The Two Cultures of Data Science

By Peter Rush

Terminological Groundwork

This textbook page is quite conceptual and philosophical. It is about differing perspectives on data analysis, and the varied language you may encounter when interacting with data scientists from disparate fields.

Data science is highly interdisciplinary, and this has led to great variance in the terminology used by data scientists. Historically, distinct disciplines have used different terms or phrases for similar concepts. However, these divergent linguistic conventions go beyond language and reflect underlying perspectival and philosophical differences.

Statistical modelling is applied across a wide variety of research contexts, and with different purposes. The specific research context in which a data scientist works generally determines the language which they use, and this is linked to the purpose for which they use statistical modelling. For instance, a criminologist testing a theory of victimization might gravitate towards specific terms, whereas a computer vision researcher might gravitate towards others. The criminologist might say something like:

> "I am **fitting** a model in order to test for **significant** relationships between this set of **predictor variables** and my **outcome variable**."

The computer vision researcher might say something like:

> "I am **training** a model to recognise patterns in these **features**, in order to **classify** images as being either of faces or of natural scenes, based on these **labels**."

Both of these people are doing data analyses, but they view what they are doing from slightly different philosophical perspectives. The varying terminology they use reflects this essentially philosophical difference. The criminologist cares more about using their model to *accurately understand causal processes*. The computer vision researcher cares more about *how their model performs on the classification task*.

Let's think of these as different underlying *philosophies of data analysis*.

The Two Cultures

The statistician Leo Breiman identified these different terminologies and philosophies as falling broadly into two camps. One camp derives primarily from traditional statistics - e.g. statistics as it is practised in university statistics departments, and has filtered out into other departments which apply statistical methods, such as biological and social science. Here there is a focus on *explanation* - having a statistical model that captures some *true* aspects of the phenomena under study. Breiman called this the "Data Modelling Culture" - a data scientist applying this paradigm is trying to find a model

which is a good approximation of the *data-generating process* which exists in nature, and which produced the data they are analysing.

The second camp has arisen around *machine learning*, a much newer field than statistics. Here there is an emphasis on *predictive accuracy* e.g. making accurate predictions about new and unseen data. As Breiman puts it:

> "*In the mid-1980s two powerful new algorithms for fitting data became available: neural nets and decision trees. A new research community using these tools sprang up. Their goal was predictive accuracy. The community consisted of young computer scientists, physicists and engineers plus a few aging statisticians. They began using the new tools in working on complex prediction problems where it was obvious that data models were not applicable: speech recognition, image recognition, nonlinear time series prediction, handwriting recognition, prediction in financial markets.*"
>
> *(Breiman, 2001, Statistical Modelling: The Two Cultures)*

Breiman called this second culture the "Algorithmic Modelling Culture". This involves creating/selecting a learning rule, or algorithm, and then "training" the algorithm to achieve optimal performance on some predictive task. Here the emphasis is more on how the model *performs* rather than how well it accurately captures the true data-generating process.

At the time he was writing, in 2001, Breiman estimated that the classical methods of the Data Modelling Culture (i.e. traditional statistics) encompassed the work of 98% of statisticians, with only 2% using the methods from the Algorithmic Modelling Culture (i.e. machine learning). *Data science* as a discipline contains both of these cultures. And which culture a specific data scientist "lives in" is determined in part by the subject matter they study, the conventions in their field, and the purpose of their data analysis.

The prevalence of machine learning is increasing (especially in industrial applications), but classical statistical methods are still widely used in scientific research. Here are some general principles that apply when comparing the two cultures:

- Classical statistical methods are less computationally intensive than machine learning methods.
- Classical statistical methods give more interpretable parameters (e.g. "β is the expected change in $\hat{y}$ for a 1-unit change in the predictor variable....").
- Machine learning methods perform better with high dimensional datasets.
- Machine learning methods will (generally) outperform classical statistical methods for predictive accuracy.
- Machine learning methods (generally) give less interpretable model parameters.

Similarities between the Two Cultures

Despite their differences, there are also similarities between the two modelling cultures. To think about both points of similarity and of difference, let's describe a typical situation that a data scientist might be in, and see how methods from each of the two cultures could apply. Let's imagine that:

- We have a sample of observational units (these could be anything we want to study - people, cars, streets, rivers, geographical areas etc.).
- For each observational unit, we have scores on different variables:
    - In an ideal dataset each observational unit has a score on each variable (e.g. there is no missing data).
    - Our dataset also might include scores on each variable, for each observational unit, at different timepoints (e.g. time-series data).
- We can visualise our dataset in "dataframe space" - each row is one observational unit; each column is a different variable upon which each observational unit has a score. E.g. each row might contain the data for one person, with each column containing that person's score on a specific variable.

This is a very typical data analysis scenario. Normally, the analyst has a variable (or set of variables) that they are primarily interested in, and other variables which they think might be statistically related to the variable(s) of interest.

Pause for thought: do traditional statistical methods or machine learning methods seem more appropriate in this scenario?

Data scientists operating in both cultures will find themselves in situations like the one described above. The key difference between the two cultures, with reference to this situation, is what the data scientist wants to *do* with the data - or, you might say, how they want to *think* about the data.

For example, if the dataset contains variables related to e-cigarette use and lung health, we might care about uncovering the causal pathway which connects e-cigarette use and some lung pathology. In this case, traditional statistical methods may be more appropriate.

However, we might care about predicting whether a new person (not in the dataset) has the lung pathology, based on their scores on several other variables. In this case we do not care about *explaining* any causal aspects of the system which generated the data, we just care about getting good predictions. Here, machine learning methods may be more appropriate.

In either case, there are similarities in the model-fitting procedure. Typically, as we have seen in the last module, we define a cost function and adjust the parameters of our model until we find the smallest value of the cost function. (As mentioned previously, we want the cost function to be *cheap* i.e. have a small value). The cost function can be thought of as a measure of *error* – if it has a large value, then the model is not describing the patterns in the dataset well. The model which gives the lowest value of the cost function is the model which provides the best description of the patterns in the dataset, out of this class of candidate models.

With respect to cost functions, machine learning methods and classical statistical methods are similar: they both seek to achieve find a "good" set of parameters which give a (relatively) small value for the cost function. In some circumstances, you might see

classical statistical models like linear regression and logistic regression referred to *as* types of machine learning model. The difference is subtle, and relates to alternate views on the underlying philosophy and purpose of the data analysis, specifically regarding the following question:

*Do we care more about **explaining**, or about **predicting**?*

Concepts Exercise

We will now engage in a discussion-based exercise.

Run the cell below, it will print out a set of concepts.

Split into groups and write (on one person's laptop) definitions for each of the concepts. Try and link these to the ideas we have just discussed.

We (the instructors) will come around and discuss the concepts with you, and answer any questions you might have about them.

We will then print out some more or less simple and reasonable definitions, and will compare these to definitions you have generated.

Note: we have come across all of these concepts before, but do not be afraid to consult with google to get some clarification/input for the definitions.

In [1]:
```
# run this cell to print out some concepts
from concepts_exercise import concepts_exercise
concepts_exercise()
```

Observational unit

Artificial Intelligence

Machine Learning

Statistics

Statistical Model

Variable

Data Science

Population

Statistical Significance

Going forward

In later sessions on this course, we will cover some more advanced classical statistical methods which are generalizations of the linear model. We will look at the mechanics of these models, as well as how to use them responsibly. These generalizations extend the linear model framework to a variety of different types of outcome variable, and together these methods constitute a flexible toolkit for data analysis. Whilst we will focus mostly

on these generalized linear models, we will also cover some machine learning models towards the end of the course. It is true that machine learning methods will often produce better predictions than generalized linear models, but - assuming that a generalized linear model fits a given dataset *well enough* - generalized linear models have the advantage that they yield clearer and more interpretable parameter estimates, relative to machine learning models. Machine learning is a rapidly developing and exciting field, but interpretability is one of the many reasons that classical statistical methods still matter and are important to understand. If model fit/predictive performance between two models is *similar*, we would argue that a simpler, more interpretable model should be preferred *if your aim is explanation and inference*.

Similarly, you will encounter generalized linear models in research literature, so it is important to understand how they work when synthesizing research findings or talking about them with other data scientists. For example, below is the abstract from a recent biomedical science paper entitled *"A comprehensive analysis of all-cause and cause specific excess deaths in 30 countries during 2020*". This paper uses a generalized linear model for primary data analysis, because the aim of the research is to understand the causal processes at play in the phenomenon under study:

> "The impact of COVID-19 on mortality from specific causes of death remains poorly understood. This study analysed cause of - death data provided by the World Health Organization from 2011 to 2019 to estimate excess deaths in 2020 in 30 countries. Over-dispersed Poisson regression models were used to estimate the number of deaths that would have been expected if the pandemic had not occurred, separately for men and women. The models included year and age categories to account for temporal trends and changes in size and age structure of the populations. Excess deaths were calculated by subtracting observed deaths from expected ones. Our analysis revealed significant excess deaths from ischemic heart diseases (IHD) (in 10 countries), cerebrovascular diseases (CVD) (in 10 countries), and diabetes (in 19 countries). The majority of countries experienced excess mortality greater than 10%, including Mexico (+ 38·8% for IHD, + 34·9% for diabetes), Guatemala (+ 30·0% for IHD, + 10·2% for CVD, + 39·7% for diabetes), Cuba (+ 18·8% for diabetes), Brazil (+ 12·9% for diabetes), the USA (+ 15·1% for diabetes), Slovenia (+ 33·8% for diabetes), Poland (+ 30·2% for IHD, + 19·5% for CVD, + 26 1% for diabetes), Estonia (+ 26·9% for CVD, + 34·7% for diabetes), Bulgaria (+ 22·8% for IHD, + 11·4% for diabetes), Spain (+ 19·7% for diabetes), Italy (+ 18·0% for diabetes), Lithuania (+ 17·6% for diabetes), Finland (+ 13·2% for diabetes) and Georgia (+ 10·7% for IHD, + 19·0% for diabetes). In 2020, 22 out of 30 countries had a significant increase in total mortality." (from Alicandro, La Vecchia, Islam, & Pizzato, 2023. *A comprehensive analysis of all-cause and cause-specific excess deaths in 30 countries during 2020*. European Journal of Epidemiology)

Bibliography

Alicandro, G., La Vecchia, C., Islam, N., & Pizzato, M. (2023). A comprehensive analysis of all-cause and cause-specific excess deaths in 30 countries during 2020. European Journal of Epidemiology, 38(11), 1153-1164.

Breiman, L. (2001). Statistical modelling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus Machine Learning. Nature Methods, 15, 233-234.

Imbens, G., & Athey, S. (2021). Breiman's two cultures: A perspective from econometrics. Observational Studies, 7(1), 127-133.