Transparency and interpretability of clinical prediction models

Niels Peek
Health eResearch Centre
School of Health Sciences
The University of Manchester

Opening the Black Box seminar
Brunel University, London, 21st November 2018
Essentially, all models are wrong, but some are useful

George E.P. Box
Menu

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions
What are learning health systems?

A system becomes a learning system when it can continuously and routinely improve itself by reflecting on its inputs, processes, and outputs.

A learning health system harnesses the power of data and technology to learn from every patient, and feed the knowledge of “what works best” back to clinicians and patients to create cycles of continuous improvement.

The learning health cycle

Friedman et al., Yearb Med Inform 2017.
Example: community-acquired pneumonia

- Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US.
- It is the 6th leading cause of death, and responsible for >1m hospital admissions per year.
Example: community-acquired pneumonia

- Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US.
- It is the 6th leading cause of death, and responsible for >1m hospital admissions per year.
- If we can predict which CAP patients are at high risk of death, we can use these models to decide if a patient needs to be admitted to hospital.
Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US and responsible for >1m hospital admissions per year. If we can predict which CAP patients are at high risk of death, we can use these models to decide if a patient needs to be admitted to hospital.

Abstract

This paper describes the application of eight statistical and machine-learning methods to derive computer models for predicting mortality of hospital patients with pneumonia from their findings at initial presentation. The eight models were each constructed based on 9847 patient cases and they were each evaluated on 4352 additional cases. The primary evaluation metric was the error in predicted survival as a function of the fraction of patients predicted to survive. This metric is useful in assessing a model’s potential to assist a clinician in deciding whether to treat a given patient in the hospital or at home. We examined the error...
1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions
MACHINE LEARNING

UNSUPERVISED LEARNING
Group and interpret data based only on input data

SUPERVISED LEARNING
Develop predictive model based on both input and output data

CLUSTERING
CLASSIFICATION
REGRESSION
Supervised learning

Build Phase

Training Data -> Feature Vectors -> Estimator Algorithm

Labels

Operational Phase

New Data -> Feature Vector -> Predictive Model -> Prediction
T-Test

Logistic Regression

Elastic Net

Gradient Boosting

Deep Learning

From a presentation by Tom Liptrot
**Inductive bias**

- Inductive bias (= learning bias): The set of assumptions that a learning algorithm uses to construct a model from data

- Statistical models typically have a stronger inductive bias than machine learning methods, because they require prior specification of relevant features

- Assumption-free learning does not exist

- But we can reduce the impact of inductive bias by using more complex models

- ... at the expense of increasing variance
The bias-variance tradeoff (1)
The bias-variance tradeoff (2)
The bias-variance tradeoff (3)
Menu

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions
Explanatory models

- Explanatory models are statistical models with high bias that are exclusively used for causal explanation.
- They are used for testing causal hypotheses in observational data.
- Predominant use of data in economics, psychology, education, and other social sciences.
- In data science terms, they have a strong inductive bias.

Example: Unified Theory of Acceptance and Use of Technology (UTAUT)

Performance Expectancy

Effort Expectancy

Social Influence

Facilitating Conditions

Behavioral Intention

Use Behavior

Figure 3. Research Model

Table 14. Measurement Model Estimation for the Preliminary Test of UTAUT

(a) T1 Results (N = 215)

<table>
<thead>
<tr>
<th></th>
<th>ICR</th>
<th>Mean</th>
<th>S Dev</th>
<th>PE</th>
<th>EE</th>
<th>ATUT</th>
<th>SI</th>
<th>FC</th>
<th>SE</th>
<th>ANX</th>
<th>BI</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE</td>
<td>.92</td>
<td>5.12</td>
<td>1.13</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>.91</td>
<td>4.56</td>
<td>1.40</td>
<td>.31***</td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATUT</td>
<td>.84</td>
<td>4.82</td>
<td>1.16</td>
<td>.29***</td>
<td>.21**</td>
<td>.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SI</td>
<td>.88</td>
<td>4.40</td>
<td>1.04</td>
<td>.30***</td>
<td>-.16*</td>
<td>.21**</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC</td>
<td>.87</td>
<td>4.17</td>
<td>1.02</td>
<td>.18*</td>
<td>.31***</td>
<td>.17</td>
<td>.21**</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>.89</td>
<td>5.01</td>
<td>1.08</td>
<td>.14</td>
<td>.33***</td>
<td>.16*</td>
<td>.18**</td>
<td>.33***</td>
<td>.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANX</td>
<td>.83</td>
<td>3.11</td>
<td>1.14</td>
<td>-.10</td>
<td>-.38***</td>
<td>-.40***</td>
<td>-.20**</td>
<td>-.18**</td>
<td>-.36***</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>BI</td>
<td>.92</td>
<td>4.07</td>
<td>1.44</td>
<td>.38***</td>
<td>.34***</td>
<td>.25***</td>
<td>.35***</td>
<td>.19**</td>
<td>.16*</td>
<td>-.23**</td>
<td>.84</td>
</tr>
</tbody>
</table>

Notes: 1. ICR: Internal consistency reliability.
2. Diagonal elements are the square root of the shared variance between the constructs and their measures; off-diagonal elements are correlations between constructs.
3. PE: Performance expectancy; EE: Effort expectancy; ATUT: Attitude toward using technology; SI: Social influence; FC: Facilitating conditions; SE: Self-efficacy; ANX: Anxiety; BI: Behavioral intention to use the system.
Is it a problem if the UTAUT model is wrong?
The big target here isn't advertising, though. It's science. The scientific method is built around testable hypotheses. [...] Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). [...] But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.
Prediction models

**Predictors**
- age, sex
- diagnosis
- medical history
- clinical findings
- lifestyle
- genome
- ...

**Outcome**
- event y/n
- time to event

**Statistical models**
- statistical model (regression model)
- decision tree
- random forest
- support vector machine
- deep neural network
- ...

**Legend**
- $f$ represents the prediction function.
Example revisited: hospital admissions

- Resources are too scarce to give preventive interventions to every patient
- Prediction models can help to deploy these resources in patients with the highest risks
Example revisited: pneumonia

- Community-acquired pneumonia (CAP) is a common illness affecting >3m people annually in the US
- It is the 6th leading cause of death, and responsible for >1m hospital admissions per year
- If we can predict which CAP patients are at high risk of death, we can use these models to decide if a patient needs to be admitted to hospital
Question

Is it a problem if the risk prediction model is wrong?

Is it important that we can interpret such a model or understand its predictions?
Menu

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions
Supervised machine-learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? Models should be not only good, but also interpretable, yet the task of interpretation appears underspecified. The academic literature has provided diverse and sometimes non-overlapping motivations for interpretability and has offered myriad techniques for rendering interpretable models. Despite this ambiguity, many authors proclaim their models to be interpretable axiomatically, absent further argument. Problematically, it is not clear what common properties unite these techniques.

This article seeks to refine the discourse on interpretability. First it examines the objectives of previous
Model interpretability: What?

- The notion of interpretability is ill-defined, and has no formal meaning
- Claims regarding interpretability often exhibit a quasi-scientific character
- It is useful to make a distinction between:
  - model transparency
  - post-hoc interpretability of predictions

Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).
Model interpretability: Why?

- The formal objectives of supervised learning (test set performance) do not capture interpretability.
- The demand for interpretability arises from additional objectives related to real-world deployment.

Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).
Model interpretability: Why?

- Trust
- Causality
- Transferability
- Informativeness
- Fair and ethical decision making

Z. Lipton, ACM Magazine Queue - Machine Learning 2018;16(3).
These corollaries offer a somewhat finer depiction of what informatics is and is not, and what informaticians do.

Corollary 1: Informatics is more about people than technology.

This corollary can be seen from the "person" appearing twice in the theorem, while the information resource appears only once. This first corollary reminds us that information resources must ultimately be built for the benefit of people. This corollary also shows what informatics is not. As illustrated in Fig 2, creating resources that function as "oracles" and may be seen as competing with people—resources that seek, on their own, to be better than the person unassisted—is not a pursuit of interest in informatics.

Corollary 2: In order for the theorem to hold, the resource must offer something that the person does not already know.

This corollary helps explain why the development of effective information resources is often so challenging. What the resource offers to the person must not only be correct, it must also be informative. It must increment his/her knowledge in some significant way. Because the persons who interact with these resources typically bring to any task a high level of personal knowledge about the domain in which they are working, the requirement that the resource be informative sets a very high bar for the theorem to be satisfied.

Corollary 3: Whether the theorem holds depends on an interaction between person and resource, the results of which cannot be predicted in advance.

This final corollary reminds us that what we know about the person alone, and what we know about the resource alone, cannot tell us what will happen when the resource is deployed. The theorem can fail to hold, even though the resource has potential to be helpful, if it is used by the person in ways that do not enable the realization of its potential. This can happen because the resource is poorly designed and thus hard to use well, or because the person does not know enough about the domain to make best use of the resource.

By way of conclusion, the theorem and its three corollaries seek to establish the timbre of informatics rather than its libretto. I hope this formulation will promote understanding through simplicity, by stimulating imagination and further discussion. Sometimes less is more, and a picture is invariably worth a thousand words.
Clinical Decision Support in the Era of Artificial Intelligence

Clorint and researchers have long envisioned the day when computers could assist with difficult decisions in complex clinical situations. The first article on this subject appeared in the scientific literature about 60 years ago, and the notion of computer-based clinical decision support has subsequently been a dominant topic for informatics research. Two recent Viewpoints in JAMA highlighted the promise of deep learning in medicine. Such new data analytic methods have much to offer in interpreting large and complex data sets. This Viewpoint is focused on the subset of decision support systems that are designed to be used interactively by clinicians as they seek to reach decisions, regardless of the underlying analytic methodology that they incorporate.

With the evolution of digital and communication technologies plus innovative software methods, the ability to offer high-quality support to clinicians has resulted in impressive new capabilities and several commercial products. For example, many decision support tools are built into medical devices, creating new ways to visualize or interpret data that are provided to expert users. Artificial intelligence programs, which are increasingly based on a variety of machine learning and natural language processing methods, are especially prominent in these data interpretation and text mining settings.

Why, then, do clinical decision support systems (CDSSs) designed for direct interactive use by clinicians have challenges of credibility and adoption when the literature has been replete for 4 decades with studies that present computing systems demonstrating diagnostic accuracy that rivals the performance of expert clinicians? The reasons are varied and reflect the realities and complexities of clinical practice. Biomedical informatics and have long understood those reasons, recognizing the spectrum of capabilities and characteristics that must be incorporated into a CDSS if it is to be accepted and integrated into routine workflow:

- Black boxes are unacceptable: A CDSS requires transparency so that users can understand the basis for any advice or recommendations that are offered.
- Time is a scarce resource: A CDSS should be efficient in terms of time requirements and must blend into the workflow of the busy clinical environment.
- Complexity and lack of usability thwart use: A CDSS should be intuitive and simple to learn and use so that major training is not required and it is easy to obtain advice or analytic results.
- Relevance and insight are essential: A CDSS should reflect an understanding of the pertinent domain and the kinds of questions with which clinicians are faced to want assistance.
- Delivery of knowledge and information must be spectacular: A CDSS should offer advice that recognizes the expertise of the user, making it clear that it is designed to inform and assist but not replace a clinician.
- Scientific foundation must be strong: A CDSS should have rigorous, peer-reviewed scientific evidence establishing its safety, validity, reproducibility, and reliability.
- Health care is a particularly challenging decision-support domain: A CDSS requires strong analytical capabilities that can function in a domain where the underlying causal mechanisms are often incomplete and taut, and an approach is accordingly inevitable. A CDSS should offer advice that avoids additional data entry tasks, such as a CDSS that acquires the bulk of the data needed for a case through integration with an electronic health record (EHR). Today's EHRs have not yet reached this level because they generally lack the cross-platform transparency and standards that would be necessary for single CDSS to be tightly integrated with multiple EHR products or implementations.
- Different decision-making tasks often pose different challenges for a CDSS. For example, a system designed to assist with clinical diagnosis is very different from one that is intended to assist with therapy planning. A CDSS for diagnosis can generally be built on linksages between clinical data and gold standards for accuracy (eg, biopsies, autopsies, biomolecular markers, or surgical findings). But in formulating a therapeutic plan, especially in complex settings, there is often no gold standard, and there may be disagreement, even among experts. For example, an early study evaluated a program designed to assist with the selection of antibiotic therapy.

Despite the enthusiasm for exploring the potential of artificial intelligence and decision support in clinical settings, several complexities limit the ability to move ahead quickly.

© 2018 American Medical Association. All rights reserved.

An evaluation of machine-learning methods for predicting pneumonia mortality

Gregory F. Cooper, Constantin F. Aliferis, Richard Ambrosino, John Aronis, Bruce G. Buchanan, Richard Caruana, Michael J. Fine, Clark Glymour, Geoffrey Gordon, Barbara H. Hanusa, Janine E. Janosky, Christopher Meek, Tom Mitchell, Thomas Richardson, Peter Spirtes

*Center for Biomedical Informatics, Suite 8084 Forbes Tower, 200 Lothrop Street, University of Pittsburgh, Pittsburgh, PA 15261, USA
+Intelligent Systems Laboratory, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15213, USA
\*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
\*Division of General Internal Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
\*Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA
\*Division of Biostatistics, Department of Family Medicine and Clinical Epidemiology, University of Pittsburgh, Pittsburgh, PA 15261, USA

Accepted 14 October 1996

Abstract

This paper describes the application of eight statistical and machine-learning methods to derive computer models for predicting mortality of hospital patients with pneumonia from their findings at initial presentation. The eight models were each constructed based on 9847 patient cases and they were each evaluated on 4352 additional cases. The primary evaluation metric was the error in predicted survival as a function of the fraction of patients predicted to survive. This metric is useful in assessing a model’s potential to assist a clinician in deciding whether to treat a given patient in the hospital or at home. We examined the error

Example: community-acquired pneumonia (CAP) is a common illness affecting >3 million people annually in the US and responsible for >1 million hospital admissions per year. If we can predict which CAP patients are at high risk of death, we can use these models to decide if a patient needs to be admitted to hospital.
What Cooper et al. found

- The most accurate model was a neural network which outperformed other methods (e.g. logistic regression) by a wide margin.

- One of the methods was a rule-based method that found the rule: \( \text{HasAsthama}(x) \Rightarrow \text{LowerRisk}(x) \)

- The authors chose to deploy the rule-based model, and left out this rule.

- Several authors have since argued that prediction models must be **intelligible** and **editable**.

Question

Why did the rule-based method infer that asthma patients were at low risk?
Opening the black box

- If managed in the same way, asthma patients with CAP would be at a higher risk than other patients.
- In our current care system, this risk is recognised and therefore asthma patients are managed differently.
- Their net risk is therefore lower.
Accuracy vs intelligibility

- Boosted Trees
- Random Forests
- Neural Nets
- Single Decision Tree
- Logistic Regression
- Naive Bayes
- Decision Lists

From a presentation by Rich Caruana
Citizens Juries

- A “Citizens Jury” is a public engagement process that allows policy makers to hear thoughtful input from an informed microcosm of the public

- In Feb/Mar 2019 we will organise 2 Citizens Juries (5 days each) on explainable AI

- The juries will explore the trade-offs between performance and explainability of computer algorithms

- Scenarios in clinical medicine, criminal justice, and professional recruitment will be considered
Menu

1. Context: learning health systems
2. Basics of supervised learning
3. Explanatory vs prediction models
4. Interpretability of prediction models
5. Conclusions
Conclusions (1)

- All models are wrong, but some are useful
- Prediction models are a radically pragmatic, “end-of-theory” use of data to engineer systems
- Their core purpose is to make predictions for future, unseen instances – not to increase our understanding
- But at the interface with humans, the need arises to provide interpretability
Conclusions (2)

- Model interpretability is still a poorly defined notion.
- It is ultimately something that should be studied by psychologists, not computer scientists.
- To understand a model, we must understand its relationship with the real world.

![Diagram of model interpretability process]
Thank you