NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

03-20

When is it Data Science and When is it Data Engineering?

Noel Cressie

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

When is it Data Science and when is it Data Engineering?

Noel Cressie
National Institute for Applied Statistics Research Australia
University of Wollongong, Australia
Email: ncressie@uow.edu.au

Abstract

Professor Bradley Efron was awarded the 2019 International Prize in Statistics and delivered his address on 19 August 2019 to the World Statistics Congress in Kuala Lumpur, Malaysia. This is a written discussion of his paper, "Prediction, Estimation, and Attribution," based on an invited discussion I gave in Kuala Lumpur immediately following Professor Efron's address.

1. Introduction

The International Prize in Statistics (IPS) is awarded by the International Statistical Institute (ISI) every two years, to individuals whose work has made a seismic breakthrough in Statistics. The IPS rivals in stature to the Fields Medal in Mathematics, although it is more recent, and the ISI is to be highly commended for initiating the prize. Brad Efron's *bootstrap* stands out among all the amazing contributions made in Statistics in the last 50 Years, and Brad himself is a singularity in the profession. He is known for his deep thinking about important science problems, elegant development of solutions with uncertainty quantified, and clear expositions of the statistical science. The IPS Laureate delivers a lecture at the World Statistics Congress (WSC). Brad's lecture at the 2019 WSC in Kuala Lumpur, Malaysia, was very forward-looking and concentrated on prediction (a more general concept than forecasting). Towards the end of my discussion, I move away from the specifics of Brad's very fine paper to a more philosophical place, where I suggest that prediction is looking more like Data Engineering than Data Science.

2. Prediction with an extra component of variation

I really liked the way Brad trichotomized statistical analysis into prediction (inference on random quantities), estimation (inference on fixed but unknown parameters), and attribution (includes hypothesis testing, model choice, and classification). I would like to add to this, a fourth "-ion," *decision* (in the presence of uncertainty), which I discuss in later sections.

Brad posed the prediction problem in terms of a "surface-plus-noise" paradigm, in which the surface is a latent process where the scientific truths reside. In his paper, he gives a decomposition that I go along with, until it stops short of what I need for spatial, temporal, and spatio-temporal prediction. In broad-brush strokes, his model for data y is:

$$y = s(x, \beta) + \epsilon, \tag{1}$$

where $s(x, \beta)$ is a "surface" and ϵ is (independent) "noise." I believe there is an extra component that should be included in (1), which is uncertainty in the representation of the scientific truths. That is, if the truths are represented by w (say), then the degree of veracity of the surface $s(x, \beta)$ is captured by the equation,

$$w = s(x, \beta) + \xi, \tag{2}$$

which is the "process model" in a hierarchical statistical model (HM). Importantly, y are observations on w not on $s(x, \beta)$:

$$y = w + \epsilon, \tag{3}$$

which is the "data model" in an HM. Combining (2) and (3) results in:

$$y = s(x, \beta) + \xi + \epsilon, \tag{4}$$

which reduces to (1) if the process-model error, ξ, is identically zero. Interpreted in terms of linear regression, where $s(x, \beta) = x^{'}\beta$, (4) tells us that the difference between the data and the surface is

$$y - x^{'}\beta = \xi + \epsilon,$$

and hence if $x^{'}\beta$ is a fixed effect the error budget is:

$$var(y) = var(\xi) + var(\epsilon). \tag{5}$$

Misinterpretations of uncertainties are possible by not recognizing equation (5): The budget may not add up to the total variance if $var(\epsilon)$ is determined from

independent calibration; or, if the noise-variance is assumed to be $var(y)$, over-smoothing will occur. In the model I use for spatial and spatio-temporal prediction, (4) is my staring point.

Usually, I want to predict w (where the scientific truths reside), for which the predictor,

$$\hat{w} \equiv E(w|y) = E(s(x,\beta)|y) + E(\xi|y),$$

is commonly used, along with its uncertainty, $var(w|y)$. In spatial, temporal, and spatio-temporal prediction, the process-model error ξ has statistical dependencies that contribute to both $E(w|y)$ and $var(w|y)$. Further, x could be random, a common example of which is the errors-in-variables problem. There are instances (e.g., when considering the output from numerical climate models) when it is the surface $s(x,\beta)$, not w, that one wants to predict. In this case, ξ is filtered out and one would use the predictor, $E(s(x,\beta)|y)$, with uncertainty, $var(s(x,\beta)|y)$.

The next section discusses how different queries about scientific truths in w usually lead to different optimal predictors.

3. Querying the predictand

The *predictand* is the random quantity w (or some function of it) to be predicted. In spatio-temporal statistics, w is often a stochastic process that is observed incompletely and noisily, resulting in data y. A *predictor* of w is a statistic, namely a function of the data y that here we write as $\delta(y)$. We are doing this statistical analysis for a reason, presumably so that we can answer a question about how our world works.

In the presence of uncertainty, there is no universal optimal predictor of w that answers every question about it, which I now discuss briefly. A prediction can be looked upon as a *decision* in a decision-theoretic context, and hence predictions are made based on cost-benefit (utility, U) and risk (probabilities conditioned on the data y). There are many possible predictions δ, and we need some way to compare them. The utility tells us how well we do if the predictand is w, and we use δ to predict it; we write the utility as $\mathrm{U}(w,\delta)$. We write the risk as $\Pr(w|y)$. Then weight the utility by the risk:

$$U(w,\delta)\Pr(w|y). \tag{6}$$

Recall that w is the random process we are trying to predict: In many of the problems I work on, w is a high-dimensional spatial or spatio-temporal random field for which I have incomplete and noisy data y. Then for each decision δ,

$\mathrm{U}(w, \delta)$ is an ensemble of utilities, where we can weight each one by its posterior probability (as we do in (6)). Integrating (6) yields the *posterior expected utility*,

$$\int U(w, \delta) \Pr(w|y) dw. \tag{7}$$

Then maximizing (7) with respect to δ over a given predictor space yields an optimal predictor δ^* that is a function of the data y.

Now suppose a different question is asked that involves making inference not on w but on a given function $g(w)$. A different utility function might be used, but suppose for the sake of argument that here it is the same. Then (6) is replaced with

$$U(g(w), \delta_g) \Pr(w|y), \tag{8}$$

where δ_g is a generic decision about $g(w)$. Integrating (8) yields,

$$\int U(g(w), \delta_g) \Pr(w|y) dw, \tag{9}$$

which when maximized yields δ_g^* that is a function of y. However, those who would like a simple life might base their prediction of $g(w)$ on

$$\tilde{\delta}_g(y) \equiv g(\delta^*(y)), \tag{10}$$

which is different from $\delta_g^*(y)$.

Now, if $\delta^*(y)$ is a great predictor of w, then wouldn't $g(\delta^*(y))$ be a great predictor of $g(w)$? Yes, if there were complete certainty in w but, modulo death and taxes,

- our world is uncertain;
- our attempts to explain the world (science) are uncertain;
- our measurements of our uncertain world are uncertain.

For example, in the simple case where w is a random variable, the commonly used utility that has a squared-error penalty yields the optimal predictor, $\delta^*(y) = E(w|y)$, of w. Now consider prediction of $g(w) = exp(w/a)$, where a is a known positive constant that has the same units as w. This problem occurs when, for example, soil or mineral properties are modeled on the log scale but inferences are needed back on the original scale. Then,

$$\delta_g^*(y) = E(exp(w/a)|y) \neq exp(E(w/a|y)) = g(\delta^*(y)),$$

because $g(\cdot) = exp(\cdot/a)$ is non-linear. In fact, because of Jensen's Inequality,

$$\delta_g^*(y) \geq g(\delta^*(y));$$

and note that $\delta_g^*(y)$ is unbiased for $g(w)$. Hence, the "shape-shifting" predictor $g(\delta^*(y))$ is consistently biased low for $g(w)$, has inferior posterior expected utility, and should be avoided despite its one-size-fits-all appeal.

4. What metrics are we using in the Twenty-First Century?

"What we measure affects what we do. If we have the wrong metrics, we will strive for the wrong things": Joseph Stiglitz, Nobel Laureate in Economics (*Financial Times*, 13 September 2009).

Statistics has sometimes been described as a discipline that translates variability into uncertainty. We have all seen how our careful attempts to get the probability space just right, so that the uncertainty can be expressed with auditorial rigor, are ignored or misinterpreted by scientists and policy-makers. Really fast computers can ingest really big data sets and predict what look like really interesting signals. But are the signals real? Scientists (including statistical scientists) like to assess their predictions with the LG (Looks Good) "metric," which works well when the signal is both strong and simple.

Statistical scientists can derive probabilities (denoted Pr(.)) for how different the predicted signal is from no signal. Statistics has "metrics" that include bias, variance, mean-squared (prediction) error, Pr(Type I error), Power = 1 – Pr(Type II error), p-value, and so forth. In simulation experiments where the signal is known and from which the simulated data are generated, or in cross-validations, we can obtain empirical measures (or metrics), such as the false positive rate, false negative rate, false discovery rate, and false non-discovery rate.

For example, the false positive rate can be interpreted as an estimate of the conditional probability, Pr(a signal is declared | no signal is present). Both the scientific method and the legal system are predicated on this type of probability, where there is a presumption of the *status quo* in Science or of innocence before the court. In statistical hypothesis testing, this is known as Pr(Type I error), and in the legal system it corresponds to Pr(conviction | innocent). Another metric is the false negative rate, where there is a presumption of a signal being present, that is, Pr(no signal is declared | signal is present). In statistical hypothesis testing, this is known as Pr(Type II error), and in the legal system it corresponds to Pr(acquittal | guilty).

To a software engineer, these metrics number among many that could be computed to evaluate an algorithm, but to a statistician the choice of metric depends

on the question being asked. The statistician is trained to look for variability over sub-populations of the evaluated metric which, amongst other things, might indicate prejudicial behavior of an algorithm. However, our training typically comes up short when an answer is needed to the following type of question: Is a false positive rate or a false negative rate of 2% equally acceptable for a population of 250 people as it is for a population of 25 million? This is an example of where the decision-theoretic machinery in Section 3 could be used to establish cost-benefit in the presence of uncertainty and then used to quantify the choices available.

5. Robo-debt: Australian government demands its citizens prove their innocence

The Australian social-services agency, Centrelink, has data on the earnings of Australians receiving various types of welfare payments. Its records are not perfect, and it looks for data from other sources, such as the Australian Tax Office (ATO), to provide a cross-check. If there is a discrepancy, a "please explain" notice is issued, at least that was the case up to the first half of 2016. In the middle of that year, a new scheme was introduced: It went from issuing about 20,000 notices annually to issuing 142,634 debt notices in the year following 1 July 2016, with a requirement that recipients prove there was no overpayment to them by Centrelink.

"Robo-debt," as it was dubbed by the media, was devised under a conservative Australian government, where Artificial Intelligence (AI) was used to catch unentitled welfare recipients and Human Intelligence (HI) was all but suspended. Clause 39 of Magna Carta Libertatum (signed in 1215) recognizes individual rights over and above the arbitrary will of a government, and it formed the basis of that part of the Fourteenth Amendment of the US Constitution that guaranteed "due process of law" to its citizens. This right to due process has, over the last 800 years, included the presumption of innocence, and it is the cornerstone of most democracies. In Australia, from mid-2016 and beyond, that presumption was suspended in approximately 700,000 cases, using an algorithm that was flawed and a scheme that had a presumption of guilt at its core.

Robo-debt was introduced to save the government money by removing HI and relying on AI and "big data." The robo-debt algorithm was based on matching the Centrelink database with an ATO database. An uneven wages profile, common in the gig economy, was treated differently by the two agencies, which prejudiced the casual and under-employed sector of the economy. The debt notice that was sent included instructions that failure to pay would be pursued by private debt collectors or by garnishing the individual's ATO tax returns. (I don't know of any credit notices being issued under the scheme.)

There was a right of appeal that involved a lengthy, cumbersome, and psychologically stressful process for a sector of the population that was already financially distressed. Moreover, even during the appeal, active collection of alleged debts proceeded. In media interviews, the government minister responsible quoted an approximately 1% error rate without clarifying exactly what that metric was. Subsequently, it was revealed that these "errors" were clerical data-entry errors and, in fact, the false positive rate of the scheme was on the order of 20%.

In the case of robo-debt, prediction failed because the algorithm was flawed and the metric was ill-defined. What the government did with that prediction was shocking and, in late 2019, the Australian Federal Court found the scheme to be illegal.

6. Science and Engineering

Broadly speaking, science tells us "why" and Engineering tells us "how." We keep them separate in our academies and colleges. Is this a left-over from the Twentieth Century? In our daily life, do we no longer care why something works, as long as it works?

Up to now, Engineering has been an important vehicle for society's uptake of Science's discoveries. Machine Learning, which looks for patterns in very big data sets and classifies attributes according to those patterns, is often (incorrectly) equated with AI, which will allow us to do everything in our lives faster, smarter, and cheaper... until it won't because we have been incorrectly classified!

An AI algorithm is trained on a "typical" data set, but it is almost certainly subject to unconscious bias that may or may not show up when applied to the general population. AI will tend to reinforce prejudices in the data in a manner similar to a Bayesian analysis where the posterior is used as a prior for the same or a similarly unrepresentative data set. Are the error rates spread evenly across all demographics and, if not, is it because of unconscious bias?

In conclusion, Brad's paper has embraced Data Science and forcefully argued for statistical thinking to be an essential component. The latter part of my discussion is purposely binary, so that as statisticians we continue to ask whether what we are doing is Data Science or Data Engineering.

Acknowledgment

This work was supported by an Australian Research Council Discovery Project DP190100180.