Strength in diversity
-
the advance of data analysis

David J. Hand
Imperial College London

d.j.hand@imperial.ac.uk

September 2004

Data analysis is

- ***the science of discovery in data***
- ***of processing data to extract evidence***
- ***so that we can make informed decisions***

The theory and methods, not of any particular discipline itself, but of how to find things out

- **a new technological discipline**
- **a merger of several existing disciplines**

statistics, computer science, pattern recognition, AI, machine learning, ...

Parallel work with different emphases ⇒ tensions, benefits and a potential for synergy

We (data analysts) live in very exciting times
We live in the *most exciting* of times

Why so exciting? ***The computer***

Replace *months of error-prone hand calculation* by
split second production accurate results

Changing the role of the data analyst:
- from concern with arithmetic manipulation
- to concern with interpretation and meaning

Two impacts
- can easily do what we did before
- can do entirely new things

Evolution of data analysis

Origins - in the mists of time

e.g.1: King David I of Scotland, 1150 AD, defined the inch as the average of the width of the thumbs of a big man, a medium man, and a small man measured at the base of the thumbnail.

e.g.2: The rood is an old British unit of length connected to the foot.

The surveyor should stand by a church door on Sunday. When the service ends, he should “bid sixteen men to stop, tall ones and short ones, as they happen to come out ...”. ...These chosen men should be made to stand in line with “their left feet one behind the other”. The sum of the sixteen actual left feet constitutes the length of “the right and lawful rood” and the sixteenth of it constitutes “the right and lawful foot”.

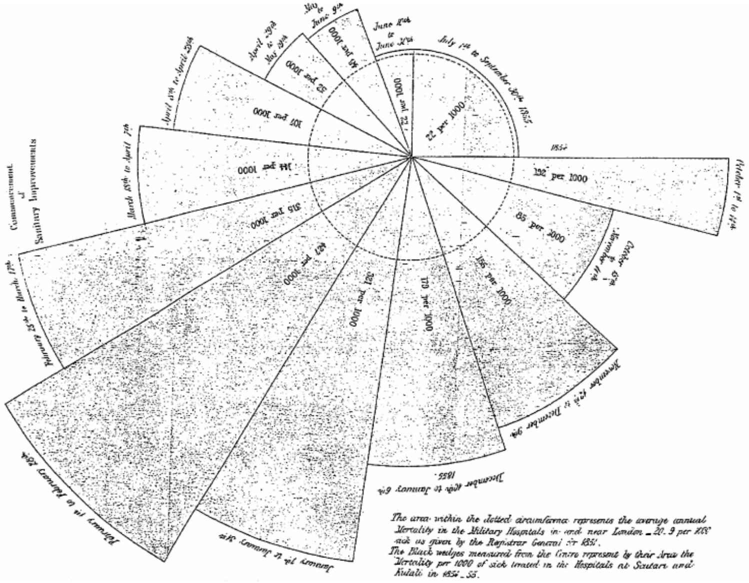
16th century account

Statistics: at first the only data analytic discipline

- the Royal Statistical Society established in 1834 - Babbage

- almost no mathematical methods or probability theory for the first 50 years

- graphics: e.g. Florence Nightingale Coxcomb plot, 1858



- 1885 paper by Edgeworth appeared which included *‘probability, the normal curve, use of the modulus, and the fluctuation, discusses n or n-1 as divisor, ...discusses significance tests, use of the median, parametric versus non-parametric tests, describes normal and Poisson approximations to the binomial, ... and deals with the tendency of a mean towards normality’*

- 1893 papers: bivariate normal (Edgeworth)
  correlation coefficient (Galton)
  standard deviation and skewness (Pearson)

- then more and more mathematics

- *Journal of the American Statistical Association* launched in 1888

So, roots in and before the 19th century

But, the discipline of statistics is essentially a 20th Century science

- Fisher (1890-1962)
- Jeffreys (1891-1989)
- ES Pearson (1895-1980)
- Ramsey (1903-1930)
- Savage (1917-1971)
- de Finetti (1906-1985)
- Neyman (1894-1981)

etc etc etc

Development driven by application domains (until recently)

- *agriculture* - experimental design
- *medicine* - survival analysis, graphical models
- *psychology* - factor analysis, linear structural equation models, item response theory
- *ecology* - ordination, multidimensional scaling
- *speech recognition* - hidden Markov models
- *robotics* - reinforcement learning

Once invented, the ideas are applied elsewhere

And then the computer arrived

⇒ changes in statistics

⇒ advent of new data analytic disciplines, with new perspectives on data analysis

- *database technology*: storage and manipulation of data
- *machine learning*: modelling natural and artificial learning
- *computational learning theory*: theoretical
- *pattern recognition*: for practical classification
- *data mining*: for large data sets

Today’s data

- automatic data acquisition

- size of data sets: Gigabytes, Terabytes, ...
    - number of cases
    - number of dimensions

Requires a different approach to analysis
- automation
- requires the tools of the computer

- complexity of data sets

- dynamic data sets
    - need for rapid analysis
        - speech recognition
        - commercial decisions

- different types of data
    - continuous and categorical
    - qualitative
    - fuzzy?
    - text, image, signal, ...
    - metadata

Data quality has always been a key issue

- gains new prominence when much analysis is *automatic* or mediated by machine

The role of mathematics

Statistics has been mathematically based since c.1900

It is seen as a mathematical science by many

- it has mathematics at its roots
- it is often based in mathematics departments
- but cf engineering, surveying

But not by everyone

Resistance to regarding stats as mathematics on two counts:

- that understanding the application domains plays a key role
- the importance of computational influences

Mathematics:

Assume properties of nature and deduce consequences

Data analysis:

Observe consequences (=data) and deduce properties of nature

To be a good data analyst one needs to understand the data

(Mathematics prodigies, but no such thing as a data analyst prodigy)

Mathematical software: Maple, Mathematica, etc.
vs
Data analytic software: Splus, SAS, SPSS, ...

'The main danger, I believe, in allowing the ethos of mathematics to gain too much influence in statistics is that statisticians will be tempted into types of abstraction that they believe will be thought respectable by mathematicians rather than pursuing ideas of value to statistics. One origin of this temptation is undoubtedly the siting of statisticians working in Universities in Departments of Mathematics; the pressure on the statisticians to develop their researches in directions thought to be acceptable to mathematicians may then become too strong to be easily resisted. However, there is little doubt that it ought to be resisted, for the two disciplines have very different objectives.'

John Nelder

'As a result of the would-be mathematicians in statistics, it has been dominated by useless theory and fads. . . If statistics is an applied field and not a minor branch of mathematics, then more than 99% of the published papers are useless exercises. (The other colleagues in statistics I have spoken to say this is an exaggeration and peg the percentage at 95%. Either way it is significant.) The result is a downgrading of sensibility and intelligence. . . But among all of the trash, there are a few places where theory has been useful.'

Leo Breiman

Statistics has suffered from the perception that it is a branch of mathematics

Perhaps other data analytic disciplines have gained from this:

- data mining: name suggests practical usefulness
- Six Sigma: in place of experimental design

Several cultures separated by a common language

Contrast 1:

- Statistics emphasis on *models*
- Machine learning emphasis on *algorithms*

Contrast 2:

- Mathematics brings emphasis on *rigour*
- But: It also brings *caution* and *risk aversion*:
    - do not publish your theorem until you can prove it
    - do not publish your algorithm without a convergence proof

Contrast 3:

- Statistics emphasis on *inference*
- Data mining emphasis on *description*

Natural, given the origins of the disciplines?

- statistics: how does nature work?
- data mining: what is in my database?

Inference builds a model for how the data arose
Inference assumes no distortions in the data
Or that the distortions are included in the model

Data mining has traditionally tended to ignore such issues

- DM rush in where stats fear to tread
  - possibly achieve more
  - but risk falling flat

Differences in emphasis example 1: *Supervised classification*

Computer	**Statistics**
Brain model?	To classify objects
Perceptron	LDA, logistic regression
Adaptive est'n	Batch, iterative
Training set	Design set
Emphasise non-overlapping classes	View in terms of overlapping densities
Error rate criterion	Separability criterion
Overfitting	Model form prevents overfitting
Implementation	Math rigour
Tackle tough problems	Rigorous solution to easier problems

Recursive partitioning methods

- **computing**:
  - emphasis on interpretation, links to rule-based systems

- **statistics**:
  - emphasis on predictive accuracy

Simple linear methods

Statistics

Regression ➔ LDA
Logistic regression ➔ GLMs

Based on clear underlying model

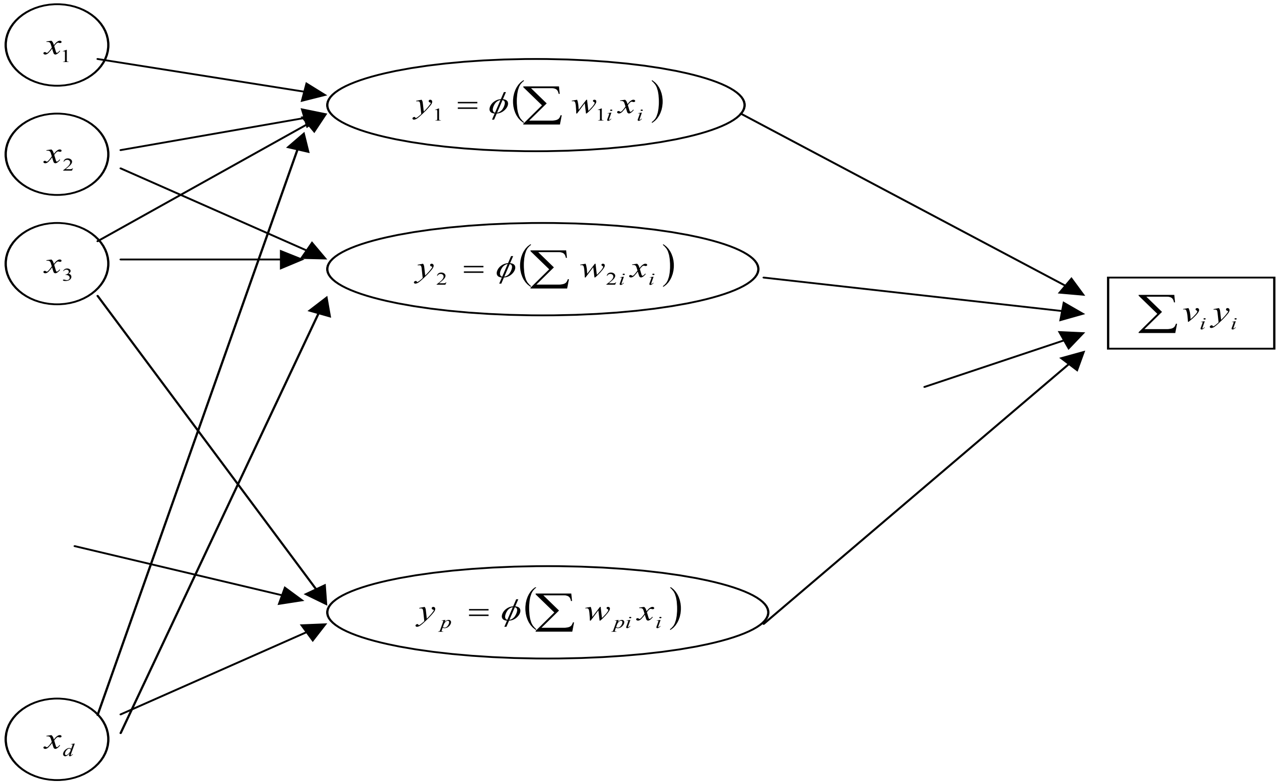
Computational

↗ neural networks

Perceptron →→

↘ support vector machines

Neural networks


x_1
x_2
x_3
x_d
$y_1 = \phi\left(\sum w_{1i} x_i\right)$
$y_2 = \phi\left(\sum w_{2i} x_i\right)$
$y_p = \phi\left(\sum w_{pi} x_i\right)$
$\sum v_i y_i$

$$p(\omega_1 \mid \mathbf{x}) = \sum \beta_i \phi_i\left(\sum \gamma_j \eta_j\left(\sum \delta_k x_k\right)\right)$$

Originated in *computing*

Sold as a black box tool for finding models even if know nothing about the data

- generally, if you do know something about the data you can do better
- automatic data analysis? c.f. variable selection

Classical errors: overfitting

Provided stimulus for deeper theoretical understanding

–e.g. deeper understanding of generalisability

Wide range of tools, as powerful as ANNs and SVMs, now developed within statistics

- projection pursuit
- generalised additive models
- multivariate adaptive regression splines
- nonparametric methods (e.g. kernel)
- . . . . .

Support vector machines

ANNs generalise linear functions using linear combinations of nonlinear transformations of ...

SVMs stick to linear functions – but in a space defined in terms of transformations of the *x*

(Very high dimensional space – problems sidestepped by some clever mathematics)

For classification:

- find the linear separating surface which best separates the classes
- which has the greatest *margin*

Rapprochement with statistics in progress

Simple linear models:

Perceptron vs *logistic regression* for classification

Perceptron:

- focus: decision surface
- criterion: error count
- estimation: error-correction algorithm

Logistic regression

- focus: ***overall model*** for $P(1 \mid x)$
    - classify by comparing $\hat{P}(1 \mid x)$ with a threshold
- criterion: likelihood $\log L \propto \sum_{i=1}^{n} \log \hat{p}(0 \mid \mathbf{x}_i)$
- estimation: iteratively weighted least squares

So:

- perceptron is finding a solution to the problem we want to solve

- log reg is finding an indirect solution

- perceptron optimises criterion of interest

- log reg optimises some other criterion

- log reg optimises a global criterion

⇒ perceptron is better for the classification problem

BUT: only if the costs are known (equal) a priori

If these are not known a priori, or are not equal, log reg is better

Differences in emphasis example 2: *pattern discovery*

Patterns are *anomalous local features* in a distribution

The data: a sample or population of cases

Signs of patterns: unexpected local dense regions of data points

Key problems:

(A) finding such features amongst millions of data points
- and with thousands of variables

(B) deciding if such features are real or chance

Data mining: most emphasis on (A) *[description]*

e.g. what percentage of my employees earn more than €*x* p.a.?

Statistics: most emphasis on (B) *[inference]*

e.g. what percentage of my employees are likely to earn more than €*x* p.a. next year?

Search:

- brute force search infeasible (and always will be)
- use relationships between potential patterns
- a priori algorithm
- sequential relationships

Inference

Sequence of *N* binary r.v.s $X_i, i=1,...,N$

$H_0: \quad X_i \sim F_0, \quad i=1,2...,N-m+1$, indep

$H_1: \quad X_i \sim F_1, \quad i=t,2...,t+m-1 \quad \text{and} \quad X_i \sim F_0 \quad \text{otherwise}$, indep

Define $\quad Y_t = \sum_{i=t}^{t+m-1} X_i$

The *scan statistic* is $\quad S_m = \max_{1\le t\le N-m+1} Y_t$

The critical region will be that set of values of S_m which exceed some value *k*, determined by the desired significance level.

Distribution $P\left(S_m \ge k\right)$ tough because of dependencies

New types of tool still seem to emerge from one paradigm or the other

Statistics:

- generalised additive models
    - powerful, flexible. cf neural nets

- multivariate adaptive regression splines
    - powerful, flexible. cf neural nets

- computationally intensive methods
    - jackknife
    - leave-one-out cross-validation
    - bootstrap

- MCMC

Computation:

- rule systems – ideas from brain modelling
- pattern detection
- genetic algorithms
- simulated annealing (but c.f. MCMC methods)
- boosting

Unification and synergy

Bayesian

- *updating*, the very basis of learning
- theory from statistics
- MCMC, practical

Theoretical understanding

- overfitting

Boosting

- originated in computer science
- it worked, but why did it work?
- statisticians showed it to be a kind of additive model
- and this deeper understanding led to improved boosters

‘***'There's no point in collaborating with people who know the same things as you'***

Martin Crowder

This is where the synergy comes from

The same is true for the interplay between disciplines

Progress has not stopped

I expect to see exciting developments

- *application areas: biotechnology, financial services, www,*
- *new and different kinds of data*
- *dynamic data*
- *large data sets leading to theoretical progress*
- *new ideas for fat data*
- *more elaborate models*
  - *but we must avoid going too far*

‘Greater statistics can be defined simply, if loosely, as everything related to learning from data, from the first planning or collection to the last presentation or report. Lesser statistics is the body of specifically statistical methodology that has evolved within the profession - roughly, statistics as defined by texts, journals, and doctoral dissertations. Greater statistics tends to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside of professional statistics. Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics.’

John Chambers

John has called the discipline of data analysis ‘greater statistics’, but I am sure we can all recognise what we do in his description. What we call it is not important.

‘What’s in a name? that which we call a rose

By any other name would smell as sweet.’

From *Romeo and Juliet*

END