in the world \textbar ExtremeTech." http://www.extremetech.com/extreme/131962-google-compute-engine-for-2-millionday-your-company-can-run-the-third-fastest-supercomputer-in-the-world.

GoogleInc. 2012. "Behind the Compute Engine demo at Google I/O 2012 Keynote - Google Compute Engine — Google Developers." https://developers.google.com/compute/io.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer.

Ma, Justin, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2009. "Identifying suspicious URLs: an application of large-scale online learning." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 681–688.