The Formation of the Statistical Learning Paradigm and the Field of Machine Learning, *circa* 1985–2000

Cosma Rohilla Shalizi*

Begun 8 November 2020, last LaTeX'd April 29, 2023

NOTES TOWARDS A DRAFT: Comments welcome, but please do not quote

Abstract

Machine learning (ML) is, today, the name of an important field within computer science. It's widely applied in industry, and has made important in-roads in nearly every quantitative discipline; much of what non-computer-scientists call "artificial intelligence" or "algorithms" is, in fact, machine learning. In one sense ML is the heir to wide-ranging efforts, dating back to the 1950s, to build machines, especially digital computer programs, which could be said to "learn" (in various senses of the word). Many of those early efforts would however no longer be intelligible to current researchers as part of the field. Instead, a specific paradigm crystallized between about 1985 and 2000, arguably even between 1990 and 1995. Crucially, this paradigm involved the selective adoption by computer scientists of ideas from statistics, relating to prediction, decision-making, and inductive inference. This article traces the rise and content the statistical learning paradigm, which has continued to give ML its core concepts and disciplinary identity, even as the field has churned through many specific techniques and technologies.

During the late 1980s and 1990s, the field of machine learning underwent a scientific revolution, very much in the mold of Kuhn (1970)[1]. Before, it was a pre-paradigmatic endeavor; by the end, there was a well-defined, widely-shared paradigm, one which still pretty much defines machine learning as an academic field. Prior to this period, efforts to construct and theorize machines that learned showed no consensus on what constituted a good contribution, or on how to evaluate contributions. Starting in this period, the typical, readily-legible paper in machine learning follows the **statistical learning** paradigm. The ideal type of such a paper

*Departments of Statistics and of Machine Learning, Carnegie Mellon University; External Faculty, Santa Fe Institute

[1] I am aware that Kuhn's scheme has been challenged in many ways, from its conceptual foundations to its historical veracity, both fairly shortly after publication (e.g., Toulmin 1972) and later (Donovan *et al.*, 1988). I neither want to defend all of the book, nor enter into this controversy. But my reasons for thinking that at least this episode fits his scheme will, I hope, become clear.

1. introduces a new class of "machines", defined as classes of mathematical functions mapping inputs[2] to outputs[3];

2. selects a particular function from this class of functions by solving a mathematical optimization problem involving the inputs and outputs of the function, averaged over a (hopefully) representative "training" data set; and

3. demonstrates superiority over predecessors and alternatives by evaluating the learned "machine", in terms of that optimization problem, on new, "testing" data, or used the statistical technique of cross-validation in the absence of separate testing data.

Ancillary contributions under this paradigm could be mathematical theory saying something about how well such a pipeline could be expected to work under various assumptions (largely relying on the statistical / probabilistic field known as "empirical process theory"), algorithms for actually doing the optimization in step 2 (or theory about such optimization procedures), applications to particular concrete cases, etc. This is a paradigm because it provides a repeatable template for what a good research contribution should look like, along with standards for evaluating and, especially, comparing contributions.

The class of mathematical functions called "neural networks" or "multi-layer perceptrons" were the first instances of this paradigm, swiftly followed by "kernel machines" and "support vector machines", which severed any pretense of biological or psychological inspiration, and were justified purely in terms of the paradigm, i.e., their statistical performance. A large number of different types of machines have risen and fallen in prominence within the field (neural networks have come back in to vogue in the last decade, as "deep learning"), but the paradigm has remained in place.

The formation of this paradigm involved computer scientists taking on conceptual tools, most notably decision theory, cross-validation and empirical process theory, which had been previously developed for other purposes in statistics. Indeed, the whole formation of machine learning as a separate discipline, with its own publications, career tracks, internal traditions, etc., can be seen as a process of computer scientists *selectively* adopting statistical ideas, and retaining the sub-set of "learning" problems which fit best with the paradigm and excluding the others (some of which have however continued under labels like "knowledge discovery in data bases").

1 Pre-Paradigmatic Machine Learning

- Samuels' work on checkers in the 1950s (brought to wide attention by, e.g., Wiener (1961, 1964))

- The first edition of Holland (1992), from 1975, was certainly seen *at the time* as a book about learning by machines, but would now be classified as being

[2] Also called "features", "covariates", "attributes", "regressors", etc.

[3] Variously "responses", "labels", "predictions", "actions", "decisions", etc.

about optimization or *maybe* artificial intelligence, but not learning in the ML sense; similarly Holland *et al.* (1986) is clearly about *learning* (the subtitle is "Processes of inference, learning, and discovery"), but it *never* evaluates any algorithm or method by out-of-sample prediction performance (as opposed to, say, being able to re-discover Vitruvius's argument in favor of a wave theory of sound, or matching experiments in animals on operant conditioning).

- Or, again, look at Mitchell (1993) developing a problem to solve letter-sequence analogies ($abc : pqr :: aabbcc$:?) based on examples. Clearly, this was a study of *learning*, but also clearly *not* something that fits with the new paradigm. (What was the loss function? the distribution over examples? the evaluation in terms of risk or cross-validation?)

- So there were other imaginable directions for learning. There were also imaginable directions for AI without learning: the "expert systems" approach, presuming extensive explicit human input of formalized knowledge (Feigenbaum *et al.*, 1988; McCorduck, 2004) [[MORE FORMAL CITATIONS]] — forms a contrast case, very prominent in the 1980s

- Pre-paradigm work on neural networks: dates back to McCulloch and Pitts (McCulloch, 1965), plus later developments up to Minsky and Papert's negative results about single-layer networks; then the revival by Rumelhart, McClelland, etc., journalistically surveyed at the time by (e.g.) Caudill and Butler (1990)

- Pre-paradigmatic textbooks: Tou and Gonzalez (1974) or Hutchinson (1994) or Mitchell (1997) (the last is especially interesting because it sees that the paradigm exists but doesn't *confine* itself to them)

2 Antecedents of the paradigm

- Statistical decision theory (1920s–1950s): Neyman and Pearson (1933) distinguished between different *kinds* of error and their costs, and derived *optimal* statistical procedures; Neyman on "inductive behavior"; Wald's statistical decision theory, and the notion of "statistical decision functions"; Giocoli (2011) reviewing how "*homo economicus* became a Bayesian statistician"; the notions of "loss function" and "risk"

- Glivenko and Cantelli (1930s) extending the law of large numbers from a *single* function of the data, to a *uniform* result over *infinitely many* functions of the data [[citations]]; work by Kolmogorov, Donsker, etc. on convergence of distribution functions

- Vapnik and Chervonenkis (Vapnik and Chervonenkis, 1971; Vapnik, 1979/1982) on Glivenko-Cantelli type results for other classes of functions, and the link to loss functions and "risks" (what was the context of this, within Soviet probability and statistics?)

- Introduction of cross-validation in to the formal statistical literature by Stone (1974); Geisser (1975), acknowledging earlier informal uses; recognition of the generality of the approach and its use for selecting *one* model among others by Geisser and Eddy (1979)
- Empirical process theory (codified early by Pollard (1984), as well as Pollard (1989, 1990), and Wellner [CITES]; but drawing on earlier work by Donsker et al.), providing tools for Glivenko-Cantelli or Vapnik-Chervonenkis style results, rapidly applied to neural networks (Anthony and Bartlett, 1999) [[dig up citations to original papers from there]]
- Contributions from physicists or ex-physicists: Watkin *et al.* (1993) emphasizing average-case results rather than bounds; Michael Jordan as an exemplary figure moving from physics to learning theory and applications

3 Rise of the Paradigm

[[TODO: Content analysis of the NIPS/NeurIPS proceedings between the beginning and c. 2000, looking at uses of selected originally-statistical terms and concepts, especially risk as expected loss, cross-validation, bootstrap, empirical process theory. When did people *stop* having to explain, e.g., cross-validation?]]

- Valiant (1984) introduces the idea of "probably approximately correct" learning, which is the same as the statisticians' idea of "consistency" but as applied to the *error rate* rather than the *parameters* ("risk consistency")
- Appearance of paradigmatic textbooks: Kearns and Vazirani (1994), Hastie *et al.* (2001) [[others?]]
  - Special attention to the works of Vapnik (Vapnik, 1998, 1995), which were widely acknowledged as important if somehow a bit odd
- Items from the 1990s showing that cross-validation was still a new thing that needed to be explained, without consensus on how to do it (e.g., Shardanand and Maes (1995), the first paper on recommendation systems)
- The appearance and blooming of "support vector machines" in the hands Vapnik, Cortina, etc., with explicit definition of the function class / machine, the optimization problem (and algorithms for solving the problem), and evaluation by cross-validation, complemented by theory about how well we can expect the models to perform [[CITATIONS]]

4 Points to make somewhere

4.1 Shared data and leader-boards

Role of the UCI repository of standardized data sets. NIST (and DARPA?) pushing benchmarks and performance on shared data sets to assess progress.

4.2 Statistical learning vs. optimization

On the one hand a tendency to regard the optimization as, so to speak, somebody else's problem, i.e., exploring specific optimization methods came to be seen as not really part of machine learning. (Unlike, say, earlier and parallel work on genetic algorithms and evolutionary optimization, where the process of improving the score on some objective function was thought of as a process of learning.) Against that, it was important that the optimization problems posed specific combinations of architecture and loss function were tractable ones, for which there were efficient algorithms; or that one could show how to transform problems in to forms where there were efficient algorithms. Outstanding case: the ease of optimizing support vector machines, via tractable convex programming problems, as opposed to optimizing neural network weights by back-propagation (= gradient descent).

4.3 Statistical learning vs. statistics

ML has not just been a re-labeling of statistics[4], because the up-take of statistical ideas was highly selective. Statisticians' traditional concerns with the interpretation of parameters in the models, and quantify uncertainty about those parameters, fell pretty completely by the wayside; even quantified uncertainty about predictions was at best a minor consideration. Model mis-specification was never an issue. (There was no pretense that the models could ever be correctly specified.) Even the applied statistician's worries about measurement quality, representative samples, biases in the data-collecting process, etc., were largely set to one side.

4.4 Assorted quotation materials

> To begin with learning machines: an organized system may be said to be one which transforms a certain incoming message into an outgoing message, according to some principle of transformation. If this principle of transformation is subject to a certain criterion of merit of performance, and if the method of transformation is adjusted so as to tend to improve the performance of the system according to this criterion, the system is said to *learn*. (Wiener, 1964, p. 14, emphasis in the original)

This would still be recognizable to a contemporary researcher, though they'd find the emphasis on "messages" odd (unless perhaps they specialized in information-theoretic bounds). But then the examples which follow!

> It is not difficult to design a machine which exhibits the following type of learning. The machine is provided with input and output channels and an internal means of providing varied output responses to inputs in such a way that the machine may be "trained" by "trial and error" process to acquire one of a range of input-output functions. Such a machine,

[4]By contrast, I'd argue that the more recent phenomenon of "data science" *is* just such a re-labeling exercise — at its best.

> when placed in an appropriate environment and given a criterion of "success" or "failure" can be trained to exhibit "goal-seeking" behavior.

(Minsky, Dartmouth AI proposal, p. 8), http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf

References

Anthony, Martin and Peter L. Bartlett (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge, England: Cambridge University Press.

Caudill, Maureen and Charles Butler (1990). *Naturally Intelligent Systems*. Cambridge, Massachusetts: MIT Press.

Donovan, Arthur, Larry Laudan and Rachel Laudan (eds.) (1988). *Scrutinizing Science: Empirical Studies of Scientific Change*. Dordrecht: Kluwer Academic. Reprinted 1992 (Baltimore: Johns Hopkins University Press) with a new introduction.

Feigenbaum, Edward, Pamela McCorduck and H. Penny Nii (1988). *The Rise of the Expert Company: How Visionary Companies Are Using Artificial Intelligence to Achieve Higher Productivity and Profits*. New York: Times Books. Foreword by Tom Peters, expert systems catalog by Paul Harmon.

Geisser, Seymour (1975). "The Predictive Sample Reuse Method with Applications." *Journal of the American Statistical Association*, **70**: 320–328. doi:10.1080/01621459.1975.10479865.

Geisser, Seymour and William F. Eddy (1979). "A Predictive Approach to Model Selection." *Journal of the American Statistical Association*, **74**: 153–160. doi:10.1080/01621459.1979.10481632.

Giocoli, Nicola (2011). "From Wald to Savage: *homo economicus* becomes a Bayesian statistician." E-print, Munich Personal RePEc Archive, MPRA/34117. URL https://mpra.ub.uni-muenchen.de/id/eprint/34117.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Holland, John H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition, Ann Arbor, Michigan: University of Michigan Press, 1975.

Holland, John H., Keith J. Holyoak, Richard E. Nisbett and Paul R. Thagard (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, Massachusetts: MIT Press.

Hutchinson, Alan (1994). *Algorithmic Learning*. Oxford: Clarendon Press.

Kearns, Michael J. and Umesh V. Vazirani (1994). *An Introduction to Computational Learning Theory*. Cambridge, Massachusetts: MIT Press.

Kuhn, Thomas S. (1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 2nd edn.

McCorduck, Pamela (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Boca Raton, Florida: CRC Press, 2nd edn. First edition, San Francisco: W. H. Freeman, 1979.

McCulloch, Warren S. (1965). *Embodiments of Mind*. Cambridge, Massachusetts: MIT Press.

Mitchell, Melanie (1993). *Analogy-Making as Perception: A Computer Model*. Neural Network Modeling and Connectionism. Cambridge, Massachusetts: MIT Press.

Mitchell, Tom M. (1997). *Machine Learning*. New York: McGraw-Hill.

Neyman, Jerzy and Egon S. Pearson (1933). "On the Problem of the most Efficient Test of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London A*, **231**: 289–337. doi:10.1098/rsta.1933.0009.

Pollard, David (1984). *Convergence of Stochastic Processes*. Berlin: Springer-Verlag. URL `http://www.stat.yale.edu/~pollard/Books/1984book/`.

— (1989). "Asymptotics via Empirical Processes." *Statistical Science*, **4**: 341–354. URL `http://projecteuclid.org/euclid.ss/1177012394`. doi:10.1214/ss/1177012394.

— (1990). *Empirical Processes: Theory and Applications*, vol. 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, California: Institute of Mathematical Statistics. URL `http://www.stat.yale.edu/~pollard/`.

Shardanand, Upendra and Pattie Maes (1995). "Social Information Filtering: Algorithms for Automating "Word of Mouth"." In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, vol. 1, pp. 210–217. New York: ACM Press. doi:10.1145/223904.223931.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions." *Journal of the Royal Statistical Society B*, **36**: 111–147. URL `http://www.jstor.org/stable/2984809`.

Tou, Julius T. and Rafael C. Gonzalez (1974). *Pattern Recognition Principles*. Reading, Massachusetts: Addison-Wesley.

Toulmin, Stephen (1972). *Human Understanding: The Collective Use and Evolution of Concepts*. Princeton, New Jersey: Princeton University Press.

Valiant, Leslie G. (1984). "A Theory of the Learnable." *Communications of the Association for Computing Machinery*, **27**: 1134–1142.

Vapnik, Vladimir N. (1979/1982). *Estimation of Dependencies Based on Empirical Data*. Berlin: Springer-Verlag. Translated by Samuel Kotz from *Vosstanovlyenie Zavicimosteĭ po Émpiricheckim Dannim* (Moscow: Nauka).

— (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1st edn. doi:10.1007/978-1-4757-2440-0.

— (1998). *Statistical Learning Theory*. New York: Wiley.

Vapnik, Vladimir N. and Alexey Y. Chervonenkis (1971). "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability and its Applications*, **16**: 264–280. doi:10.1137/1116025. Translated by B. Seckler.

Watkin, Timothy L. H., Albrecht Rau and Michael Biehl (1993). "The Statistical Mechanics of Learning a Rule." *Reviews of Modern Physics*, **65**: 499–556. doi:10.1103/RevModPhys.65.499.

Wiener, Norbert (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.

— (1964). *God and Golem, Inc.: a Commentary on Certain Points where Cybernetics Impinges upon Religion*. Cambridge, Massachusetts: MIT Press.