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*

- cancer genomes + google compute
  - begin with someone/something that does have all the data -- Google and certain genomic scientists work together on cancer genomes
  - 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.

- 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 keep talking about people, companies and scientists without saying much about who they are. Can we easily identify who does machine learning? - the computer scientists - Ng: stanford Phds - programming collective intelligence

- the wonderful people
  - Hilary
  - 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

- Jaron Lanier
- occupydata
- animation
- text/coding/reproducibility

References