university of groningen

NWO
Netherlands Organisation for Scientific Research

Society of Learning Algorithms
HLRS 2019

Data-driven Science and Undercover Theory

The case of clustering

⋆

Jan-Willem Romeijn
University of Groningen

Bacon’s epistemo-entomology

The theme of this talk is nicely captured in Bacon’s Novum Organon:

> *[Scientists] have been either empirics or dogmatical. The former, like ants, only heap up and use their store, the latter like spiders spin out their own webs. The bee, a mean between both, extracts matter from the flowers of the garden and the field, but works and fashions it by its own efforts.*
>
> Francis Bacon, The New Organon [Book One], 1620.

Machine learning may seem the work of ants. It focuses on collecting data and “letting those data speak for themselves”.

Bees, not ants

As most machine learning experts will tell you, this popular idea of machine learning is mistaken.

The general inevitability of inductive bias is well-known. But it is still a challenge to identify it in concrete cases.

Plan of talk

1. Machine learning in science
2. Concerns over reliability
3. Data-driven psychopathology
4. Learning from a fruit machine
5. Uncovering inductive assumptions
6. Automated text allocation
7. The epistemology of data science

1 Machine learning in science

Consider some examples of machine learning methods in the sciences:

- Psychiatrists use automated classification, e.g., hierarchical clustering, to come up with subtypes of heterogeneous diseases like depression.
- Linguists employ distant reading methods to disclose a corpus of texts, e.g., they use statistical methods to allocate the texts to clusters of similar ones.

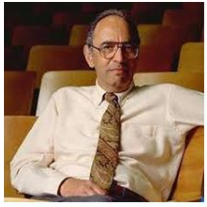
Machine learning in science (continued)

- Biomedical researchers employ automated causal discovery to identify the causal structure of mechanisms of gene expression in the cell.
- Astronomers use kernel methods to analyze the structure of galaxies, classifying stars according to their likely material composition.
- Sociologists interlink large data repositories with the aim of identifying connections of variable collected in separate studies.

In these cases the impact of theoretical starting points is difficult to trace. Does that matter?

Methodological concerns

The different nature of the new methods puts the continuity with existing theory under pressure.




And that the new methods are “black-boxed” makes it hard to hold machine-learning research accountable and motivate policy with it.

Transparency

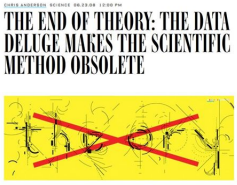
Continuity and accountability can be linked to transparency: we need to get a handle on the implicit assumptions in machine learning.

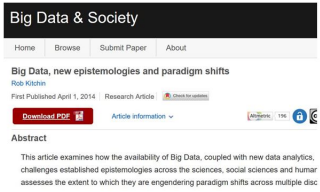
- If the assumptions implicit in the machine learning methods are uncovered, we can relate them to earlier models.
- A clear statement of the assumptions will allow us to criticize the methods and explain the results.

So we have to do the work of uncovering inductive assumptions in machine learning methods.

2 Concerns over reliability

Several machine learning researchers have proclaimed the “death of theory”.





This presents a separate motivation for transparency. We want our methods to be reliable.

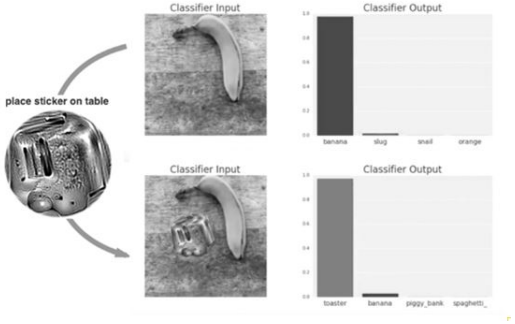
No free lunch

All inductive methods are in some way dependent on theoretical starting points: “there is no free lunch”.

If we have no control over the implicit assumptions of our methods, we do not know their conditions of applicability.

Inevitable inductive bias

As illustrated by so-called adversarials, machine learning methods are vulnerable to highly unexpected error.



We have to gain insight into the inductive assumptions to gain control over the reasons for misfiring and “debug”.

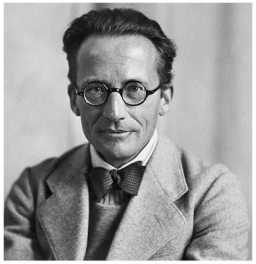
At NIPS 2017

Rahimi sparked a fierce debate by deeming machine learning the “new alchemy” and calling for an active “rigor police”.

This debate runs parallel to the philosophical one on inductive assumptions and the use of theoretical concepts.

“Anschaulichkeit”

The development of quantum mechanics offers an interesting example of the need for intelligibility.



Whether for epistemic, metaphysical or pragmatic reasons, scientists seem to prefer theories that provide insights.

Wish list

In sum, despite the attractiveness of theory-free methods, we want methods to. . .

- allow continuity in research,
- facilitate accountability,
- be understandable and communicable,
- have clear application criteria,
- avoid erratic mistakes.

For this we need clarity on the assumptions. How to reconstruct those?

3 Data-driven psychopathology

Psychiatric classification and sub-typing can be assisted by automated clustering methods.

Format: Abstract Send to

Biol Psychiatry Cogn Neurosci Neuroimaging. 2016 Sep;1(5):433-447.

Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders.

Marquand AF[1], Wolfers T[2], Mennes M[2], Buitelaar J[3], Beckmann CF[4].

Author information

1 Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen; Department of Cognitive Neuroscience , Radboud University Medical Centre, Nijmegen; Department of Neuroimaging (AFM), Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London.

2 Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen.

3 Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen; Department of Cognitive Neuroscience , Radboud University Medical Centre, Nijmegen; Karakter Child and Adolescent Psychiatric University Centre, Nijmegen, The Netherlands.

4 Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen; Department of Cognitive Neuroscience , Radboud University Medical Centre, Nijmegen; Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (CFB), University of Oxford, London, United Kingdom.

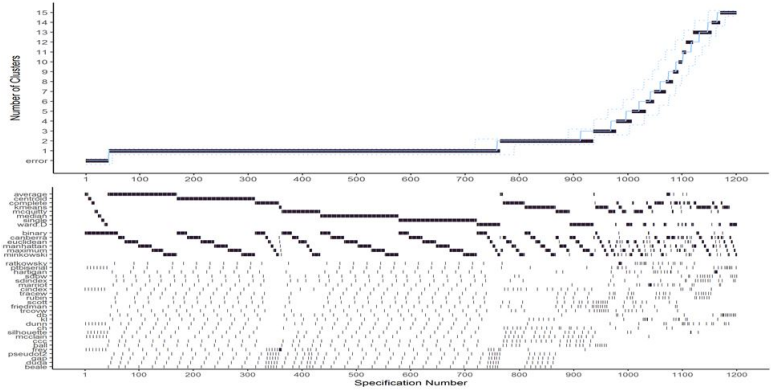
Abstract

Heterogeneity is a key feature of all psychiatric disorders that manifests on many levels, including symptoms, disease course, and biological underpinnings. These form a substantial barrier to understanding disease mechanisms and developing effective, personalized treatments. In response, many studies have aimed to stratify psychiatric disorders, aiming to find more consistent subgroups on the basis of many types of data. Such approaches have received renewed interest after recent research initiatives, such as the National Institute of Mental Health Research Domain Criteria and the European Roadmap for Mental Health Research, both of which emphasize finding stratifications that are

Do the methods identify patient groups that are distinct for the purpose of prediction and intervention?

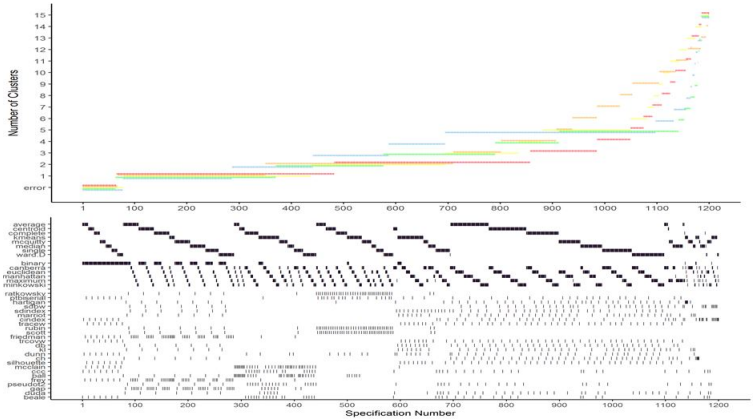
Specification curves

In a large comparison of clustering methods, Beijers et al. (manuscript) did not find much stability in the attempted clusterings.



Specification curves (continued)

When repeating the procedures for simulated data that were constructed to allow for easy detection, the same failures obtain.



The defects of automated clustering

We must not write off the use of data-driven methods in psychopathology but there are serious problems.

- There is wide variation and little overlap among the results of clustering subtypes of mental disorders.
- The comparison does not point to any particular specifications as being most adequate.
- The theoretical choices do not relate to the clustering outcomes determined by them in a conspicuous way.
- Variance, noise variables, and outliers all contribute to the failure of the clustering.

4 Learning from a fruit machine

Inductive logic is arguably a precursor of machine learning. Consider sampling pieces of fruit Q_i:

Carnapian predictions are made on the basis of data alone:

$$P(Q_{n+1} = a|Q_1 \ldots Q_n) = \frac{n_a + \lambda/k}{n + \lambda},$$

where the number of possible results $k = 4$ and we might choose $\lambda = k$.

Analogy effects

Carnap gradually admitted more flexibility in the prediction rules. A good example is analogical prediction, e.g.,

$$P(Q_{n+1} = a | Q_1 \ldots Q_n) \quad = \quad \frac{n_{\{a,c\}} + \mu/2}{n + \mu} \times \frac{n_a + \lambda/2}{n_{\{a,c\}} + \lambda}.$$

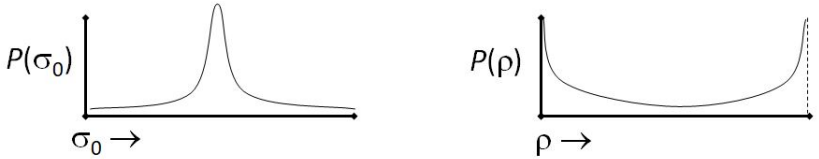
If $\mu < \lambda$, apples and bananas affect our expectation of cherries differently:

$$P(Q_{n+2} = c | Q_1 \ldots Q_n \wedge Q_{n+1} = a) \quad > \quad P(Q_{n+2} = c | Q_1 \ldots Q_n \wedge Q_{n+1} = b).$$

The literature offers numerous other systems that provide a handle on similarity in the data.

The use of models

It is helpful to redefine analogical prediction in Bayesian terms, by a prior over multinomial distributions: $P(H_\theta)$ with $\theta \in \langle \rho, \sigma_0, \sigma_1 \rangle$.



Translating prediction rules into Bayesian models is illuminating. Can we translate machine learning methods in the same way?

Putnam’s curse

There is a striking parallel between adversarials and so-called unlearnable sequences in inductive logic.

- Putnam (1963) challenged Carnap’s project by constructing a sequence that, relative to a set of prediction rules, is not predictable.
- If some rule assigns a high probability to an observation, the sequence wille have some other observation as its continuation.
- The formal learning theory developed after Putnam might shed light on the actively researched issue of adversarials in machine learning.

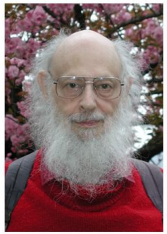
5 Uncovering inductive assumptions

Philosophy and statistics have seen many more unsuccessful attempts to rid inductive inference from its theoretical starting points.

We can learn from these examples of data-driven science. Where did the implicit theoretical assumptions go to hide?

Universal prediction

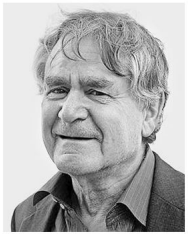
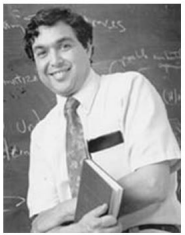
Sterkenburg (2017) offers an in-depth analysis of Solomonoff's idea of universal prediction, i.e., of considering all possible data patterns in prediction.



The predictions rest on the assumption of a highly skewed prior over all semi-computable measures. And in the end they fall prey to Putnam's curse.

Fiducial argument

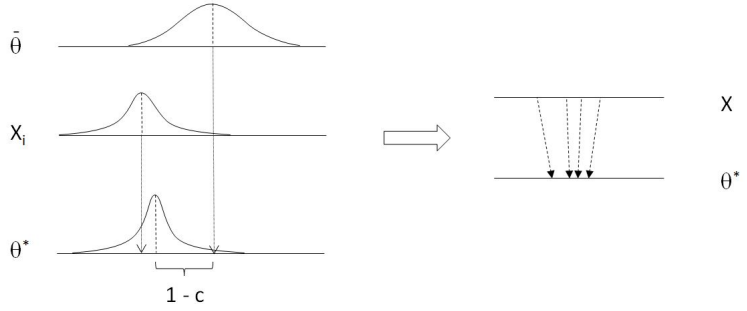
Fisher attempted to generate probabilistic conclusions about statistical hypotheses on the basis of data only.

But. . . his argument rests on the assumption of an improper implicit prior, projected onto the hypotheses via a functional model.

Shrinkage estimators

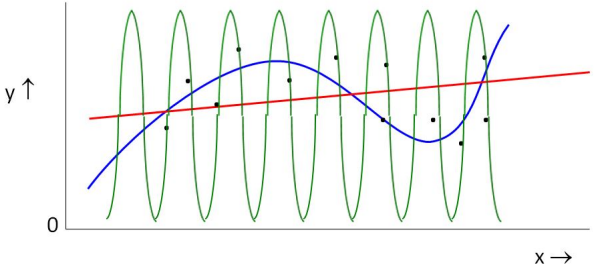
James and Stein (1957) derive that maximum likelihood estimators can be improved if we consider a collection of estimation problems.



As Efron and Morris (1977) already show, the improvement rests on an implicit empirical prior.

Model complexity

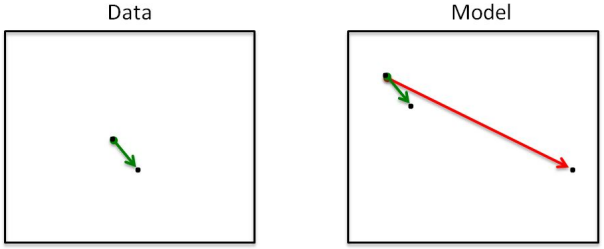
The sine model below offers a perfect fit while only using three free parameters.



This is an instant model selection hit. Fourier analysis trumps Taylor expansions.

Robustness and degeneacy

Nudging the data space cause the best estimate to change radically. But the real problem is that the model is degenerate.



This will show up in a properly approximated marginal likelihood: the prior term will act as substantial penalty.

Bayesian statistics

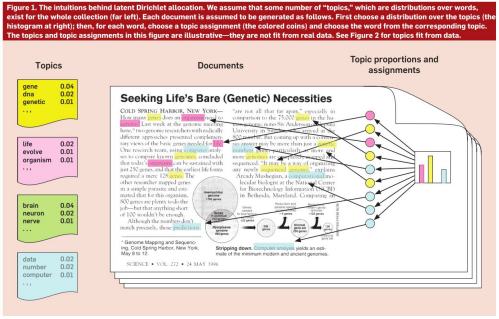
Modeling assumptions and prior opinion are made explicit. For conceptual purposes, Bayesian statistics is most useful.



Through the notion of exchangeability, even De Finetti's version of Bayesian inference rests on an assumed structure in the data.

6 Automated text allocation

Historians and linguists employ “distant reading” to grasp large corpora.



Latent Dirichlet Allocation is a model-based tool for sorting items in the corpus in a data-driven way.

Robustness analysis
Once again we investigated the behavior of the method on unprepared data and on data in which the clustering was already built in.

- For the unprepared data taken from Wikipedia, the resulting clusters showed only small internal coherence, and they were hard to interpet.
- For the data obtained from a labelled news feed, varying the specifications of the LDA model led to a range of non-nested clusterings.
- Minimizing the distance to the true distribution did not offer a handle on the choice of specifications.

Comparison to clustering methods

Much like clustering, the results of the LDA are “hit-and-miss”. The ability of the tool to identify the correct text groupings is doubtful.

Table 5.12: The result of BBC(5,10,10)

Topic #1	**Topic #2**	**Topic #3**	**Topic #4**	**Topic #5**
0.004*"new"	0.004*"mr"	0.005*"year"	0.004*"technology"	0.005*"mobile"
0.004*"mobile"	0.003*"new"	0.004*"mr"	0.004*"mr"	0.005*"new"
0.004*"mr"	0.003*"use"	0.004*"music"	0.004*"year"	0.004*"year"
0.004*"users"	0.003*"technology"	0.004*"technology"	0.004*"games"	0.003*"use"
0.003*"first"	0.003*"software"	0.003*"new"	0.003*"music"	0.003*"technology"
0.003*"technology"	0.003*"time"	0.003*"mobile"	0.003*"game"	0.003*"games"
0.003*"game"	0.003*"first"	0.003*"digital"	0.003*"world"	0.003*"users"
0.003*"digital"	0.003*"net"	0.003*"net"	0.003*"time"	0.003*"get"
0.003*"make"	0.003*"many"	0.003*"games"	0.003*"digital"	0.003*"game"
0.003*"year"'	0.003*"world"'	0.003*"many"'	0.003*"new"'	0.003*"software"'

The crucial difference between the cases of clustering and LDA is in the deployment of a model. This can guide the debugging.

7 The epistemology of data science

Philosophy of science can help to introduce machine learning methods into science in a responsible way.

- Machine learning will very likely transform our sciences so we will have to focus attention there.
- Preliminary studies suggest that the outcomes of machine learning methods suffer from failures of robustness: unless assisted, they overfit.
- To improve on the assistance, our primary goal should be to identify the assumptions inherent in machine learning.

Making assumptions explicit

The foregoing suggests how we can uncover inductive assumptions inherent in the new machine learning methods. The rough idea is:

- Identify modeling assumptions by translating between predictive systems and Bayesian statistics.
- Consider the assumptions inherent in how the sample space and the space of hypotheses is constructed.
- Connect the specifications of the machine learning method explicitly to the input of the Bayesian model.

Why again?

Uncovering these assumptions is an important task for the philosophy of science.

- It will help to integrate the new methods into existing and more theoretical approaches.
- Similarly it will improve on the communicability and public acceptance of machine learning results.
- And it will make it easier to hold researchers accountable and criticize their conclusions.
- Most importantly, it will help to apply methods correctly and guard against unreliable inferences.

Thanks for your attention

Help from Talitha Anthonio and Lian Beijers is gratefully acknowledged.
Slides of the talk will be available at http://www.philos.rug.nl/~romeyn.
For comments and questions, email j.w.romeijn@rug.nl.