STATISTICAL LITERACY IN THE DATA SCIENCE WORKPLACE

ROBERT GRANT
Kingston University, St George's, University of London, BayesCamp Ltd
robert@bayescamp.com

ABSTRACT

Statistical literacy, the ability to understand and make use of statistical information including methods, has particular relevance in the age of data science, when complex analyses are undertaken by teams from diverse backgrounds. Not only is it essential to communicate to the consumers of information but also within the team. Writing from the perspective of a statistician who later taught himself about data visualisation and machine learning, I consider some pitfalls for communication and drivers of behaviour within the team. Recruiters and managers also play a part in creating a workplace where speed and novelty are sometimes over-valued. Statisticians have a duty to educate and shape this exciting new workplace.

Keywords: *Statistical literacy; Data science; Machine learning*

INTRODUCTION

You will probably have seen many definitions of data science (DS). I like one of unknown origin that is quoted by David Taylor in a post on the website KDnuggets: "Work that takes more programming skills than most statisticians have, and more statistics skills than a programmer has" (Taylor, 2016). This post then goes on to contrast and poke fun at a series of increasingly elaborate Venn diagrams that try to define DS (most of which are not Venn diagrams at all). There is some truth behind the humour. If we can generalise DS at all, it is the collaboration of people from statistics and machine learning / computer science (ML / CS) backgrounds. This means they tend to misunderstand one another somewhat, and the boss who has put that team together probably understands neither. Generally, there is a lot of hype and high expectations, as can be seen from the diagrams that throw in communication skills, domain expertise, and IT know-how and suggest that the ideal data scientist will be one person with all of these in abundance. A unicorn indeed (Taylor, 2016; Noyes, 2016)! I want to consider what statistical literacy might mean in this melting pot of clever people from different backgrounds. As a first step, it helps to understand the experience of people working around DS.

It might help to say briefly what my personal preferences are, so the reader can decide if I am just giving vent to biases. I studied mathematics and tried to ignore the obligatory computing courses, then studied statistics, and then more recently understood the value of those programming skills and taught myself some more. In the last couple of years, I have taught myself some machine learning methods. I still call myself a statistician, though I have bills to pay and might not hold out much longer. (Another of those DS definitions is "The field of people who decide to print 'Data Scientist' on their business cards to get a salary bump.")

Statistics Education Research Journal, 16(1), 17-21, http://iase-web.org/Publications.php?p=SERJ
© International Association for Statistical Education (IASE/ISI), May, 2017 – Invited essay

I use a lot of visualisations, including interactive online ones, some ML / CS techniques, and Bayesian modelling, so I am regarded as a bit of a fringe act by more conservative statisticians. I mostly work in health applications and interact with domain experts who have a little knowledge of statistics and regularly ask me to nudge, cajole and occasionally torture the data until the *p*-value pops under 0.05. For them, I provide disappointment as a service. I am like a dentist, who tells it straight and sometimes has to do stuff the client does not like at the time but will be better off for in the long run. I think ML / CS people are less inclined to disappoint like that; the expectation is on them to solve any problem in any setting. I worry about the crisis of replication and poor practice in science and feel that education is the real problem. I teach introductory classes, and try to reform the old-fashioned curriculum that starts with probability theorems and ends at ordinary least-squares regression. I think that a better schooling would start with regression, trees and other predictive methods and work backwards to the fundamentals, involve some philosophy of science about inference, substitute simulations for algebra, and – if the students would allow it – dump simple hypothesis tests entirely (after all, they are special cases of predictive models). Apparently I am not good at hiding my feelings, which is of concern every time a student says, “I just need to know what button to push in SPSS.” In other words, I picture myself modestly as close to the ideal data scientist and yet (or perhaps therefore) annoy everyone.

People often talk of tribes of science and their different cultures and languages. This is especially true of the statistics – ML / CS interface. Here are some metaphors I have heard used – some people like these but others are irritated by them.

- “Machine learning is punk rock statistics” – especially for the do-it-yourself ethos and supposed inclination to break the rules (source unknown).
- “ML / CS people are the cowboys to the statistical Indians” – I take this to refer to extant Hollywood stereotypes, not real people; it is not clear which group is being approved of, though the implication of being chased off one’s homeland caught my eye as a statistician (Yee Whye Teh of Google DeepMind and University of Oxford).
- “Statistics is a classical music education and machine learning is jazz” – ML / CS value innovation more, and are not so concerned about why things work or how long-lasting they might be (I made this up in 2015; countless variants could be made).

There is some truth to these differences. Anyone starting out in their career will learn norms of behaviour and aspiration from their colleagues and mentors but also from the publicly visible role models in their field. In statistics, these role models are likely to be established academics and to a lesser extent government (official) statisticians. In ML / CS they are often much younger, as online examples of your work with the latest methods has become a standard way of landing a job interview. Notably, though, lack of statistical literacy can appear in self-published spare-time projects and affect novice readers, for example, giving arbitrary integer codes to a nominal variable, representing these in binary notation, and then making each binary digit a dummy variable (McGinnis, 2015). Also, tech companies increasingly employ ‘evangelists’ and similar outreach-focussed job titles, who popularise their products with inspiring use cases.

Consider the incentives and rewards for these groups. The statistician working in academia, government, pharmaceuticals, etc., is not expected to be a great communicator or innovator, but they are expected to get things right every time. Correspondingly, there are plenty of sticks but few carrots, and they are incentivised to play it safe and not to waste time learning new methods. ML / CS people, by contrast, are rewarded for innovation and keeping abreast of the latest methods and computing tools. The workforce has greater churn than statisticians and so the repercussions of poor predictions from their models matter less. When these groups meet, there is potential for a clash. A DS team may also include web and graphic designers, bringing another layer of conflicting views.

Understanding DS is also a problem for employers. The shortage of data scientists is such that I get approached by recruiters quite often for obviously irrelevant jobs. This is a widespread problem – I am told – and shows how difficult it is for companies to identify and attract the right DS team, especially when recruiters do not understand what it entails. Once, the project was so unusual and high-profile that I went along and had an interview in the classic Silicon Valley mould (in that I spent many hours solving odd puzzles and being escorted to the toilet, but at no point did anyone say it was an interview, or that there was a job or a project or what the team wanted). This gave me a new appreciation of how hard it is from the company's perspective. The people I met were mostly programmers and clearly were unsure what to ask me or look for in my answers. I kept warning them that I was no programmer, and was reassured that they did not want a programmer, but they were obviously sworn to secrecy. Given all the mystery, my best guess is that they wanted the rest of the Venn diagram in one person, but in the end I clearly was not that three-quarter unicorn. I do not blame them for this; I think they had been sold the DS dream like so many other employers.

So, I think there are two aspects to statistical literacy in this setting: the familiar literacy among consumers, employers and clients, but also literacy among the DS team members themselves.

STATISTICAL LITERACY AMONG CONSUMERS, EMPLOYERS AND CLIENTS

There are some additional challenges for literacy in using the findings of a mixture of statistical and ML methods. Firstly, uncertainty is often ignored in many ML tools, and point estimates are presented alone. This may be because the tool optimises some loss function and does not provide a sampling or posterior distribution, but it is also the case that ML / CS courses simply do not teach this. Bootstrapping is the first-line choice for most complex inferences, and much research shows it to be easily understood by novices and hence the general public too. The bootstrap is well within the ability of a DS team, so it is surprising not to see it used more. However, fully Bayesian methods would be better suited when uncertainty comprises more than just sampling error. In some cases, such as the much-vaunted deep learning, fitting the model may take so long that running it again many times is simply out of the question. A current hot research topic is amending the model so that it provides uncertainty along with point estimates using the same optimisation algorithm, for example, Korattikara, Rathod, Murphy, and Welling (2015). The statistically literate user might wonder whether such a model has converged to an unbiased answer, but this is generally not discussed in polite DS company.

The lack of established asymptotic properties is a problem for many greedy and heuristic algorithms employed (often with great success) in DS. In such cases, what is presented as a result is often an acceptable local optimum, but again this is rarely acknowledged to the consumer. Some tools are presented as black boxes, where there is no simple formula to fit more predictions, and the only way to predict for future cases on the basis of the past is to re-run the program. In fact, there is always a formula, although it might be unfeasibly difficult to communicate. Regardless of its opacity, the literate consumer can always ask which observations in a test dataset are predicted well or poorly, and learn a lot from that. Perhaps because of the black boxes, it is common to hear talk of magic, wizardry, dark arts and such terms. Although they are used light-heartedly, they accumulate into damaging templates for people to think about (or not think about) DS methods. To be fair, statistics has long had the same problem; I have a standing policy (unenforced) of expelling from my office anyone who calls me a guru.

In many commercial settings, the management passes down questions to a DS team who are regarded as successful if they return a clear answer that can drive decisions, and do so quickly, possibly before the management meeting ends upstairs (Noyes, 2016). This is a situation unfamiliar to the slow-paced statisticians in the team and causes the team to become more gung-ho over time as cautious people go unrewarded and drift away. Really, it is the statistical literacy of management that needs to be tackled, and this starts with the DS team itself. If they keep creating a reassuring image of certainty, they will continue to be asked for it. Tom Davenport (2015) has helpfully identified the intermediary role of a translator, and divided DS people into "light" and "heavy quants".

The ubiquity of trade secrets has already been discussed, but this clearly contributes to a lack of understanding and hampers literacy. Interestingly, some widely used tools such as Google's TensorFlow and Amazon's DSSTNE have been released as open-source (Google Brain Team, 2016; Amazon, 2016). It is safe to assume that this happens when the companies have moved on and do not fear that the old product will give clues to the new one, but it helps to demystify the methods – at least with one step behind the current development. Finally, the high-dimensional nature of data sets and parameter spaces, alongside the black box, makes visualisation exceptionally difficult. An inspiring counter-example is TensorFlow Playground (Carter, Smilkov, Viegas, & Wattenberg, 2016).

STATISTICAL LITERACY IN THE DATA-SCIENCE TEAM

Colleagues of different backgrounds stand to learn a lot from each other in a DS team, not just about the mechanics of different analytical methods. Statisticians can teach ML / CS people about Bayesian and likelihood-based inference. They can debunk some of the reverence that ML / CS courses teach for eminently fallible methods like generalized linear models or principal components analysis. They can question assumptions of linearity and independence, and urge simple and fast methods in the place of the latest fashion; we have a wonderful history of make-do from the days before personal computers, and once again the data we have and the methods we use outstrip the capacity of our machinery. Statisticians have plenty of experience of exploratory data analysis and preliminary plotting to uncover structure in the data, which is ignored by ML / CS people in the belief that the method will find the best prediction from any starting point.

For example, neural networks will work much better with cleverly chosen functions of the predictor variables than by simply plugging in the raw data, but because of the black box, one can only discover this by trying it out. They can also bring Bayesian methods to bear on problems where second-hand data introduce biases and additional sources of uncertainty, and can cast doubt on work with poorly-defined research questions or that strays into multiple testing and related problems. On the other hand, there are some incredibly useful computer-intensive methods that are widely employed in ML / CS and not in statistics, such as *k*-fold cross-validation (for avoiding overfitting in models), random forests (a predictive method that averages across many decision trees), and boosting (combining many runs of a model with weighting that emphasises poorly fitted data). Statisticians should not expect to have a protected place in DS out of respect for their professional identity; we need to work for it and expand our own literacy.

CONCLUSIONS

Societies of statisticians and ML / CS people need to inform recruiters about successful DS teams, removing some of the corporate secrecy. It is in our own best interests. They could also provide some form of accreditation for individuals such as managers and recruiters who are peripheral but crucial to the success of DS. It may be possible to accredit whole organisations too, along the lines of ISO 9001 accreditation for management processes. As a profession, statistics also needs to start talking about and researching statistical literacy in this broader DS context.

We need to contribute to demystifying new methods. More data visualisation (dataviz) and method visualisation (methodviz) in the spirit of the TensorFlow Playground (Carter, et al., 2016) would be useful, and this should be user-tested on as wide a range of potential users as possible. Similarly, we need accessible, inspiring – and above all, short – books and websites for young people who are interested in DS as a career (of whom there are now many), written by experts from their own experiences.

Personally, I try to avoid grumpiness about having my territory encroached upon by youngsters who have never worked out the second derivative of a log-likelihood function, but at the same time not getting over-excited about how trendy and socially rewarding new ML methods are. It can feel galling to a Bayesian statistician – sometimes mocked as a cult – to see such mysterious methods as deep learning adopted unquestioningly by research funders, policy makers and business. I find it helpful to reflect on their incentives and other drivers of behaviour. But I urge everyone who has a background similar to mine to engage with this exciting trend; for statistically-trained people, an excellent starting point is the textbook by Hastie, Tibshirani, and Friedman (2013).

REFERENCES

Amazon (2016). Deep Scalable Sparse Tensor Network Engine (DSSTNE). A library for building deep learning. [Online: github.com/amznlabs/amazon-dsstne]

Carter, S., Smilkov, D., Viegas, F., Wattenberg, M. (2016). *TensorFlow Playground*. [Online: playground.tensorflow.org/]

Davenport, T. (2015). In praise of ‘light quants’ and ‘analytical translators’. *Features. StatsLife*. Royal Statistical Society. [Online: www.statslife.org.uk/features/2233-in-praise-of-light-quants-and-analytical-translators]

Google Brain Team (2016). *TensorFlow*. [Online: www.tensorflow.org/]

Hastie, T., Tibshirani, R., Friedman, J. (2013). *The elements of statistical learning: Data mining, inference and prediction*. 10th edition. New York, Berlin: Springer. [Online: statweb.stanford.edu/~tibs/ElemStatLearn/]

Korattikara, A., Rathod, V., Murphy, K., & Welling, M. (2015). *Bayesian dark knowledge*. (Pre-publication) [Online: arxiv.org/abs/1506.04416]

McGinnis, W. (2015). Beyond one-hot: an exploration of categorical variables. *KDNuggets News*. [Online: www.kdnuggets.com/2015/12/beyond-one-hot-exploration-categorical-variables.html]

Noyes, K. (2016). Why being a data scientist ‘feels like being a magician’. *PCWorld*. [Online: www.pcworld.com/article/3128320/why-being-a-data-scientist-feels-like-being-a-magician.html]

Taylor, D. (2016). Battle of the data science Venn diagrams. *KDNuggets News*. [Online: www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html]

ROBERT GRANT
Faculty of Health, Social Care & Education, St George’s Hospital,
London SW17 0RE
UK