INVESTMENTS & WEALTH MONITOR

A reprinted article from July/August 2021

Machine Learning for Investment Management

By Richard Roche, CAIA®

INVESTMENTS & WEALTH INSTITUTE®

Machine Learning for Investment Management

By Richard Roche, CAIA®

Despite our growing familiarity with robo-advisors, the investment management industry is still in its early days of adopting artificial intelligence (AI) and its subset, machine learning (ML). But adoption of these tools is inevitable because ML bests humans at finding patterns in data, and we can exploit these new-found patterns to improve our investment strategies.

Indeed, very few asset managers have yet to incorporate ML techniques for alpha seeking, position-sizing, risk management, trade execution, or portfolio implementation and management. This means that portfolio managers, analysts, and financial data professionals and programmers have plenty of opportunity to advance their careers and help clients using these techniques.

This article explores some of the applications of machine learning to the world of investment management. It also presents a primer of sorts on the topic of machine learning to familiarize investment managers with how it works.

MACHINE LEARNING AND ASSET MANAGEMENT

Computers are hardly new to investment management. By the 1960s, machine-readable data was widely available. But automation of tasks—including securities trading and portfolio selection—is not machine learning.

A broad framework for ML applications in asset management has been proposed by Derek Snow of FirmAI (Snow 2020). Snow suggests that financial ML applications be divided loosely into four streams. The first concerns asset price prediction, where researchers attempt to predict the future value of securities using ML models. This is how an asset manager might use an ML algorithm—a sequence of well-defined, computer-implementable instructions—to discover unique sources of alpha.

The second stream involves forecasting forward-looking company fundamentals such as earnings, earnings before interest, taxes, and amortization, or free cash-flow prediction (Alberg and Lipton 2018). A 2017 study using deep neural networks (DNNs)—the most sophisticated form of ML—to forecast future fundamentals showed a significant improvement in compounded annual returns (17.1 percent annual with DNNs versus 14.4 percent for a standard factor model). The second stream of ML applications also includes forecasting hard or soft financial events such as analysts' earnings surprises, regime changes, corporate defaults, and mergers and acquisitions.

The third stream entails the prediction or estimation of values that are not related directly to the price of a security, such as idiosyncratic security, sector, or market volatility, credit ratings, and factor quantiles. The fourth and last stream comprises using ML techniques to solve traditional optimization and simulation problems such as position sizing and portfolio optimization.

Marcos López de Prado, a practitioner and frequent contributor to the literature of quantitative machine learning, recommends that asset managers use a separate optimizer or bet-sizing ML model from the alpha capture model (López de Prado 2018, slide 7). An alpha capture model is used to make buy-or-sell decisions—which side of the market—long or short. Snow's fourth stream helps to solve the position-sizing decision or portfolio optimization.

WHAT IS ARTIFICIAL INTELLIGENCE? WHAT IS MACHINE LEARNING?

Artificial intelligence is the technology that computers use to perform tasks that require human intelligence, such as image recognition or spoken language translation. Computers are able to perform these tasks based on explicit instructions—programming provided by humans.

AI is used for fraud detection, spam filtering, speech synthesis, and to generate the recommendations you get on Netflix and Amazon. It gives us ever-more-accurate weather forecasts (Blum 2018), medical diagnoses and treatments (Topol 2019), and facial recognition. Tens of thousands of AI applications are embedded deeply in the infrastructure of every industry (Kurzweil 2006), including investment management.

Machine learning takes AI to the next level. Using machine learning, a computer deploys algorithms in order to detect patterns in data and "learn" to solve complex problems.

Amazon defines ML as "a collection of algorithms that can learn from and make predictions based on recorded data, optimize a given utility function under uncertainty, extract hidden structures from data, and classify data into concise instructions" (Gravell 2019). Stated another way, ML is the ability of computer software to improve its performance by exposure to data without the need to follow explicitly programmed instructions.

Machine learning has redefined computer science by shifting the question from how to program computers to how to get computers to program themselves. That's because with traditional programming, a human tells the computer exactly what to do. But with ML, a human provides an algorithm that the computer uses to figure out how to solve a specific problem by trial and error.

Machine learning is an intersection of computer science, statistics, and probability estimates. The core artifact of ML execution is a mathematical model that describes how an algorithm processes new data after being trained with a subset of historic data. Once an ML algorithm has been trained on historical data, the output of the process is known as a model. The goal of training is to develop models capable of making predictions from observed datasets.

John Wihbey, a Northeastern University professor, says that machine learning is "based on doing ridiculous amounts of multiplication problems very fast" (Callahan 2018).

To be clear, automation of tasks is not artificial intelligence, nor is it machine learning. You can automate a computer to perform certain operations based on a fixed set of rules. For instance, you can manually program an algorithm to sell an asset if the asset price drops by a certain amount (that is, a stop loss). But even giving large numbers of complex rules to a machine does not represent machine learning. It is merely an automation of tasks. The computer will hang the first time it encounters a situation that does not exactly match the program that governs the automation.

In machine learning, the computer is given input (variables and datasets) and output that is a consequence of the inputs. The machine then finds or learns a rule that links the inputs to the output. Because machine learning is heavily oriented toward predictive modeling and inference, ML algorithms often are called prediction machines. This capability makes ML algorithms ideal tools to use in predicting capital markets' direction and economic trends. ML can detect patterns and make predictions that improve over time—which is why it's called learning. An example would be going from 58-percent to 59-percent accuracy in image recognition to the mid to high 70s and so on.

MACHINE LEARNING CATEGORIES AND TECHNIQUES

J.P. Morgan's Quantitative and Derivatives Strategy (QDS) group has produced an "Overview of Machine Learning Methods" that contains useful descriptions of machine learning (Kolanovic and Krishnamachari 2017). There are numerous ways to classify ML and varying degrees of granularity, but QDS classifies ML methods in asset management into three broad categories: supervised learning, unsupervised learning, and deep learning (see table 1).

SUPERVISED LEARNING

Supervised learning is a learning model built to make predictions given unforeseen input. It depends on annotated data: images, audio, or text painstakingly labeled by people. Supervised learning indicates the presence of a supervisor as a teacher. The labeled data is fed into computer algorithms thereby teaching the algorithms what to look for. After ingesting thousands or millions of labeled data, the algorithms become expert at recognizing what they have been taught to see.

After training the model, the machine is provided with a new set of examples (data) so that the supervised learning algorithm analyzes a test or validation dataset. Ideally, the algorithm produces a correct output from the labeled data.

Supervised learning was the original ML method and is still the most widely used. However, its use is constrained to relatively narrow domains defined largely by available labeled training data. This is often called artificial narrow intelligence. This supervised learning algorithm equals or exceeds human capabilities or efficiency—but in only one specific domain, such as handwriting recognition

Table 1

CLASSIFICATION OF MACHINE LEARNING TECHNIQUES

Machine Learning/Artificial Intelligence						
Supervised Learning		Unsupervised Learning		Deep Learning		Other Approaches
Regression	Classification	Clustering	Factor Analysis	Time Series	Unstructured	Reinforcement Learning
Lasso, Ridge, Loess, KNN, Spline, XGBoost	Logistic, SVM, Random Forest, Hidden Markov	K-means, Birch, Ward Spectral Cluster	PCA, ICA, NMF	Multilayer Perceptron (MLP) Convolutional Neural Nets (CNN) Long Short-Term Memory (LSTM) Restricted Boltzman Machine (RBM) Active Learning		Semi-Supervised Active Learning

Source: J.P. Morgan Macro QDS

or as an email spam filter. It uses classification algorithms and regression techniques to develop predictive models.

An example of supervised learning can be found in postal automation. In 1996, the University at Buffalo helped the U.S. Postal Service (USPS) to implement machine-reading of handwritten addresses, which has saved USPS billions of dollars in labor costs over more than 20 years. Today, 99 percent of all mail is sorted without manual intervention, including 98 percent of all handwritten mail (Govindaraju 2016).

UNSUPERVISED LEARNING

Unsupervised learning is the opposite of supervised learning. Unlabeled data is used because training sets do not exist. None of the data is presorted or pre-classified, which means unsupervised learning algorithms are more complex and processing is more time-intensive. Unsupervised learning classifies a dataset by discovering a structure through common elements in the data. This makes unsupervised learning attractive where data is cheap and labels are either expensive or unavailable.

Unsupervised learning algorithms group data based on similar attributes, trends, patterns, or relationships. Unsupervised models include clustering techniques and self-organizing maps. Different unsupervised learners use various strategies to divide data into cohorts or clusters. Some methods are relatively straightforward, rapidly dividing data based on common attributes or similarities. Other unstructured learners are more complex, using two-step clustering where group assignment is made after a second pass through the dataset.

Finding emergent topics in an evolving document collection is a good application for unsupervised learning because you don't know in advance what those topics might be. Detecting faults or anomalous instances in a long time series is another example where you would want to find out if something went wrong, and if it did, when and why (Wittek 2014).

One of the earliest applications of unsupervised learning was PageRank, the original system of measuring the importance of websites and webpages based on links. Rather than determining relevance only from the intrinsic content of a document or from hyperlinks to the document, PageRank used a probability distribution to determine the importance based on the extrinsic relationships between documents. The PageRank algorithm was an application of unsupervised learning that scored the relative importance of webpages without any human supervision.

Supervised and unsupervised learning are considered classical machine learning methods. The distinction between supervised and unsupervised learning typically is well understood. If the data is not labeled with a response—for example, a loan status—then unsupervised learning is the correct choice to discover patterns and lower the dimensionality of the data (Dixon and Halperin 2019, 3).

Conversely, the presence of a label suggests supervised learning is more appropriate. In supervised learning, a "teacher" provides the feedback or supervises the right output for each datapoint in a training set. Although supervised learning is still the dominant force in machine learning in finance, deep learning has been making enormous strides.

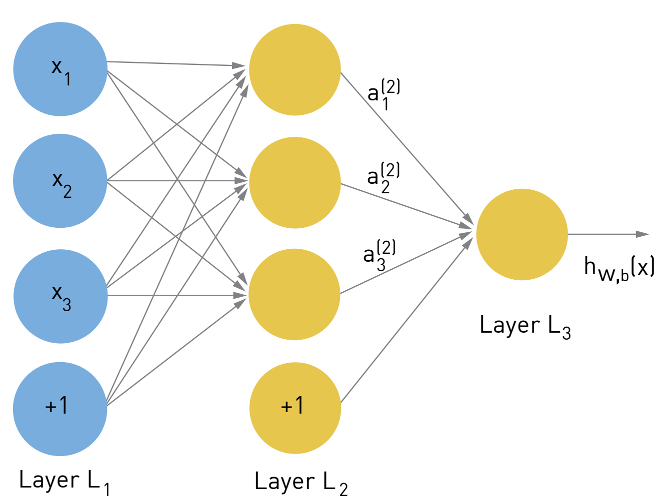
DEEP LEARNING

Deep learning is a branch of machine learning inspired by biology that attempts to replicate the structure and decision-making functions of the brain. An artificial neural network (ANN) analyzes a dataset by processing it through multiple layers (neurons). ANNs extract data features by using multiple filters on overlapping segments of the data.

In its simplest form, an ANN has three layers: an input layer, one or more hidden layers, and an output layer. The interior layers are "hidden" because they are not directly observable from the system's inputs and outputs. The hidden layers are where the "ridiculous amounts of multiplication" occur by applying probability statistics for data discovery.

A neural network with more than one hidden layer is referred to as a deep neural network or DNN; some "very deep neural networks" have 1,200 layers (Huang et al. 2016). When using neural networks or DNNs for stock-index prediction, roughly five or so hidden layers may be preferable

Figure 1 DEEP NEURAL NETWORK



Source: Sangeet Moy Das, Towards Data Science, Deep Learning 101 to avoid model overfitting (Rasekhschaffe and Jones 2019).

In figure 1, Layer L_1 is the input layer. Layer L_2 is the hidden layer where the probability calculations take place. Layer L_3 is the output layer.

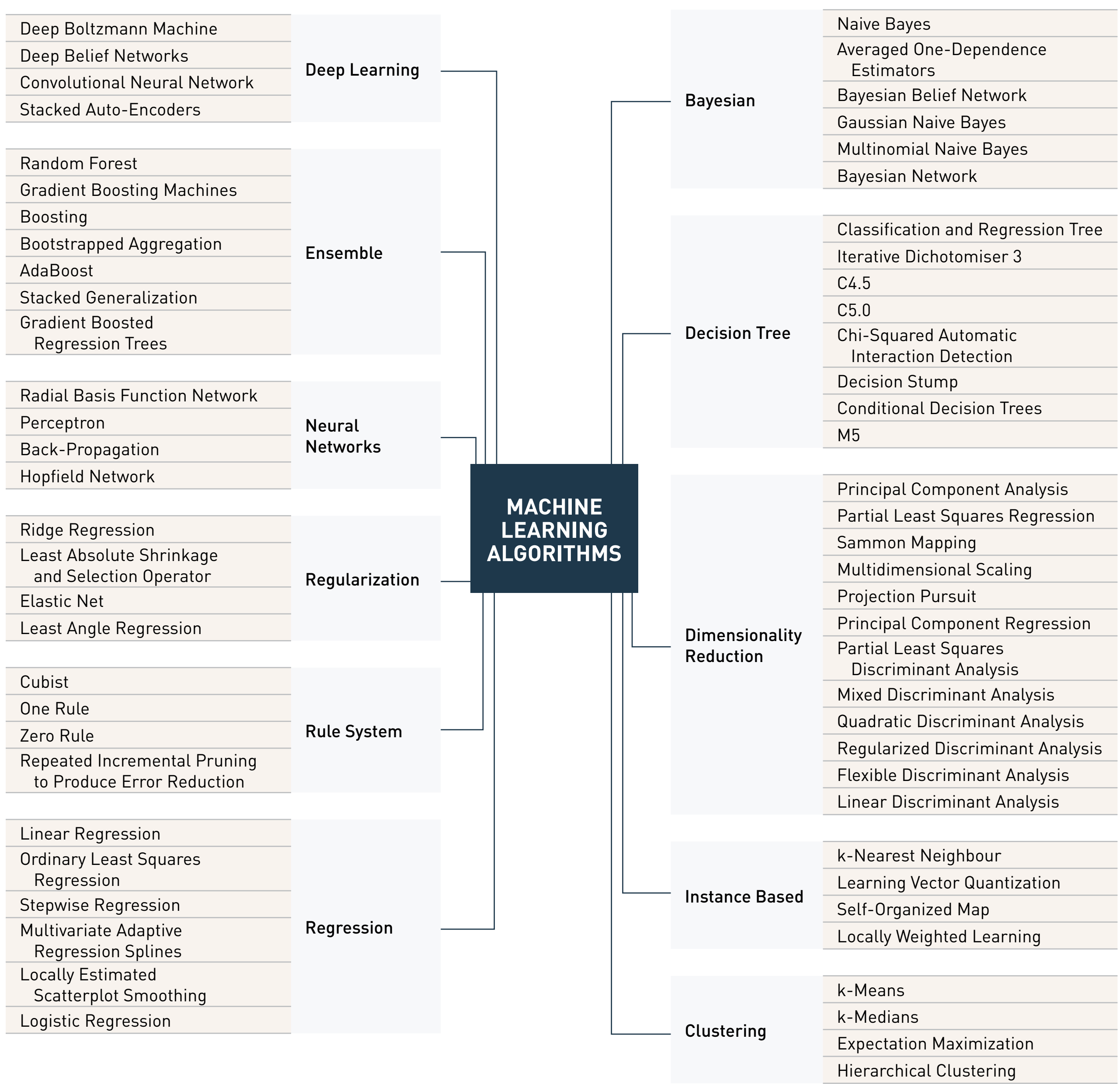
The J.P. Morgan QDS ML taxonomy is also helpful in identifying the tasks ML methods are most accomplished at performing. Note in table 1 that supervised learning is used for regression and classification whereas unsupervised learning excels at clustering and factor analysis.

Within deep learning, representative machine learning algorithms include a multilayer perceptron, such as a convolutional neural net (widely used in image recognition) or a backpropagation neural network (backprop or BP, a classic, extensively used feed-forward mechanism). Operations of BPs can be divided into two steps: feed-forward and backpropagation of data.

Feed-forward. In the feed-forward step, an input pattern is applied and its effects propagate, layer by layer, through the network until an output is produced. The networks' actual output value is compared to the expected

Figure 2

MACHINE LEARNING ALGORITHMS OR MACHINE LEARNERS



Source: Snow (2020)

output and an error signal is computed. The output errors are then transmitted back to the hidden layer that contributed to the erroneous output. This process is repeated (iterated), layer by layer, and the errors are used to improve output accuracy.[1]

Backpropagation. Backpropagation, short for "backward propagation of errors," sends data backward through a network to update an algorithm. For example, in 2012, Geoff Hinton ("The Godfather of Deep Learning"), Alex Krizhevsky, and Ilya Sutskever pioneered a backprop modification called Dropout that randomly drops out neurons in order to prevent any neuron from relying excessively on the output of any other neuron (Hinton et al. 2012). It forces the model to rely on the population behavior of all its inputs. Dropout has yielded remarkable improvements on difficult problems in image recognition and speech using well-known datasets.

Deep learning is behind most prominent accomplishments of ML in the past eight to 10 years. Image recognition, speech recognition, language translation, and autonomous driving rely on deep learning algorithms. Deep learning is used to process datasets of satellite images, sequencing the genome, natural language processing for investor sentiment, and facial recognition. Many asset managers already reap the benefits of deep learning, without actually implementing algorithms used in ANNs.

There are dozens and dozens of ML algorithms or machine learners (see figure 2). No one machine learner is the best choice for all investment research projects. ML practitioners and quantitative analysts should be familiar with a broad range of ML models and their applications. In addition to understanding methods of investment analysis, practitioners need to understand relevant datasets and financial assets used to trade the data signal. But the first step in tackling a dataset is to make an educated guess about which method of analysis is expected to yield the best results.

CONCLUSION

Why has AI and ML finally been adopted in so many everyday applications and commercial enterprises? A confluence of favorable hardware and software developments has combined with changes in human behavior and available data to set the stage for widespread diffusion of ML algorithmic use.

Automation has gotten better, cheaper, and less avoidable. It's still a little mysterious, but it's no longer foreign. Most of us interact with dozens of ML applications on a daily basis. We have casual conversations with Siri and Alexa in our homes, cars, and offices. Waze directs us through traffic and warns us of speed traps.

Now applications for various forms of artificial intelligence and machine learning are presenting themselves universally, and the investment-management industry will not be excluded. Asset managers that learn about these new tools and how they can be used to optimize financial transactions stand to benefit.

Richard Roche, CAIA®, is a managing director at Little Harbor Advisors, LLC., a Boston-area sponsor of quantitative investment strategies. He is the founder of Roche Invest AI, LLC, a consultancy that specializes in research, integration, and promotion of artificial intelligence, machine learning, and utilization of alternative investment data in the investment management industry. Contact him at rroche@littleharboradvisors.com.

ENDNOTE

1. See https://www.cse.unsw.edu.au/~cs9417ml/MLP2/BackPropagation.html.

REFERENCES

Alberg, John, and Zachary C. Lipton. 2018. Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals (April 26). https://arxiv.org/abs/1711.04837v2.

Blum, Andrew. 2018. *The Weather Machine: A Journey Inside the Forecast*. New York: HarperCollins.

Callahan, Molly. 2018. Machines Are Writing News Stories. Does That Mean Robots Will Replace Reporters? *News@Northeastern* (May 11). https://news.northeastern.edu/2018/05/11/machines-are-writing-news-stories-does-that-mean-robots-will-replace-reporters/.

Dixon, Matthew Francis, and Igor Halperin. 2019. The Four Horsemen of Machine Learning in Finance (September 15). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3453564.

Govindaraju, Venu. 2016. Reading the Handwriting on the Letter: UB research that gave new life to the US Postal Service (May). https://www.buffalo.edu/content/dam/www/research/pdf/Postal-Automation-Highlights_20160516.pdf.

Gravell, K. 2019. Increase Operational Efficiency and Improve Client Experience with Machine Learning. SimCorp. https://www.simcorp.com/en/insights/white-papers-and-reports/machine-learning-for-investment-managers.

Hinton, Geoffrey, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors (July 3). https://arxiv.org/pdf/1207.0580.pdf.

Huang, Gao, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. 2016. Deep Networks with Stochastic Depth (July 28). https://arxiv.org/abs/1603.09382v3.

Kolanovic, Marko, and Rajesh T. Krishnamachari. 2017. Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing. J.P. Morgan Quantitative and Derivative Strategies (May). Figure 39, page 52. https://www.cfasociety.org/cleveland/Lists/Events%20Calendar/Attachments/1045/BIG-Data_AI-JPMmay2017.pdf.

Kurzweil, Ray. 2006. *The Singularity is Near: When Humans Transcend Biology*. New York: Penguin Books.

López de Prado, Marcos. 2018. Ten Financial Applications of Machine Learning (Seminar Slides) (June 16): slide #7. https://ssrn.com/abstract=3197726.

Rasekhschaffe, Keywan, and Robert Jones. 2019. Machine Learning for Stock Selection. *Financial Analysts Journal* 75, no. 3 (third quarter): 70–88.

Snow, Derek. 2020. Machine Learning in Asset Management—Part I: Portfolio Construction–Trading Strategies. *Journal of Financial Data Science* 2, no. 1 (winter): 10–23.

Topol, Eric. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Wittek, Peter. 2014. *Quantum Machine Learning: What Quantum Computing Means to Data Mining*. New York: Elsevier.

CONTINUING EDUCATION

To take the CE quiz online, go to www.investmentsandwealth.org/IWMquiz

5619 DTC Parkway, Suite 500
Greenwood Village, CO 80111
Phone: +1 303-770-3377
Fax: +1 303-770-1812
www.investmentsandwealth.org