Algorithms in the Justice System: Some Statistical Issues

Royal Statistical

Society

November 08, 2018

1. Introduction

This contribution to the discussion of the use of algorithms in the justice system is submitted on behalf of the Royal Statistical Society (RSS). It has been prepared by the RSS Section on Statistics and the Law and addresses some relevant statistical issues. In particular, we draw attention to the prevalence of dangerously misleading statistical arguments in the current literature, and the consequent need for expert statistical input. We also present some recommendations regarding the use of software systems in analysing and producing legal evidence.

2. Statistical issues in risk assessment

Many algorithms in current use, in areas from insurance pricing to medical prognosis, involve some form of classification or prediction. In the justice system, Actuarial Risk Assessment Instruments (ARAIs) are algorithms that purport to predict recidivism. ARAIs have been used in the US as aids to bail, sentencing and probation decisions.

For use on a new case, such an algorithm will be fed with suitable input information on that case, will process that information somehow, and will deliver its prediction for the case. This might be as a categorical prediction, such as “Will recidivate”, or, more realistically and usefully, as a score (possibly but not necessarily in the form of a probability) indicating the risk of or uncertainty about the outcome. The initial construction of the algorithm will have involved some sort of processing of “training data” from many individuals.

The construction and use of an algorithm raises many statistical questions, including:

Transparency Are the details of the training dataset and the internal computational methods available and suitable for scrutiny?

Training data How were the training data generated and/or selected? Was any randomisation or experimentation involved? What population (if any) can the data be regarded as being representative of?

Variables How were the input variables selected and validated? How appropriate are they for the task at hand? Was the right output variable measured? How accurate were the measurements?

Analysis Was a logically sound analysis of the data conducted? What uncertainty attaches to the algorithm's outputs, and how should this be quantified?

Testing What empirical assessments of the performance of the algorithm have there been, and what range of contexts do these cover?

Generalisability How well can the system be expected to perform in new contexts, especially under intervention? (See Barabas *et al*. (2018) for criticisms of the use of predictive ARAIs to guide interventions.

The above issues are not unique to the justice system and have been recognised and discussed extensively in the more general statistical literature. However, a substantial literature has arisen specifically in the judicial arena, targeted on certain statistical aspects of ARAIs. The issues revolve around the statistical assessment of the empirical performance of an algorithm, treated as a "black box", in its application to new cases. Such assessment is clearly a vital ingredient in assessing the accuracy and fairness of an algorithm. Unfortunately, many of the most influential contributions to this specialist literature exhibit statistical errors and misunderstandings so severe that they could seriously endanger the fair administration of justice.

We illustrate this with two examples, the first focused on fairness and the second on accuracy, where the literature has been particularly misleading. The issues are somewhat technical, as well as logically subtle. We summarise essential points, referring to relevant literature for further details.

2.1. Fairness: Is COMPAS racially biased?

COMPAS is a proprietary recidivism ARAI from Northpointe, Inc. It takes as inputs the answers to 137 questions, including age, sex and criminal history, that are either answered by the individual or extracted from records, and (using an algorithm whose details are not in the public domain) outputs a score between 1 to 10 that purports to indicate how likely the individual is to reoffend (a higher score corresponding to a higher probability of recidivism). Race is not used as one of the inputs, although some of the other inputs could be regarded as associated with race.

There has been intensive statistically focused debate as to whether or not the COMPAS system is racially biased. This was initiated by an article (Angwin *et al*. 2016) on the investigative journalism website ProPublica https://www.propublica.org. This article has been widely construed and cited as showing that the COMPAS algorithm embodies a built-in bias against black people. However, a more careful analysis exposes serious problems with this claim. [1]

The basis for the ProPublica claim is as follows. Suppose we look at those who *did not* later go on to recidivate, and ask: What proportion of these received a medium or high risk score (between 5 and 10) thus, wrongly, suggesting they would probably recidivate, and likely leading to harsher sanctions. We can do this calculation separately for black and white individuals: the associated proportions were found to be 58% for blacks and 33% for whites. This discrepancy, say ProPublica, means that the system is biased against blacks: black non-recidivators are more likely to be penalised than whites. They argue that, in a fair system, these rates should be essentially the same for both racial groups.

The article was followed by a rebuttal from Northpointe (Dieterich *et al.* 2016) (see also Flores *et al.*, 2016), which pointed out that, for each of the 10 values of the COMPAS risk score, the proportion of blacks with that score who went on to recidivate was very close to the corresponding proportion of whites—meaning, they argue, that there is in fact no racial bias. Note that this argument involves an interpretation of "fairness" different from ProPublica's. The approach used by Northpointe is standard in equal employment and related discrimination cases in the U.S. and many other countries.

There has been a good deal of further wrangling in the literature as to whether or not the COMPAS algorithm displays bias. The argument revolves around the question: What should constitute an appropriate statistical indicator of racial bias? Should we (for each race) look at those eventually non-recidivating, and enquire as to their previously measured test score, as done by ProPublica? Or should we (as done by Northpointe) look at those receiving some specified test score, and enquire as to whether or not they later recidivate? The difference between these approaches has logical parallels in the infamous "prosecutor's fallacy" in the criminal courts, which involves confusing the probability of obtaining incriminating evidence, if the defendant is indeed innocent, with the probability that the defendant is innocent, given that the incriminating evidence was obtained. We can look at the situation in two different directions, and these can give very different answers.

A nice overview and analysis of the COMPAS problem was given in a Washington Post article (Corbett-Davies *et al.*, 2016). [2]This points out that if-as is indeed the case-the overall recidivism rate differs between blacks and whites,[3] then, as a mathematical necessity, we can not have "fairness" in both directions at once. See Fry (2018); Chouldechova (2017); Kleinberg *et al.* (2017) for more on this.

[1] Note that we are not claiming that the COMPAS system is satisfactory in all respects. Indeed it can be and has been criticised for failing on a number of the criteria listed above in Section 2.

[2] Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Washington Post Monkey Cage, online at https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.2b2b9590da1c.
See also Corbett-Davies *et al.* (2017) for a wide-ranging and thoughtful account of general statistical issues relating to fairness, using COMPAS as an example.

[3] This very fact might itself be regarded as an indicator of unfairness, but that would be down to biases in the wider justice system and in society at large, so not specific to the COMPAS algorithm.

Which then is the right direction in which to look in order to assess fairness? To address this question, it is helpful to make a distinction between two apparently similar but subtly different paradigms, namely *classification* and *prediction*. This distinction parallels that between *diagnosis* and *prognosis* in a medical context, which can be used as an analogy. In diagnosis, the patient is already suffering from some disease, which gives rise to various signs and symptoms that can be used to argue, backwards, as to what that pre-existing disease might be. In prognosis, however, the doctor argues, forwards, from the observed current condition of the patient to his likely future outcome. Since the event of recidivism is an uncertain future outcome, not a pre-existing condition, it is prognosis (prediction), not diagnosis (classification), that more closely parallels the problem of predicting recidivism.

As a simple example of a classification/diagnosis problem, consider a medical test for a disease, with two possible readings, positive and negative. However, the test is not perfect. When applied to a patient with the disease it (correctly) registers positive 95% of the time, while when applied to a healthy patient it (wrongly) registers positive 5% of the time. Importantly, these proportions can be taken as fixed characteristics of the apparatus, relevant no matter for which patient, and in what context, it is applied. But when we have applied it to a patient and obtained a positive reading, we have to look at the probabilities "backwards",[4] and then the implication of the reading will depend on the context. Suppose, for example the underlying rate of the disease is 10% for men, and 1% for women. Then the probability of disease in an individual with a positive reading will be about 68% for a male, but only 17% for a female. In spite of these differences, the apparatus could not be accused of unfairly discriminating between the sexes: it is classifying fairly because, for any patient, male or female, with or without the disease, it exhibits a stable (context-free) behaviour in the way it outputs its reading. However, that reading is not itself to be taken at face value as a robust indicator of risk, but must first be mathematically transformed, taking account of the relevant base rate, into the associated probability (which then *is* a robust indicator of risk).

The situation for prediction/prognosis is the exact reverse of the above. Now, instead of diagnosing a pre-existing disease, let us consider the task of predicting death (in say the next year), in the light of whether or not the patient currently has a certain dangerous condition- but taking no account of the patient's sex. Such a predictive approach could be considered "fair" (sexually non-discriminatory) if we found a stable proportion of deaths, for those exhibiting the condition, irrespective of the patients' sex (and likewise for those not exhibiting the condition.) But it would be simply inappropriate here to look at the situation backwards, and consider (for each sex) the proportion having the condition, in those who later do, or do not, die.[5] In contrast to the diagnostic example above, here the output of the method yields a probability for the actual

[4] This involves an application of "Bayes's theorem."

[5] Suppose we nevertheless did so, in a case where—equally for both sexes—the probability of death for those patients having [resp., not having] the condition was 95% [resp. 5%]; and 10% of men, and 1% of women, had the condition. Essentially the same mathematics as before would now show that the probability of having had the condition, in those who later die, will be 68% for a man, 17% for a woman. Just such an inequality constitutes the essence of the ProPublica criticism of COMPAS. However, here we can clearly see that it is merely an artifact of an irrelevant backwards analysis and the fact that more men than women have the condition; it is not an argument against fairness.

event of interest, which is directly relevant (and, ideally, unaffected by sex differences in the base rate) without any need for further processing.

The COMPAS system is predictive in nature, and its performance should therefore be judged according to the principles of the previous paragraph. It will be racially unbiased if the proportion eventually recidivating, for those assigned a given risk score, is essentially the same in both racial groups. Since this is indeed so, the ProPublica criticisms, which mistakenly analyse the problem as one of classification[6], are ill-founded.[7]

2.2. Accuracy: Are risk assessment algorithms too imprecise to be useful in individual cases?

In a series of papers, Hart *et al*. (2007); Cooke and Michie (2010); Hart and Cooke (2013) (henceforth HMC) contend that the statistical uncertainty attaching to the individual risk estimates of an ARAI is necessarily much too great for such instruments ever to be useful. While numerous objections have been raised to these claims (Harris *et al*. 2008; Hanson and Howard 2010; Imrey and Dawid 2015; Mossman 2015), the HMC criticisms, which are couched in a superficially convincing rhetorical style, have been highly influential (for example a special journal issue, Singh and Petrila (2013), has been devoted to HMC's arguments) and largely accepted as valid. But if they were valid, they would affect, not just the judicial system, but all the other areas, such as insurance and medicine, where risk prediction is used. That insurance companies have proven profitable for hundreds of years should raise some doubts as to the credibility of this critique. In fact, the HMC analyses are riddled with so many serious statistical misunderstandings, both of overall logic and of detailed technicalities, that they must be totally rejected.[8]

The major problem[9] with the HMC "analysis" is that it confuses:

(i). Uncertainty about a future event (recidivism of a given individual, Joe say), which is appropriately quantified as a probability, between 0 and 1; and

(ii). Uncertainty about the value of that probability, when it is estimated from a necessarily finite database, matched with Joe on a necessarily finite collection of

[6] Inappropriate conflation of problems of classification and problems of prediction is widespread. Thus Flores *et al*. (2016) say "We chose AUC-ROC as it is recognized as a standard measure in assessing diagnostic accuracy of risk assessments and has properties that make it not affected by base rate". But these "properties" hold only when the probability of receiving a certain score, conditional on actual outcome, is constant across different contexts and base rates. While this is a sensible requirement for a classification method, it is not so in a predictive context, where (ideally) it is the probability of the outcome, conditional on the score obtained, that is not affected by the base rate. Consequently, it is simply not meaningful to apply AUC-ROC or similar measures to a prediction problem. For more on this see Levy (2018).

[7] However, a separate analysis using the correct predictive approach (Corbett-Davies *et al.,* 2017) does identify bias in COMPAS - against women.

[8] Again, we are here simply arguing against bad statistics, and not prejudging the general issue of the usefulness of risk assessment algorithms.

[9] -among many others. See Imrey and Dawid (2015) for further deconstruction of the variety of spurious philosophical and mathematical arguments presented by these authors.

more or less relevant variables.

For example, in an attempt to explain their point that individual risks can never be precisely identified, Hart et al. (2007) write as follows, using what they term “confidence intervals”[10] (CIs) to describe uncertainty:

> To illustrate our use of Wilson’s method for determining group and individual margins of error, let us take an example. Suppose that Dealer, from an ordinary deck of cards, deals one to Player. If the card is a diamond, Player loses; but if the card is one of the other three suits, Player wins. After each deal, Dealer replaces the card and shuffles the deck. If Dealer and Player play 10 000 times, Player should be expected to win 75% of the time. Because the sample is so large, the margin of error for this group estimate is very small, with a 95% CI of 74–76% according to Wilson’s method. Put simply, Player can be 95% certain that he will win between 74 and 76% of the time. However, as the number of plays decreases, the margin of error gets larger. If Dealer and Player play 1000 times, Player still should expect to win 75% of the time, but the 95% CI increases to 72–78%; if they play only 100 times, the 95% CI increases to 66–82%. Finally, suppose we want to estimate the individual margin of error. For a single deal, the estimated probability of a win is still 75% but the 95% CI is 12–99%. The simplest interpretation of this result is that Player cannot be highly confident that he will win or lose on a given deal.

The situations considered here are all of type (i), concerning intrinsically uncertain future events: the outcomes of the game on some number (large or small) of future plays. However, since the probabilities involved are fully specified by the game, there is no uncertainty of type (ii). On 10 000, 1000 or 100 future deals, the actual success rate will vary randomly about its target value of 75%: the so-called“ confidence intervals” described for these cases are intended to give some idea of the possible extent of that type (i)random variation in the success rate. But on a single deal the actual success rate can only be 0 (which will be the case with probability 25%) or 100% (with probability 75%). This binary uncertainty is well described by these two probability values; but cannot be use- fully described by any “confidence interval,” let alone the above one of 12–99%, based on a totally misconceived and inappropriate formal application of Wilson’s formula. But all this discussion is in any case irrelevant to the point at issue, which is about the type (ii) accuracy we can achieve in estimating a risk. For that we would need an example where the probabilities were not known in advance, but were estimated from past data- for example, for predicting the outcome of the toss of an unfair coin, with an unknown bias towards heads. Given enough data on the results of past tosses of the coin, we can get a precise estimate of the probability of a head on the next toss, which is what we want. What we cannot get, or reasonably expect, is certainty as to the outcome of the next toss. HMC confusedly and wrongly conclude, from this obviously irreducible uncertainty in the outcome, that we cannot get a good estimate of the probability value itself.

[10] Their usage of this term is not in accordance with the standard statistical definition.

3. On the use in Court of algorithms embedded in software

Another way in which algorithms are now entering the justice system is in the form of scientific or technical software used in the production of legal evidence. For example, there are software programs that purport to analyse a complex DNA profile from a crime scene. This may be a mixture of DNA from several individuals, in varying quantities some or all of which may be minute. The software can be used in the investigative stages, and its output may be presented as evidence in court. A typical output will be a numerical probabilistic assessment of whether a specified individual was a contributor to the crime sample. There are currently at least 7 such systems in use: DNAmixtures and KinMix; Euroformix; likeLTD; Foresim; Lira; TrueAllele; and STRmix. They vary in the input data they use (*e.g.,* just the locations of alleles, or also the amounts of DNA seen there), the modelling assumptions made, and the analyses provided. Consequently different systems may yield different answers on the same data. While these could all be "correct" in their own terms, such discrepancies could obviously baffle a jury.

The following are some important considerations surrounding the use of software that can generate evidence for use in court or other judicial systems (*e.g.* immigration):

Availability Some systems are freely available, others are costly commercial products. For example, access to STRmix for purely academic purposes costs $6000 plus $1000 per annum (costs for casework application not disclosed). Such costs are a serious barrier, both to access by a defence team, and to academic investigation of properties and performance.

Transparency The details of commercial systems are typically secret. To allow fair assessment, there should be full and detailed documentation available illustrating the model, the algorithms used to implement it, as well as results from proficiency tests (see below). Courts should prioritise reasoned defence requests for disclosure of computer source code above the commercial interests of the supplier.[11]

Scientific basis The programs require scientific scrutiny to determine:

(i). Whether the methods are scientifically valid. This should include determining circumstances in which they may yield unreliable results.

(ii). Whether the software correctly implements the methods published.

(iii). Evidence needs to be given on the foundational validity of methods/algorithms across a broad number of possible settings.

(iv). Black-box methods should not be allowed.

[11] See Imwinkelried (2017) for a discussion of the tension between commercial confidentiality and legal disclosure, with particular reference to the use of TrueAllele in US criminal cases.

Uncertainty It is usually appropriate, and should then be required, to admit to and suitably describe the uncertainty inherent in conclusions. For DNA evidence this should be based on the likelihood ratio. Uncertainty quantification is especially needed for systems used to analyse pattern evidence, *e.g.* fingerprints, bitemarks, firearms, footprints, facial recognition, speech recognition, iris scans, and for trace evidence analysis, *e.g.* drug traces, gunshot residue, where this is not commonly done.

Validation The performance of the system needs to be validated and tested for reliability by appropriate validation studies. These should be conducted by external bodies. Models, software and data used both for validation studies and in real cases should be made available for examination and should be clearly explained in the report that is presented in court.

Conclusions

We have considered only a few of the many statistical issues relating to the behaviour and use of algorithms in the justice system. As highlighted by the examples in Section 2, this particular area seems to allow (and has allowed) great scope for generating dangerously misleading statistical analyses. A proper understanding of the statistical properties of algorithms, which is vital to their use and usefulness, requires the application of expert statistical knowledge and understanding. The Royal Statistical Society will be pleased to act as a source of advice in this area and will aim to be proactive and interactive in directing attention to situations where such advice is required (which may well not be obvious to non-experts).

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks. ProPublica, online at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barabas, C., Dinakar, K., Ito, J., Virza, M., and Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *Proceedings of Machine Learning Research*, **81**, 1–15.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidi- vism prediction instruments. *Big Data*, **5**, 153–63.

Cooke, D. J. and Michie, C. (2010). Limitations of diagnostic precision and predictive util- ity in the individual case: A challenge for forensic practice. *Law and Human Behavior*, **34**, 269–74.

Corbett-Davies, S., Pierson, E., Feller, A., Huq, A., and Goel, S. (2017). Making fair decisions with algorithms. Optimization and Fairness Symposium, UC Berkeley, 17 November 2017, online at https://samcorbettdavies.files.wordpress.com/2017/11/making-fair-decisions-with-algorithms.pdf

Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc. Research Department, online at https://www.documentcloud.org/documents/2998391-propublica-Commentary-Final-070616.html.

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And its biased against blacks." *Federal Probation*, **80**, 38–46.

Fry, H. (2018). *Hello World: How to be Human in the Age of the Machine*. Doubleday, London. "Justice" chapter.

Hanson, K. R. and Howard, P. D. (2010). Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior*, **34**, 275–81.

Harris, G. T., Rice, M. E., and Quinsey, V. L. (2008). Shall evidence-based risk assessment be abandoned? *British Journal of Psychiatry*, **192**, 154.

Hart, S. D. and Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences and the Law*, **31**, 81–102.

Hart, S. D., Michie, C., and Cooke, D. J. (2007). Precision of actuarial risk assessment instruments. Evaluating the 'margins of error' of group *v.* individual predictions of violence. *British Journal of Psychiatry*, **190, suppl. 49**, s60–5.

Imrey, P. B. and Dawid, A. P. (2015). A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy*, **2**, (1), e1029338, 1–18. DOI:10.1080/2330443X.2015.1029338.

Imwinkelried, E. J. (2017). Computer source code: A source of the growing controversy over the reliability of automated forensic techniques. *DePaul Law Review*, **66**, 97–132. Online at https://via.library.depaul.edu/law-review/vol66/iss1/6

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67, (ed. C. H. Papadimitriou), pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany.

Levy, D. (2018). In machine learning predictions for health care the confusion matrix is a matrix of confusion. *Statistical Thinking*, online at http://www.fharrell.com/post/mlconfusion/.

Mossman, D. (2015). From group data to useful probabilities: The relevance of actuarial risk assessment in individual instances. *Journal of the American Academy of Psychiatry and the Law Online*, **43**, (1), 93–102. http://www.jaapl.org/content/43/1/93.abstract.

Singh, J. P. and Petrila, J. (ed.) (2013). Special Issue: Methodological Issues in Measuring and Interpreting the Predictive Validity of Violence Risk Assessments. *Behavioral Sciences and the Law*, **31**, 1–164. Wiley