5-Minute Guide

MACHINE LEARNING

ATTIVIO®

ATTIVIO®

5-Minute Guide to

Machine Learning

Discussions of machine learning can quickly get lost in the weeds of computational statistics and other arcane disciplines. But here's a simple definition of machine learning by Arthur Samuel, who coined the term while working at IBM. Machine learning is the "field of study that gives computers the ability to learn without being explicitly programmed." [2]

***Machines can learn too.** Central to machine learning is the idea that with each iteration, the algorithm will learn from the data.*[1]

Samuel used games—specifically checkers—to test his theories about machine learning. His checkers playing application is considered the first self-learning program. [3] Leslie Valiant, a British computer scientist, also contributed to the theory of machine learning with his concept of "probably approximately correct" (PAC) learning. [4]

Machine Learning is Everywhere

Tech blogger Jeffrey Walker makes the point that we're already surrounded by examples of machine learning, such as Google's page ranking system, photo tagging on Facebook, customized product recommendations from Amazon, and automatic spam filtering on Gmail. [5] He also predicts that machine learning will further accelerate the move away from relational databases. Walker cites the rise of triplestore databases that can much more easily handle the types of complex queries processed via machine learning. [6]

1 http://www.sas.com/content/dam/SAS/en_us/doc/conclusionpaper1/statistics-machine-learning-at-scale-107284.pdf
2 https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf
3 https://en.wikipedia.org/wiki/Arthur_Samuel
4 Ibid.
5 https://www.monitis.com/blog/7-reasons-why-you-need-to-get-onboard-with-machine-learning/
6 Ibid.

How Does AI-Powered Search Use Machine Learning?

A lot of cooks stir the broth that is AI-powered search. The executive chef in this analogy is text analytics. And one of the most important sous chefs is machine learning.

It's all about relevance. AI-powered search solutions provide increasingly relevant answers to implicit and explicit queries by combining self-learning technologies such as indexing, natural language processing, and machine-learning.

All search at some point can be reduced to math. But understanding relevance based on context and intent is a math problem of a much higher order than simply returning all the documents that contain a keyword or phrase. And that's where machine learning comes in.

Users want more than just relevant results from their search queries. They want their search tool to "understand" what their queries mean based on context. In other words, know the difference between what was expressed in the query and what was intended. Machine learning helps accomplish this in two ways.

Enriching Data

Data enrichment involves two types of machine learning—batch based (supervised) and online (unsupervised).

Batch-based learning

Batch-based data enrichment starts with a model. For example, batch-based machine learning can automate classification of document types. If you want to distinguish between an inter-departmental email, an HR memo, a newsletter, and a trade request, you create a sample set of each document type. The larger the sample set, the more accurate the model.

Within each class of documents, the machine learning application looks for the key words, visual patterns, and metadata that distinguish one document from another. To test the accuracy of the model, the application runs against a second set of sample documents. When the application makes a classification error, the model can be corrected manually, increasing accuracy.

***Search—the new data curation.** Steven Nicolaou, principal consultant with Microsoft, notes that as the cost of storage decreases and data volumes soar, it's cheaper (and faster) to use search rather than data curation to do the heavy lifting of finding the right data.[7] And the technology that does a lot of that heavy lifting is machine learning.*

Online learning

But suppose you don't have examples of what you want the computer to detect? Online or unsupervised machine learning lets the computer build a model as it sifts through the data—learning as it goes.

Data scientists often apply this type of machine learning to outlier detection and predictive analytics. Outlier detection is still a form of classification, one in which the "none-of-the-above" category is the most valuable, especially in risk and compliance applications. With predictive analytics, the computer tries to infer what may happen in the future based on the data it sees now.

In both types of machine learning, there is a constant feedback loop between inputs and outputs.

"The fundamental goal of machine learning is to generalize beyond the examples in the training set."[8] **Pedro Domingos,** Computer Scientist

Everything that's learned about a piece of content is added as metadata, which makes it valuable and accessible for search.

Financial services institutions have been among the most enthusiastic adopters of AI-powered search-based applications. Machine learning helps these applications manage and monitor the volume and ambiguity often associated with global financial transactions and the communications that accompany them.

Increasing Relevancy

Machine learning applied to relevancy helps users get the best results for a search query. It can support a relevancy model with thousands of parameters, which goes far beyond the capacity of any individual or group of individuals.

Machine learning programs can track what people search for and what they click on and then match that to the parameters of the relevancy model. As data accumulates, the machine learning program can reverse engineer an algorithm to match the accuracy of the results. With continued click tracking, more data feeds back into the model or the model builder, which creates greater relevancy downstream.

7 http://www.docurated.com/enterprise-search/enterprise-search-14-industry-experts-predict-future-search6 Ibid.
8 http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

"Developing the right model to fit the data is like Goldilocks. We want the fit to be not too much, not too little, but just right."

- Wayne Thompson,
Manager of Data Sciences
Technologies, SAS

Driven by machine learning, relevancy continually improves and users are more satisfied, which promotes engagement and productivity. Moreover, machine learning can easily support multiple relevancy models. For example, you could train one model based on input from employees in HR and another from those in sales. Each model would incorporate a different set of preferences that could be refined over time.

We're the Teachers

At its core, AI-powered search is a document-driven decision support system. And, whether it's for an internal search capability within the enterprise or a commercial application such as product matching, AI-powered search relies on machine learning to mimic "human cognitive capabilities like perception, language processing and visual processing." [9]

Paradoxically, machine learning is really no different than the way humans learn. We learn from our mistakes (acquire data inputs); modify our actions accordingly (tune the model), and proceed (run the model again). So, the next time you run a search and quickly find just what you were looking for, remember—you're also teaching a machine how to do its job.

ATTIVIO®

Attivio is the leading AI-powered search & insight platform company. Our mission is to put search at the core of every enterprise. By easily integrating every data source into one always-learning platform, Attivio enables Fortune 500 enterprises to answer the most complex questions, by providing immediate insight across all of their information. Our solutions power innovation, drive operational efficiencies, and transform business outcomes at scale, and feature industry-leading capabilities including natural language processing, machine learning, analytics, and knowledge graphing.

For more information, **please visit www.attivio.com.**

9 https://www.datanami.com/2016/08/22/search-engines-get-machine-language-boost/