How can an Artificial Intelligence do Statistics?

David Duvenaud
University of Cambridge
dkd23@cam.ac.uk

James Robert Lloyd
University of Cambridge
jrl44@cam.ac.uk

1 Overview of our work

In the McKinsey Global Institute report *Big data: The next frontier for innovation, competition, and productivity* it is claimed that

> "The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data."

With such a large potential demand for data analysis, it is natural to ask to what extent this analysis can be automated. In February 2013, statistician Andrew Gelman provided his thoughts on the question, *How can an artifical intelligence do statistics?* via his blog (http://andrewgelman.com/):

> "In the old-fashioned view of Bayesian data analysis as inference-within-a-supermodel, its simple enough: an AI (or a brain) just runs a Stan-like program to learn from the data and make predictions as necessary. But in a modern view of Bayesian data analysis—iterating the steps of model-building, inference-within-a-model, and model-checking—here, its not quite clear how the AI works. It needs not just an inference engine, but also a way to construct new models and a way to check models. Currently, those steps are performed by humans, but the AI would have to do it itself, without the aid of a homunculus to come up with new models or check the fit of existing ones. This philosophical quandary points to new statistical methods, for example a language-like approach to recursively creating new models from a specified list of distributions and transformations, and an automatic approach to checking model fit, based on some way of constructing quantities of interest and evaluating their discrepancies from simulated replications."

In summary, an artificial statistician would have to automate and iterate over the following three steps:

- Statistical model construction
- Inference
- Model checking

In addition, any automatic system would have to be able to report its findings in simplistic terms to have a wide impact. Automatic statistical inference is currently an active area of research known as probabilistic programming but as for the other procedures "Currently, those steps are performed by humans".

In our research we have begun to address the problem of automating the process described above. Following the work of [1] we have demonstrated an automatic model construction procedure for non-linear non-parametric regression [2]. This was achieved by defining an open-ended space of regression models via a generative grammar and then searching this space greedily.

We have subsequently demonstrated that the models produced by this system can be automatically described in simple natural-language in the form of statistical reports [3]. We provide examples of these reports; the final sections demonstrate our preliminary investigations into automatic model-checking using techniques described in [4].

2 Discussion

An optimistic view of our work was posted on Andrew Gelman's blog later in February 2013:

> "I feel so lucky to be around during this exciting era. Imagine being stuck with formalisms such as Wald's and Savage's hopeless attempts to shoehorn statistical reasoning into the formats of decision theory and game theory. Those guys were brilliant but they just didn't have the tools to do the job. Not that I think today's researchers have the last word, by any means, but it's so satisfying to see forward motion in modeling, computing, and also conceptual frameworks."

Perhaps a more measured view is that of our coauthor, Joshua Tenenbaum, also published on the blog:

> "These methods might seem computationally expensive ... perhaps compared to what people are used to when they build or fit only one or a small number of models. But when you consider the size and scope of the space of models that is searched, and the fact that all steps of model construction, evaluation and search are automatic, it doesn't seem like such an expensive process. In my experience, working statisticians, machine learners, and data scientists rarely if ever explore such a space so systematically in large part because it seems impractically expensive to do so (in terms of both their own time and computation time, as well as perhaps other scarce resources)."

We believe that this line of research could have an impact on any area that relies upon data analysis, including quantitative finance.

References

[1] R.B. Grosse, R. Salakhutdinov, and J.B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. In *Uncertainty in Artificial Intelligence*, 2012.

[2] David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, June 2013.

[3] James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of additive nonparametric models. In *Constructive Machine Learning Workshop, Advances in Neural Information Processing Systems*, December 2013.

[4] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.