Opinion

The role of feature engineering in a machine-learning world

Opinion

Artificial Intelligence (AI) continues to be the next great topic of debate. In fact, Microsoft, Amazon, IBM, Google and Face book announced on Thursday, Sept 29 the formation of the Partnership on Artificial Intelligence to Benefit People and Society. Within the predictive analytics discipline, though, we tend to use the term "machine learning" as our reference point for artificial intelligence. Much of our thinking in this area has focused around the role of the practitioner or craftsman versus the machine and the concept of machine learning. Yet, machine learning has now evolved into the usage of higher levels of mathematics and computer science with the most recent level being deep learning.

Machine learning vs. deep learning

When the term "machine learning" became a more "in vogue" term within the business community, discussion arose about the power of the machine replacing the practitioner. If one googles the term "machine learning", one will observe a variety of very familiar techniques such as multiple regressions, logistic regression, decision trees, neural nets, etc. These techniques have existed for decades and have never posed threats to the role of the predictive analytics practitioner. In fact, technology has facilitated the utilization of these techniques which has essentially increased the demand for these type of practitioners. The demand for practitioners in fact has grown dramatically due to the increasing need for human people to apply their intellectual capital in adopting these technologies along with the utilization of data in creating the analytical file to solve the business problem at hand. This has always been at the core of predictive analytics.

But deep learning is now the latest term to enter the vernacular of our industry and once again the familiar refrain regarding the exact role of the practitioner remains. But how does deep learning differ from machine learning. Based on much of the literature, deep learning is a higher level form of machine learning as the machine has much higher self-learning or cognitive type capabilities. But even with all these definitions, what are the implications for the practitioner. Specifically, what are the functions that can be automated by the machine versus those that still require human intellectual capital. In the predictive analytics process, the concept of feature engineering or variable selection is arguably the most important part of the process. The push to automation has included more of these feature engineering techniques which can be used within this process. For example, the practitioner can use techniques such as factor analysis, decision trees, correlations, etc. as mathematical routines to aid in the feature engineering process. Previous articles have discussed the merits and advantages of each of these techniques. But in the Big Data era, we potentially now have hundreds of millions of records and thousands of potential variables. If one thinks about fraud or credit loss, a .5% increase in lift can achieve millions of dollars in profitability due to large volumes of activity. The traditional techniques discussed above may simply be too onerous and inefficient within the feature engineering process. Can the advancements achieved in deep learning aid in this process.

Volume 4 Issue 2 - 2016

Richard Boire
Environics Analytics, Canada

Correspondence: Richard Boire, Environics Analytics, 33 Bloor Street East Suite 400, Toronto, ON M4W 3H1, Canada, Email Richard.boire@environicsanalytics.ca

Received: October 16, 2016 | **Published:** October 26, 2016

Advancements in technology

Yet, in understanding these advancements, it is important to understand what has transpired within data processing technology. Obviously, the biggest advancement has been the shift towards "distributed" processing from "sequential" processing which was pioneered by Google in being able to develop their search engine technology. This type of technology is excellent if one is simply designing reports or creating visualization tools from the data. But if one is developing predictive analytics solutions, other technologies are required such as in-memory processing which essentially speeds up the processing of complex mathematical calculations. Many of the leading software providers in advanced analytics offer this functionality within their products. A good example of the benefits of this technology is the development of fraud models at one major U.S. bank whereby just the data processing component of the model consumed 36 hours. With in-memory processing, the time was reduced to less than 90 minutes. Through this functionality, the practitioner can in effect utilize these more traditional type techniques with in-memory processing but not have to undergo long delays while awaiting the results of a given technique.

Addressing the analytical challenges with deep learning

Having addressed the data processing problem, the second challenge is analytical as we need to address the question of whether or not we are overlooking specific trends or patterns within this mass volume of data. Through the effective use of neural nets technology, deep learning can detect patterns which were otherwise seemingly overlooked. Historically, certain key variables or features were overlooked as the mathematics was trying to fit the data to a given distribution. In the case of neural nets, there is not that bias as the technology will detect linear and non linear patterns in the data. Of course, the "knock" against neural nets is that the output is hard to explain in terms that a given business stakeholder will understand. Yet, these features, if difficult to explain, produce better solutions, do we really care about explanation in terms of the features which are being used in the solution? Explanation may be irrelevant if solutions are working but what happens when they fail? Validation of these features becomes a real challenge given their complexity due to the non linear nature of neural net technology. How do we open up the hood of these technologies to really understand what worked vs. what did not work.

In our increasing world of Big Data, do we simply trust that the machine is doing the right thing? As our discipline advances, newer algorithms and advancements in deep learning will be developed that offer better and more efficient ways in building a predictive analytics solution. But how do we effectively determine that this is indeed the case. For many seasoned practitioners, the traditional or older techniques are still the preferred techniques due to both performance lift and explainability to the business stakeholders. But as advancements continue to push the boundaries of deep learning, the practitioner's toolkit will continue to expand. Practitioners need to be willing to try these new techniques but within an environment that analyzes whether or not there is indeed any incremental impact in lift. Once these models are deployed, we can certainly determine the effectiveness of a model but our real challenge will be to identify the causes and reasons for when a given model fails to deliver its expected results. For instance, can we evaluate the inputs that went into a given model? Certainly, the literature is very sparse in this area in examining model failures that use very advanced techniques. Predictive analytics practitioners need to consider all tools in their arsenal but a certain level of bias will set in unless we can understand how to examine these advanced models in more detail. Otherwise, it remains a black box, which is simply less likely to be used by practitioners. In a sense, the real challenge with the more recent advancements in machine learning is to reduce the "black box" nature of its solutions thereby increasing the comfort level of practitioners in using these techniques.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

Citation: Boire R. The role of feature engineering in a machine-learning world. *MOJ Proteomics Bioinform.* 2016;4(2):204–205. DOI: 10.15406/mojpb.2016.04.00117