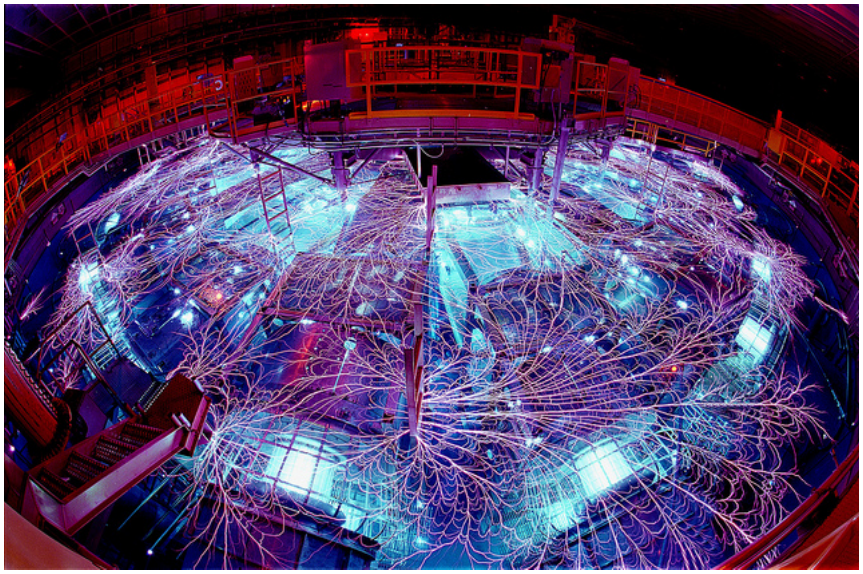
TECHNOLOGY / GLOBAL

What Machine Learning Can and Can't Do

26 Mar 2015 5:00am, by Mark Boyd



+

As machine learning products continue to target the enterprise, they are diverging into two channels: those that are becoming increasingly meta in order to use machine learning itself to improve machine learning predictive capacity; and those that focus on becoming more granular by addressing specific problems facing specific verticals.

And while the latest batch of machine learning products across both these channels may reduce some pain points for data science in the business environment, experts warn that machine learning can't solve two issues regardless of the predictive capacity of the new tools:

- Solving unique problems for a particular business use case, and
- Cleaning the data in the first place so that it is valuable in a machine learning workflow.

Machine Learning Tackles Context

Last year, new machine learning market entrants focused on speeding up processes around mapping the context that a machine learning algorithm would need to understand in order to predict needs in a given business situation. For example, if a voice translation machine learning product was listening in to a customer service call in order to more quickly help the call operator surface the appropriate solution-based content, the first job of the machine learning product would be to create an ontology that understands the customer call context: things like product codes, industry-specific language, brand items and other niche vocabulary.

Products like MindMeld and MonkeyLearn built automatic ontology-creators so the resulting machine learning algorithm had a higher degree of accuracy without the end user first having to enter a whole heap of business-specific data into the product to make it work. Others, like Lingo24, created their own specific vertically-based machine learning engines for industries like banking and IT so that their machine learning translation service could apply the right phrase model to the right situation.

The people developing those products recognized that to be accurate, even off-the-shelf machine learning products require a lot of customization and data science leg work to be an effective tool in any given business use case.

Now, in the first quarter of this year, the latest generation of machine learning tools are aiming to speed up the next bottleneck in the machine learning and predictive analytics pathway: speeding up the process of data modeling for data science in general, and solving particular pain points for particular verticals.

Machine Learning for Data Modeling

The data modeling stage often requires data scientists to iterate multiple data models and run them against historical datasets in order to identify the most accurate predictive models. The process is so slow and cumbersome that a Reddit Q&A sought out productivity hacks for how to use the downtime while waiting for a machine learning model test to be completed (fitness was surprisingly popular as a way to fill in the time: pushups, stretches, or batching up enough data modeling jobs to allow time to get out of the office and go rock climbing were all popular responses.)

Last month, Skytree released Skytree Infinity 15.1 aimed at automating data modeling processes, while also analyzing when it is best to run big data machine learning activities.

"In data science, creating models is an iterative process," said Martin Hack, chief product officer at Skytree. "You create models, run them, compare the results against historical accuracy, and then put the most accurate into production. So there are usually three steps: train, tune and test. What we have done is combine this into one. It's potentially a huge time-saver for data scientists, and reduces time-to-market for data models."

The new feature in Skytree's latest version provides an auto-modeling tool. Users set their optimal parameters and Skytree will do all the iterative data modeling itself until a single data model emerges with the most consistent accuracy.

> "On any given day, our customers might have been producing hundreds or thousands of models," Hack said.

The feature was created in conjunction with existing customers who had an early version of the software. Hack confirms the auto-modeling feature was tested for business cases including fraud detection, determining and reducing insurance rates, and in marketing applications for the segmentation and scoring of customers.

Machine Learning for Knowing When to Run Data Models

Skytree's new release also includes a feature aimed at predicting the computing resource costs of actually running large-scale machine learning data model experiments.

As data models draw on ever-expanding volumes of data, Hack believes the need to use machine learning to understand the costs of the modeling process will help enterprise decide where the right payoff is:

"Our model management tools record everything: What processes have I done? How was the performance on a specific model as it evolved through the data science process? We call this essential model quality, and you absolutely want to be able to see what resources the data model application is using, all the way down to the CPU changes." Hack adds:

> Computation and data science can go hand in hand. Ultimately you are going to see a model view and which model worked best and how much resources each model is using. Even Hadoop itself is realizing it needs to have more allocation-aware/resource-aware systems.

Machine Learning for Serving Content

Cloud application delivery service Instart Logic recently released their latest product, which they say is the industry's first machine learning product aimed at speeding up web applications.

Their SmartSequence tool optimizes how HTML and JavaScript code should be loaded in web browsers and mobile devices. SmartSequence is an algorithm that determines the optimal number of samples required to collect and analyze the required code/content to be delivered for optimal performance. The approach is also horizontally scalable, and the expansion on resources will be similar to adding additional hardware capacity when traffic increases.

SmartSequence collates data on a customer's web application usage, and then starts figuring out how to improve performance. "It depends on the type of code that the SmartSequence system is processing [HTML or JavaScript], but to get started we need to generally see between 6 to 12 requests for the object through our system," explains Peter Blum, vice president of product management. "The request is going to result in some back-end analysis of the code itself plus information we get back from the real consumption of that code, by end users' browsers."

Blum said once the algorithm samples some actual requests, it starts getting smarter and can notice when the end user's behavior patterns change.

To create the machine learning tool, Blum draws on a data tech stack as well as their own created tools: “We use a number of existing solutions such as R, MatLab, Hadoop and Hive, but for the production implementation we ended up building some of our own technology around this due to the specific use case and the fact that it’s a core part of our distributed architecture. We do make a point of adopting existing open source technology into our solutions as part of our service.”

Blum also said Instart Logic has built-in architecture to minimize the computing resources required when running the SmartSequence algorithm.

At a high level, the company has a cloud-client architecture, Blum said. The cloud is a set of globally distributed serving locations. And then the client component of the Instart Logic solution is a thin JavaScript-based virtualization client that injects automatically into a customers’ web pages as they flow through the system.

The client side component is responsible for measurement and monitoring, Blum said. It can, for instance, gain an understanding of how the code is consumed and executed by the end users’ browsers. It beacons this information back to the cloud portion of the service for analysis and learning. It is learning across a subset of the website loads.

On the cloud side, the company has a tiered system with essentially a full proxy that will send and receive data between the service and the end users’ browsers, and will also communicate with customers’ backend web server infrastructure. That’s where the SmartSequence technology lives. A compute tier receives the profile information from the end users’ browsers and does all the analysis and learning. Once SmartSequence comes up with the right optimization of code it passes this over to the full proxy tier, so that future requests benefit from the learning around what code to send up front versus what only needs to be sent as needed.

> Our system is much more compute intensive than a traditional web delivery service, so we have deployed more raw compute as part of our architecture.

Machine Learning for Predicting Problems

In the same way that Instart Logic is using machine learning to solve a particular problem — load time for web applications — cloud-based analytics service Sumo Logic is using machine learning for a similar pain point: to identify potential outliers from web engagement metrics in order to ward off potential future problems.

Unlike last year’s big machine learning plays by startups taking on text mining, voice recognition or language translation, this year’s machine learning products are more granularly focused on being a component tool within a larger workflow.

Sumo Logic said their outlier detection and predictive analytics features are focused on identifying pattern anomalies in large sets of unstructured data from both machine logs and user behavior on websites and mobile applications.

Sumo Logic starts with pattern recognition: the company looks for signatures in unstructured data and slims down the results to sizes that humans would need to look at to understand what is happening, said Sahir Azam, director of product management.

Customers are often parsing out the log data and looking at specific values, such as response time of an application, and then trying to understand the ups and downs of that metric, said Azam. Traditionally, when reviewing large amounts of machine and unstructured data for outliers, data scientists have had to set static thresholds that are either too high to identify abnormalities, or so low that there is too much noise in the system to bother trying to understand each outlier as it happens. Sumo Logic’s new feature is able to build a “band of normalcy” that accounts for seasonal variation and helps create a multi-dimensional view.

“Our big innovation is we can take this stream of data across these microservices and run them on aggregated data. The technical capability is broad based, it can be applied anywhere. Where we see the most value is the mission-critical customer-facing apps. We really aim to solve a problem for the DevOps teams and the line of business app owner.

In addition to the outlier detection tool, the predictive analytics feature then uses that machine learning to project where these trends will head in the future if left untouched, Azam said. Sumo Logic’s predictive analytics is a sister operator that will take that outlier trend and use linear progression to look at what might happen in the future. For example, trends in reduction in sales on an e-commerce site might actually be an early warning sign of latency problems.

A big use case so far is among security and compliance officers that need to detect IP addresses that are scraping website content regularly to create competitive sites, said Azam.

Other users are more traditional IT Ops administrators who are learning how to tune the SumoLogic feature to suit their business use case. Said Azam: “No machine learning is perfect. We can reduce a lot of the noise and get the visibility up, and tune the analytics for a particular application. But latency and security abnormalities vary from use case to use case and customer to customer. These tools allow a customer to get more customized. That adds a bit of time but they are getting much lower false positives.”

What Machine Learning Can’t Do: Clean the Data

But while machine learning may be helping speed up some of the grunt work of data science, helping businesses detect risks, identifying opportunities or delivering better services, the tools won’t address much of the data science shortage. At the end of the day, business users will still need a data scientist on their team to make the most of the tools, said Alon Bartur from Trifacta and machine learning author, Louis Dorard.

Alon Bartur, product manager at data transformation service Trifacta, said the main stumbling block for many enterprises wanting to start using off-the-shelf machine learning tools is the quality of the data to start with. “You need to make sure the data is correctly structured. You need to identify any biases that might exist. If you haven’t had a look at the data yourself, then you cannot take the right action,” he cautions.

Whoever is feeding this data into these tools, they still need to have confidence that the data is clean, free of biases and free of anomalies, Bartur said. That work still has to be done, whether it is done by the person who is building the data models or someone else. It still takes a critical eye to see what to ask the data and have tools that enable the user to generate models faster and help get results faster.

Bartur said that as businesses adopt multiple machine learning tools to assess data at various stages of a business process or for a particular task, they may need to restructure their data into the format suited to that machine learning tool.

A lot of people struggle with cleaning the data, Bartur said. Then it may be cleaned, but it may need to be in a different format in order to run it through a machine learning tool. As the volume of sources is increasing, this becomes more of a problem. There tends to be a bottleneck in cleaning the data and preparing it.

Bartur gives an example from the big data enterprise market:

> What we are seeing in the Hadoop market is that people are thinking Hadoop is the solution. They are seeing more sources of data, asking more questions of that data, and then finding the structure is too rigid to be able to get the analysis they want.

As the user look at the analytics lifecycle, there are levels of maturity: it starts with an exploratory question or hunch, and that leads to an interesting question, so it requires an investment in a centralized data view, which in turn enables more of this exploratory work. You allow a business to work out what is valuable to it.

More peopele are getting creative about their data, Bartur said. There is a maturity curve that people go along as they discover new ways of looking at it. The key is to get people to think about data in a more creative way than seeing it as a rigid model, he said.

What Machine Learning Can’t Do: Leap Over Pareto’s Principle

Author of Bootstrapping Machine Learning, Louis Dorard, said the latest generation of machine learning tools are akin to the Web of the early 2000s: “With web development, you used to have to know HTML, CSS and JavaScript. So you would need a developer that could create those websites. It was very difficult to meet that demand. But then along came WordPress, and almost anyone can use it, and it works in 80 percent of the cases, but the rest of the time you need developers. The same 80/20 rule applies to data science.

“There are going to be customers for whom these products will work, and in 20 percent of the more delicate work you will need access to a data scientist,” Dorard said.

Dorard sees this as one of the main reasons why products like Instart Logic are trying to solve a specific problem. “If many companies have the same needs, then these solutions are going to cater to these needs, but if you are doing something a bit more funny and not that usual, you are going to have to come up with your own solution.”

The McKinsey Global Institute argues that data analytics is emerging at the forefront as the competitive advantage of any business, driving productivity, growth and innovation. They warn of shortage in the U.S. alone of close to 200,000 data scientists and up to 1.5

million managers and analysts confident in making decisions based on data supply. New machine learning tools may relieve some of the burden from either laborious data science processes (like Skytree) or handle 80 percent of the workload (like Instart Logic or Sumo Logic), but data science will still be in strong demand to prepare data in the first place and to get the full value of the new tools on offer.

Feature image via Flickr Creative Commons.

HADOOP INSTART LOGIC MACHINE LEARNING PREDICTIVE ANALYTICS SKYTREE SUMOLOGIC

+

RELATED STORIES

ANALYSIS / EVENTS / GLOBAL

The New Science of Building Successful Data-Driven Apps

7 Jul 2017 10:28am, by Joab Jackson

ANALYSIS / TECHNOLOGY / TOP STORIES / GLOBAL

Salesforce's Einstein Mixes Automated AI with Business-Specific Data Models

7 Jul 2017 3:00am, by Mary Branscombe

View / Add Comments

SPONSORED FEED