Eventually Probably Approximately Correct[1]:
An Introduction to Machine Learning for Quants

Walter Alden Tackett, CFA
NxE12, LLC
*[**Presentation:** Monday Oct 16, 2107, 2:30PM @ Q-Group Fall 2017 Conference, Vancouver, BC]*

Abstract

Presently, it appears the management of Financial Investment and Risk is virtually the only area of human endeavor that has not been spectacularly transformed by some variation on Machine Learning, Big Data, Data Science, or Artificial Intelligence. These terms don't seem entirely synonymous and the distinction is not clearly stated by those reporting upon or promoting them, even as they declare that all firms must have an *Artificial Intelligence Strategy*. Yet it appears that *somebody* must be doing *something* right: computer algorithms have demonstrated the cunning intelligence of bluffing and deceit in Poker[2], while the Social Sciences cannot even define or agree upon what intelligence *is*.

This introductory presentation will not teach Python programming, nor will it issue a *certificate of completion:* rather, it will provide a clear definition of Machine Learning as a field governed jointly by the studies of Statistical Data Analysis and Computer Science. The methodological lineage traces directly to both Operations Research and the Systems-Theoretic Disciplines of Electrical Engineering, which over time have reciprocally adopted elements of Machine Learning. Quant Finance has similar ancestry, and many key elements of Machine Learning will be relatively familiar to the audience: examples include Optimization, Regression, Dynamic Programming, Monte Carlo Methods, Clustering, and Principal Components Analysis.

Having established common ground with quant finance, the discussion of Machine Learning's less familiar elements begins with the central tenet of *generalization*[3], which is in fact the opposite of "data mining bias," a pejorative term unique to financial economics. The approach that separates this *"Statistical Computing"* paradigm from Mathematical Statistics provides a concrete justification for its use in Quant Finance, motivating the explanation of what Computer Science is and why it is essential to competitive advantage (for those who didn't get the memo about HFT). The recursive use of algorithms that set the parameters of other parameter-estimation algorithms – collectively called meta-algorithms and hyperparameters – is standard in Machine Learning and will be illustrated using a Linear Regression example. The overview will conclude with a discussion of taxonomies, including the more general study of adaptive computing, different types of learning, and how quants may tilt their focus to algorithms, data, or both.

Breaking the *jargon barrier*: Machine Learning includes deep connections to biology and psychology, whose vocabulary has combined over many years with words normally used to describe radar signals or various computer and rocket parts, together forming an intimidating pidgin dialect. The talk assumes no prior knowledge of such arcana. Recommended materials for further study do the same (See: *Learning Machine Learning*, below). Those who follow the prescribed path will not be taken by surprise when encountering a RELU in the Receptive Field.

Endnotes

[1] Not to be confused with *Probably Approximately Correct*, the recent book by Leslie Valiant, which greatly expands upon *PAC Learning*, a foundational theory of machine learning that first appeared in *Valiant, Leslie G. 1984. "A Theory of the Learnable." Communications of the ACM 27 (11): 1134–1142*.

[2] The https://deepstack.ai web site is accessible to both specialist and non-specialist viewers. It describes the DeepStack project headed by Michael Bowling of U. Alberta, including gameplay footage, video lecture, and pre-print of the cover article from the May 5, 2017 issue of *Science*. DeepStack is a program (one of two, the other being CMU's Libratus) that has significantly outperformed human champions at Heads Up No-Limit Poker, a game of *incomplete information*, a property which places it in a category of difficulty far beyond Chess or Go. See also: *Moravčík, et al, and Michael Bowling. 2017. "DeepStack: Expert-Level Artificial Intelligence in Heads-up No-Limit Poker." Science 356 (6337): 508. May 4, 2017.*

[3] Defined as the objective of *achieving optimal performance against out-of-sample data or states*, generalization is in fact the subject of the full-page diagram preceding the first page of the first research journal article describing the first working Machine Learning program. See: *Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal of Research and Development. 3 (3): 210–229.*

Learning Machine Learning

First Step: Computer Science Overview

Christian and Griffiths, *Algorithms to Live by*.

This is a popular science and general readership book from which even professionals in the field can gain new insights. It draws from life's problems large and small: Where should I park? How long should I date before I marry? What's the best way to sort my laundry? From these life examples Computer Science is revealed as the study of *problems*, their characteristics, and the *algorithms* that address them. The book illustrates hallmarks of what makes problems easy or hard, how solutions and algorithms can employ tradeoffs to gain different types of efficiency, and how these ideas inform Economics, Political Decisions, and Behavioral Psychology. Free of equations, suitable for audiobook, and recommended for all attendees.

Machine Learning for Quants, Hands On: Starting Point

James et al., *An Introduction to Statistical Learning with Applications in R*.

For quants intending to investigate Machine Learning, this is the preferred starting point because it approaches the field from a statistical perspective that highlights similarities and differences with the foundations of Quant Finance practice. It is highly recommended to learn the material via the online Stanford class based on this book (it is a live course, but recent semesters are archived for on-demand use). Course and book are free, as are all software and data. NOTE: this class is taught by Tibshirani & Hastie through Stanford's Lagunitas program, and should not be confused with Andrew Ng's Coursera Machine Learning class. Taking both isn't a waste, but if forced to choose, *Statistical Learning* is more compatible with Quant Finance.

The Full Reference: Statistical Learning

Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning, 2ed*.

The first edition of this book was the culmination of a ten-year transformation beginning in the early 1990's that resulted in both a rigorous grounding of ML in statistical method as well as firm establishment of Statistical Data Analysis first proposed by John Tukey, where proofs about algorithms replace proofs about equations, reducing reliance on closed-form formulas and assumptions regarding data distributions. While the authors were key figures in that movement, they carefully cover the breadth of the statistical learning field and its contributors. Note: *Intro to Statistical Learning* is culled from this book.

Supplement: Hastie, Tibshirani, & Wainwright, *Statistical Learning with Sparsity*.

Addresses the special case (or, perhaps, the *common* case) of too many candidate explanatory factors and not enough data.

Machine Learning for Quants, Hands On: 800 pages of Next Steps

Goodfellow, Bengio, and Courville, *Deep Learning*.

This book assumes minimum knowledge equivalent to understanding all concepts in Christian & Griffiths, Intro to Statistical Learning, and undergrad-level calculus. Requires Python (learnable on-the-fly). Includes 200-page intro to linear algebra and differential tensor calculus. Teaches math, statistical, and learning concepts, and requires large projects using TensorFlow.