The Art of Data Science

Mr. Krishna Balasubramanian
Claro Data Science, USA
krishna@clarodatascience.com

Data science is a 21st century term although some of the techniques it encompasses, like machine learning, have been around for decades. and some, like statistics, for centuries. The growth of computational power and storage capacity and the vast array of new personal computational devices and associated online services and social networks, monitoring devices, embedded processing in automobiles or manufacture, these have all contributed to an explosion of data crying out to be interpreted. Many stories remain untold and pose a teasing challenge to a generation of emerging data scientists. This article provides an overview of the data science landscape and subsequent articles will expand on interesting aspects.

What is Data Science?

So what constitutes data science? A blend of requirements gathering, product design, programming, data gathering, data cleansing, modeling through statistics or machine learning, visualization, and presentation of the analysis. The telling of a story based in data or the generation of complex reports for the analytically minded decision maker, understanding systems with incomplete information, the building of data product. This could happen on a laptop or take large clusters with many big data sources. It is a science. One looks at empirical data and builds models to explain the data and generate new probable scenarios: predictions or forecasts. It is potentially a new way of doing science - build a black box model that lives or dies based on its performance on a well-constructed test rather than a sound theoretical foundation. And thus, if one drops the theoretical demands, the barrier for entry into the field is significantly lowered and it is possible to make useful contributions by tinkering. Take a model, play with it and get an improved model (one that performs better on the test data). The proliferation of such tinkering within modeling frameworks has led data scientists to be likened to toddlers playing with Lego blocks. And perhaps with some justification as there are a number of results which get reported with a very scant understanding of the underlying scientific context. The democratization of learning (not unlike the democratization of journalism) comes with its share of problems.

Not surprisingly, most of the available data is observational data: data that was generated as a by-product of other activity rather than that generated by a well-designed experiment. Experiments can be expensive. When starting on a new project, it is often not apparent which lines of inquiry a given data set supports. The data may contain signal for some kinds of analysis and be utterly uncooperative on other fronts. So a data scientist must begin with an exploration of the available data. Which fields are complete and which have missing values, which have strong correlations or associations, which are potential target variables one would like to predict, which can be controlled to alter the processes, which are well understood by the business, which are likely to have significant errors ... Sometimes data scientists, like marketers, are pushed to squeeze water out of dry stone. The key to keeping it sciency is to document honestly the hunch or anecdotal bridges that were needed to complete the analysis so a future researcher may advance the state of the science. It usually pays to craft an interesting story from the data.

Trying too many things with the same data set can lead to what has been termed p-hacking in the hypothesis testing context. The fivethirtyeight blog provides a demonstration of what could happen if one goes fishing in the dataset instead of setting up a hypothesis first. But again some thought shows that the problem is not with trying too many things, but with the fuzziness of the metrics and mis-statement of the robustness of the result. If you find a hypothesis supported by fishing in the data set, you could still take random test/train splits of the data to verify its robustness. Your error estimates may end up optimistic because of leakage during the training process (unless you save off a test set from the start). So don't let the concern for p-hacking discourage you from exploring the data set but ensure that you report only robust results.

This article on wired.com provides a wonderful example of a data science project. A food recommender that uses collaborative filtering. that is, it gauges the visitor's preferences and recommends items that will align with their preferences at a particular restaurant by looking at what similar visitors prefer in that situation. A diverse set of attributes, unstructured data, ingredients and other similarities between foods defined by a chef and fed to the recommender system. One can hold out a test set of users and their food choices and see if the user's choices fall within the top 5 predictions made by the algorithm for that user. But then often the joy of a recommender system is whether it generate interesting recommendations: does it tell good stories of why something was recommended? An excellent place to learn all about recommender systems is this Coursera course.

The Data Science Process

Data science projects can start at very different points. Sometimes organizations have gathered data for years and want to answer specific questions about their customers or their suppliers. They are looking to optimize business processes. The data is already clean and stashed in relational stores so one can start with an exploration to understand it and talk to the data owners about the kind of analysis they want and get to modeling pretty quickly.

But more often than not, the data is not yet gathered or not clean or spread over different data stores which may not quite align and require flexible logic to connect them. And thus begins a larger exploration of ill-defined scope. One can start by exploring the needs, potential data sources, discuss the range of analysis different data sets might support and the associated costs. And then again, sometimes a data journalist or blogger goes hunting for a story that is hidden in the data and one starts by exploring potential angles.

Model Building

Machine learning models come in many flavors. Supervised learning methods include classification where we have a labeled data set and the aim is to assign labels reliably to new cases, and regressions where the labels are continuous variables. There are many types of classifiers and many ways of improving regression models. And then there are unsupervised models used for clustering or dimensionality reduction where the purpose is to discover structure in the data. Often one uses unsupervised methods to bring out important features in the data.

It is a well-known fact of the modeling process that complex models tend to have many adjustable parameters and therefore require more data to fit. They are also harder to explain. So the rule is to go for the simplest model that does a reasonable job as a basis for the solution. Sometimes one tinkers around that with more complex models which may be used to get better predictions without a good understanding of why they work. Also remember that a complex model can fit the training data set too well and then it does not generalize well to new data. Thus it is necessary to include some defense against overfitting such as regularization.

It is important to separate the test data set and ensure that training and tuning of the model happen within the training set. The test set is to be used solely for estimating model performance.

There is a lot that remains to be said about models and the process of data science, about bias in models, data governance, privacy but it is best to first get a sense of what modeling is about by building or exploring a few models. R2d3 provides a visual explanation of a decision tree classifier. You can build a model using the drag and drop interface provided by Azure, or if you like to code, try the datacamp scikit-learn tutorial.

Quotes:

"In God we trust; all others bring data." — W. Edwards Deming

"All models are wrong but some are useful." --- George Box

"If you torture the data long enough, it will confess." --- Ronald Coase

"Statistician -- The Sexiest Job in the 21st Century." --- Hal Varian

"Data Scientist -- The Sexiest Job in the 21st Century." --- Harvard Business Review

"Why Data Scientist is being called the Sexiest Job in the 21st Century" --- import.io

"The sexiest (and last?) job in the 21st century" -- techcrunch

About the author: Krishna Balasubramanian is a member of Claro Data Science, a company which provides data science solutions for small businesses. He has twenty years of experience working in start-ups in the online and television advertising space, and has been building machine learning applications for ad delivery, audience segmentation, targeting and optimization. He taught a data science bootcamp and a Deep Learning course while at Metis/Kaplan. At Claro he has been working on projects in data privacy and NLP.