In: Emre Bayamlıoğlu, Irina Baraluic, Liisa Janssens and Mireille Hildebrandt (eds), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, 2018: Amsterdam University Press, 130-133. DOI 10.5117/9789463722124/CH22.

Imagining Data, Between Laplace's Demon and the Rule of Succession

*Reuben Binns**

Abstract

Laplace presented two models of prediction. One posits an imaginary demon with complete knowledge of the world and the laws of nature. The other was the rule of succession, which has the modest aim of putting a precise figure on justified inductive faith given limited observations and no background theoretical knowledge. Despite their non-causal, non-explanatory nature, modern computational prediction systems are often presented as approximations of Laplace's omniscient demon, rather than the rule of succession. One approach to justice in algorithmic prediction is to attempt to make it more like the former and less like the latter; from this perspective, better predictions – based on causal models – are fairer predictions. Causal models explain the actual world, with all its injustices, but justice is also partly about the ability to imagine things that are not in fact the case. While we clearly fall short of the total predictive capacity of Laplace's demon, our human faculties of imagination give us access to an infinite variety of possible alternative worlds against which the actual world can be compared; this is the basis on which to assess the justice of algorithmic predictions. And to imagine such alternative possible worlds is not (or not only) an exercise in scientific conjecture; it is an inherently political act.

Keywords: profiling, prediction, causality, justice, fairness

Introduction

What is the probability that the sun will rise tomorrow? In 1814, Laplace posed this question and a means of answering it. By Laplace's reckoning, there were 1,826,251 recorded days in human history in which the sun had risen, and none in which it had not, giving odds of 0.9999994 % that the sun will rise tomorrow. While Laplace's 'rule of succession' was a poor answer to Hume's problem of induction, it was part of a theory of probability and statistical inference which fleshed out the actuarial calculations of Bayes, and ultimately furnished the mathematical foundations of modern statistical inference and machine learning.

Perfect prediction

But Laplace is perhaps better known for a thought experiment which informed the classical definition of a deterministic universe (Laplace 1951, 4):

> *An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.*

Thus, with complete knowledge of the position and motion of every atom, and all the laws of cause and effect existing in nature, the future could be perfectly predicted like clockwork.

Laplace presents these two models of prediction; one, an impossible ideal, the other a pragmatic compromise. On the one hand, an imaginary intellect with complete knowledge (Laplace's 'demon' as it became known) encapsulates the fantasy of perfect prediction which may be metaphysically coherent but is epistemically forever out of reach. On the other, the rule of succession has a more modest aim; to put a precise figure on our inductive faith given limited observation and no background knowledge.

Modern computational systems of prediction, classification and inference are, for the most part, following the rule of succession. And this rule is expected to do more than ever; as data are generated from any corner of economic or social life, they are pressed towards the prediction or classification of some unknown, in the hope of reducing risk, increasing efficiency or exerting control. Machine learning is an exercise in fitting curves around these known data points in a multi-dimensional feature space, in such a way as to maximise the number of future data points falling on the right sides of the curves. Where Laplace plundered the historical record for observations of the sun's rising, modern data scientists mine the legacy databases of banks and welfare systems, or construct new ones out of the many digital traces we leave online.

Alongside widespread enthusiasm for data-driven decision making in the private and public sector, there are often strong concerns over its use to make consequential decisions concerning people's lives. Such concerns about algorithmic decision-making have been articulated in terms of the threats to individual dignity, procedural justice, discrimination and fairness (see e.g. Hildebrandt and Gutwirth 2008). As a result of these concerns, data protection regulation affords various rights to subjects of data-driven-decisions; foremost, to not be subject to them, but also to have their logic meaningfully communicated, and to request a human reviewer.

One way of framing these concerns, and the philosophical motivation for such legal protections, is in terms of the damaging consequences of conflating Laplace's two models of prediction. Despite their non-causal, non-explanatory nature, the insights of machine learning are often presented and treated as if they approximate those of Laplace's omniscient demon. Patterns of geolocation, mouse movements, locations within social graphs and their statistical associations with loan repayments, retail purchases or employee productivity are taken as equivalent to the demon's knowledge of the position and forces of nature.

Laplace's sunrise problem is not a fruitful machine learning problem, but it is an extreme example which illustrates the limited nature of the probabilistic knowledge that statistical methods, following the rule of succession, can furnish us with. Relying purely on observation, it eschews theory. It does not pretend to know anything about the 'forces that set nature in motion'; it merely provides us with guidance on what to believe in the absence of such knowledge. We can explain our belief that the sun will rise not by reference to astronomical theory, but by subjective degrees of belief derived from numerical operations over observations.

Laplace readily acknowledged that estimations of likelihood are merely a set of consistent rules about how to act based on limited and subjective sets of evidence. And they cannot substitute for causal models; the rising of the sun, or human behaviours like repaying a loan, are fundamentally unlike the process of drawing coloured marbles from an urn or flipping a coin. As Ian Hacking charted, probability emerged in the 17th century as a new way of knowing, and statistical regularities became elevated to a status analogous to laws of nature (Hacking 2006). But patterns at population level are not explanations for any single individual's behaviour. From the perspective of Laplace's omniscient demon, no individual person is 60% likely to default on a loan or commit a crime; they either will or they won't, in the fullness of time, depending on the precise configurations of the position of matter and the operations of natural laws. But absent such omniscience, individual behaviours are indeterminate, and only predictable at the population level.

It's in this space of indeterminacy where we act in ways we personally identify with, and where we attribute both to ourselves and to others freedom, agency and intentionality. Regardless of how we understand the notion of free will at a metaphysical level, attributions of agency persist in the face of population-level statistical regularities. This is one reason why people may object to the use of statistical prediction to make decisions about individuals. Even if every other person with a given set of features acted in a certain way, the nth person sharing those same features might act otherwise.

This theme came to the surface in recent experimental work, where we probed people's perceptions of justice in response to a variety of hypothetical automated decisions, accompanied by a range of different explanations which aim to impart meaningful information about the system (Binns et al. 2018). One participant, reacting to a decision to deny an individual a financial loan on the basis of a machine learning model trained on data from prior borrowers, argued that: 'it's unfair to make the decision by just comparing him to other people and then looking at the statistics. He isn't the same person' (Binns et al. 2018, 7)." This suggests that ML-driven decisions will always be on some level unfair, because at any point, someone might act counter to the trend. As such, we need human intervention to allow for discretion and the chance that people might act otherwise.

But there is another potential response, one more likely to be favoured by advocates of such systems; make the system better to catch the exceptions. This means finding new sources of data, building more complex models which encompass different sub-groups, or both. Any discretion that might be exercised in the case of a human reviewer treating an individual differently to the model's output, could perhaps be subsumed under the statistical model by adding more data. This strategy is compelling because it suggests that the demands of justice are ultimately in line with the goal of accuracy.

But to call only for more data is a problematic response to questions of justice. The data you might need to update the model in ways that would enable it to handle the exceptions generated by human discretion might never exist. Training data from the real world usually does not encompass the full

range of possible values for a set of features. One cannot always draw samples from all logically possible populations for various societal, economic, or even biological reasons. For instance, there may not be data on the population of prisoners who were deemed 'high risk' but were released; or of those with low credit scores who were nevertheless given loans; or of pregnant males (except in rare circumstances). This is a practical problem for machine learning in any sphere, not only those in which human lives are at stake. Laplace could not experiment with the astronomical circumstances underlying the sun's motion to say 'why', beyond induction from the past, we should expect the sun to rise tomorrow, or explain the conditions under which it would not.

But non-existent data is not just a problem for machine learning. It is also, perhaps, something we need to imagine as a pre-condition for justice in decision-making. To understand why this might be so, consider recent work on 'de-biasing' machine learning (e.g. Pedreshi, Ruggieri and Turini 2008). The problem is that models may be trained on data which reflect unjust social biases, such that certain populations are more likely to be given a certain label. Both the variables used to predict an outcome, and those used to measure the outcome itself, might be biased. For instance, an educational qualification may be a decent proxy for a job applicant's knowledge, but if the awarding institutions have structural gender biases then a model for predicting applicant's future performance using such a proxy will be unfairly biased against women. Similarly, if work performance itself is measured by managerial reviews that are also gender biased, then both the predictor variables and the outcome labels will be biased, potentially reinforcing those underlying discriminatory patterns.

Imagining data justice

While various different definitions of discrimination and fairness have been proposed for correcting such systems, taken to their logical conclusion, they ultimately require us to go beyond the data and imagine alternate states of affairs in which some discriminatory patterns do not exist. They might require us to determine what qualification the female applicant would have got in an unbiased institution, or what evaluation she would have got as an employee in a discrimination-free workplace. This requires causal models of discrimination and social injustice. But even causally understanding injustice may not be enough. We may also need to imagine what the just alternative might be, i.e. imagine the situation of the individual as they might be under a fundamentally different, non-patriarchal society. Defining fair decisions thus requires thinking about counterfactual causal scenarios in imaginary worlds (and perhaps even 'impossible' worlds, but hopefully not).

This leaves us somewhere orthogonal to Laplace's two extremes of minimal inference to subjective probabilities from incomplete data, and the 'single formula' of the all-knowing intellect. Justice is partly about the ability to imagine things that are not in fact the case; while we clearly fall short of the total predictive capacity of Laplace's demon, our human faculties of imagination give us access to an infinite variety of possible alternative worlds against which the actual world can be compared. And to imagine alternative possible worlds is as much a political act as it is an exercise in counterfactual causal reasoning.

*****Reuben Binns is a researcher in Computer Science at the University of Oxford, and a research fellow in Artificial Intelligence at the UK Information Commissioner's Office.

References

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM. Paper no. 377.

Hacking, Ian. 2006. "The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference." 2nd edition. London: Cambridge University Press.

Hildebrandt, Mireille, and Serge Gutwirth, eds. 2008. Profiling the European citizen: Cross-Disciplinary Perspectives. Dordrecht: Springer.

Laplace, Pierre Simon. 1951. "A Philosophical Essay on Probabilities." Translated by Frederick Wilson Truscott and Frederick Lincoln Emory. 6th edition. New York: Dover.

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. 2008. "Discriminaton-aware data mining." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 560–68.