Machine Learning as Philosophy of Science

Kevin B. Korb
School of Computer Science
and Software Engineering
Monash University
Clayton, Vic 3168 Australia
korb@csse.monash.edu.au

Abstract

In this position paper I consider three aspects in which machine learning and philosophy of science can illuminate each other: methodology, inductive simplicity and theoretical terms. I examine the relations between the two subjects and conclude by stressing the fundamental identity between the two disciplines.

1 Introduction

Machine learning studies inductive strategies as they might be carried out by algorithms. The philosophy of science studies inductive strategies as they appear in scientific practice. Although they have developed to a great extent independently, the two disciplines have much in common. This is slowly coming to be recognized in a number of ways, at least by some AI researchers. In particular, there has been substantial new interest in the relation between AI and statistics, as evidenced by new conferences devoted to the subject as well as the profusion of statistical subjects appearing in the major annual AI conferences *Uncertainty in AI* and the *International Conference on Machine Learning*; and the relation between statistics and the philosophy of science is of long standing (e.g., Reichenbach, 1949). What I shall argue here is that the two disciplines are, in large measure, one, at least in principle. They are distinct in their histories, research traditions, investigative methodologies; however, the knowledge which they both ultimately aim at is in large part indistinguishable. Furthermore, it is appropriate for them to begin to share research techniques as well.

In *Computational Philosophy of Science* (1988), Paul Thagard presented some similar ideas. In particular, he emphasized that philosophies of scientific method, if they have any merit, ought to be realizeable in a computer program. As a minimal test of methodological cogency, either normative or descriptive methodological proposals ought to be clear enough and precise enough to be implementable in a universal computational device. It can be, and has been (e.g., Dreyfus, 1992, Humphreys and Freedman, 1996), argued that many of the skills which humans have, including the methodological skills employed by scientists, are non-algorithmic. It would follow that Thagard's minimal test of methodological cogency is strictly incorrect. I do not believe that these anti-computational arguments succeed (see Korb, 1996, and Korb and Wallace, 1997, respectively). But even supposing that they do, meeting Thagard's test would remain a valuable desideratum: implementation of a methodology is as good a demonstration of cogency as one can hope for, and provides a ready test environment of the merits or demerits of the proposed method through computer simulation. Such implementation therefore provides for an experimental philosophy of science that goes beyond the empirical methods applied thus far within the philosophy of science, namely the observation of laboratory life in the "new experimentalism" of Franklin (1990) and others and the historical investigations of scientific practice, as in the work of Hacking (1983). There has been important recent work putting methodoligical concepts into the practice of computer programs, especially in the area of learning causal models (e.g., Glymour and Cooper, 1999); thus, there is an existence proof of the feasibility of implementing proposed methods.[1]

Since there have been advanced no successful arguments that anything crucial in scientific practice is non-algorithmic (even though there are non-defeated arguments that this is conceivable), and since there already are examples of the implementation of important and useful inferential methods, I shall assume in the sequel that all of scientific method may be implemented algorithmically. If this is wrong, then any considerations below will turn out to apply to a (demonstrably non-empty) subset of human scientific practices.

2 Meta-learning and the search for scientific method

There are infinitely many possible inductive strategies. In his *normative* problem of induction, David Hume (1739/1888) argued over 200 years ago that no single inductive strategy can be universally better than all others; machine learning researchers have only recently discovered the relevance of this result (cf. Wolpert and Macready, 1995, and Schaffer, 1994), which implies that no one machine learning algorithm (or, at any rate, no one algorithm with a fixed "bias" — i.e., fixed learning parameters) can perform optimally with respect to every learning problem (or, in every possible world). Here I shall assume that Hume's problem of induction is not solvable; although I do not know that that is so, there is apparently substantial evidence that it is so. The meta-learning problem is, having given up on finding a universal learning algorithm, to attempt to find heuristics or algorithms for selecting an optimal inductive algorithm (or bias) given a particular learning problem. Of course, the meta-learning problem cannot itself be fully soluble on our assumption, for the meta-learning algorithm, conjoined with the individual algorithms amongst which it is selecting, would solve Hume's problem. Nevertheless, assuming that there is some pattern to the relation between problem types and the inductive algorithms that are usefully employed on them, we can hope to learn that pattern just as well as we can hope to learn first-order patterns in the first place. This meta-learning work in machine learning can be applied to the *descriptive* problem of induction: identifying how, as a matter of fact, human scientists go about their inductions.

What distinguishes epistemologists from philosophers of science is their view of just what it is to attempt to understand human knowledge of the world. Philosophers calling themselves epistemologists tend to view philosophy as an analytic, a priori enterprise, believing that we can answer the important philosophical problems (or "dissolve" the more perplexing philosophical "puzzles") by better understanding the interrelations between relevant linguistically accessible concepts — hence, Strawson's (1999) unsatisfactory "dissolution" of the problem of induction. Philosophers of science tend to see philosophy of science as a meta-science,

[1]Note that I would prefer to distance myself from Slezak's slightly premature announcement of the same (Slezak, 1989), since that was in reference to Langley et al.'s (1987) work on the program BACON, which for many reasons does not embody a serious methodology for real science — for example, it copes with noisy measurements only in an egregiously ad hoc fashion.

as the (perhaps supertheoretical) scientific study of science. Many of them practice within departments of the history and philosophy of science (HPS) and actively seek to use historical case studies to inform their philosophical theories. And yet it is natural to have reservations. Giere (1973) raised the question of whether or not the HPS approach is just committing the naturalistic fallacy, by attempting to found the ought of science on its is — or its was. To the extent that philosophy of science itself aims at descriptive knowledge of human science, the objection does not apply. And even where normative knowledge is aimed at, knowledge which will allow us to judge some scientific proposals as preferable to others, it can be argued that the history of science provides relevant inductive evidence in showing us many clear examples of failures and some clear examples of successful inductions. Laudan (1987), in particular, has suggested that our knowledge concerning the most appropriate scientific method be relativized to particular goals, and that it then can be obtained by induction on the history of science. One could take this project further by considering experiments with different methods instead of using only available historical observations. Whereas conducting grand experiments in the philosophy of science by employing competing teams of scientists is hardly a realistic option, the application of Thagard's criterion by implementing competing inductive algorithms computationally enables for the first time in history the experimental philosophy of science via computer simulation (as it does also, perhaps surprisingly, an experimental ethics; see Mascaro, Korb and Nicholson, 2000).

Thus, we see that the meta-learning project in machine learning and the methodology project in the philosophy of science are one and the same, at least insofar as they are (or might be) directed at the problems raised by human science. Furthermore, computer simulation of problem environments and algorithms engaged in learning about them, which is a common technique in machine learning, opens up the possibility for philosophy of employing experimental methods, in addition to the observational and historical methods employed in the past.

3 Inductive Simplicity

"Everything should be made as simple as possible, but no simpler." (Albert Einstein)

Ockham's principle asserts that we should not multiply entities beyond necessity. Applying this to induction, we should aim for the simplest theory possible, that is, the simplest theory which can reasonably be held to account for the evidence. This contrasts markedly with the idea that optimal inductions best "save the phenomena", or "inference to the best explanation", or in statistical theory the idea of adopting that model which maximizes the likelihood on the data. With such approaches we ignore the complexity of our theories and suppose that the only epistemological criterion of value is "explanatory power" — how closely the data are represented by the theory. It has been well established in both statistics and in machine learning that the result of such over-attention to the data in hand is *overfitting:* the additional complexity accepted into our hypotheses in order to precisely fit the data in fact is fitting measurement noise, with the result that the complex model "accounts" for the existing data (with noise), but by incorporating spurious theory it fails to be optimal in accounting for new data, i.e., predictive power. In consequence, even if we are using a consistent, (statistically) unbiased estimation technique, so that it is guaranteed to converge on the truth in the (infinite) limit, in sample sizes short of that limit, the technique can be well and truly beaten by an alternative respecting Ockham (see Dai, Korb, Wallace and Wu, 1997, for an example of this). In Bayesian terms, such approaches to inference maximize likelihood to the neglect of the prior probability of theories; hence, they are insufficiently sensitive to the posterior probability of theories, since that depends on both prior and likelihood. Bayesian inference can be put into 1-1 correspondence with minimum encoding approaches to computational induction, where simplicity is equated with shorter message lengths and higher probabilities (see, e.g., Solomonoff, 1964; Georgeff and Wallace, 1984). Such inference puts Einstein's version of Ockham's Razor into application.

Ockham's Razor in this form receives whatever support accrues to Bayesian principles for scientific induction, which is not negligible (e.g., Howson and Urbach, 1993; Korb, 1992). Other considerations in its favor may also be raised. In a machine learning context, where induction implies some search of the space of hypotheses, there is no serious alternative to employing *some* variety of simplicity metric to govern the search: typically the space is infinite, even if enumerable, so hypotheses to be examined are constructed by the algorithm during the search, necessitating the examination of simpler hypotheses first since the alternative program would not halt.[2] There is also a more direct case to be made for simplicity. If complexities in our hypotheses are introduced prior to any pressure on their behalf in the evidence, then they have been introduced without any empirical guidance. So there will be no more reason to believe that they correspond to reality than that any fantasy does. On the other hand, if we are conservative a la Ockham in introducing complexities, then each additional complexity will have been introduced only when the evidence actually justifies it (or, at least, when it appears to justify it). Assuming that we are sufficiently conservative and that our inductive technique does converge on the truth in the limit (and that the evidence does not happen to mislead us in some particular case), then by applying Ockham's Razor, we may err by having too simple a theory to correspond to reality — which would imply having too small an evidential base, and so be a rectifiable error — but we will never err by having too complex a theory.

4 Theoretical terms

Scientific theories commonly employ theoretical terms, like 'electron', 'gene' and 'super-ego'.[3] These terms have an inductive importance. Hempel's "theoretician's dilemma" (Hempel, 1958) suggested that theoretical terms are necessarily dispensible: if they are useful, then they must lead to true observational consequences. But, if they do that, then by Craig's lemma (Craig, 1956), or alternatively by Ramsification (Ramsey, 1931), they can be eliminated in favor of laws expressed in the observation language (or, per Ramsey, by theories in which theoretical terms are replaced by closed variables). As Hempel noted, however, this possibility is not a real one: in the case of Ramsification, there is no ground to claim that the new theory is any more scrutible (or any less ontologically committed to theoretical entities, following Quine on ontological commitment) than the old; whereas Craig's procedure for eliminating theoretical terms typically replaces a finite theory with a theory requiring infinitely many axioms. So, in any case, the theoretical terms are at least heuristically necessary. Nor is there any theoretical ground for claiming that they engage us in no ontological commitment.

[2] More complex hypotheses can be examined before simpler ones on any metric, but no more than a finite number of times in any case.

[3] Theoretical terms are often introduced by contrast with observational terms. But that distinction is problematic in many ways. I believe a more promising account of theoretical terms is one that simply considers them as those terms introduced by a theory, as in the notion of T-theoreticity in Sneed (1979).

Machine learning has dealt with theoretical terms from its beginning. BACON (Langley, 1987), limited as it was, nevertheless could generate new variables from old (e.g., observational) variables in order to discover more perspicuous natural laws. In automated causal discovery, researchers are working on methods to learn models with latent variables. In classification research, the generation of theoretical terms is called *constructive induction*, which is an active area of research.

The machine learning of theoretical terms raises some new issues for philosophy of science. In addition to the question of just how to generate theretical terms, it poses the question of how to justify their introduction — not in the sweeping way of Hempel's dilemma, but the concrete way of coming up with criteria for preferring one theory over another in the inductive search through the hypothesis space. Machine learning thus makes specific the otherwise general concerns of philosophy of science in introducing terms which refer to non-observables.

5 Meta-evaluation

I have suggested a number of ways in which machine learning can aid the philosophy of science. One area in which machine learning has clearly lagged behind both statistics and philosophy of science is in coping with the meta-evaluation problem. That is the question of how to judge the relative merits of one machine learning algorithm over another, which is the key issue in meta-learning described above. Although there has been research in this area (e.g., Bensusan, 1999), the standard of practice in the field as a whole is abysmal (cf. Hoffman (ed.) *Empirical Methods in AI, IJCAI 1999 Workshop*). Progress in Bayesian philosophy of science and Bayesian computational statistics is very likely to find fruitful application in this problem for machine learning.

In addition to the institutional and historical differences in tradition that separate philosophy of science and machine learning, the subject matter of philosophy of science just appears radically different from that of machine learning. Philosophers of science are attempting to account for messy, obscure practices of scientists that are embedded in an extraordinarily rich cultural context, while AI researchers are engaged in writing clean, crisp, clear-cut programs implementing straightforward algorithms, yes? No.

It is true that AI has long supported a tradition of "logicism", which sought to reduce all AI problems to problems of axiomatization and theorem proving (see McCarthy, 1968, for a classic expression of this point of view; see McDermott, 1987, and Korb, 1995, for two critques thereof). In a kind of self-inflicted reductio, this approach to AI failed to find any reason or value in machine learning (Simon, 1983). However, the logicist tradition is clearly in the decline. The more active research programs in AI today use artificial neural networks, genetic algorithms, and probabilistic reasoning systems, all of which seek to implement inductive inference which copes with uncertain information and complex environments. The ultimate goal of AI is to produce an autonomous artificial agent which can cope with an a priori unknown world; hence, providing competent machine learning is a strict precondition for success.

Assuming that the pessimism of Dreyfus and other antagonists of AI is unwarranted, that, regardless of the difficulty, the human context of scientific reasoning can ultimately be represented algorithmically and scientists' inductive strategies can be implemented on universal machines, then machine learning and the philosophy of scientific method will coalesce. The case I am making is therefore not one of a marriage of convenience, nor of a love affair, but of ultimately the most intimate relationship possible.

References

Bensusan, H. (1999). *Automatic bias learning: An inquiry into the inductive bias of induction.* PhD dissertation, University of Sussex.

Carnap, R. (1962). *Logical foundations of probability*, second edition. University of Chicago.

Craig, W. (1956). Replacement of auxiliary expressions. *Philosophical Review, 65,* 38-55.

Dai, H., Korb, K. B., Wallace, C.S. and Wu, W. (1997). A study of casual discovery with weak links and small samples, *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence,* 1304-1309. Morgan Kaufmann.

Dreyfus, H. (1992). *What computers still can't do: A critique of artificial reason,* third edition. MIT Press.

Franklin, A. (1990). *Experiment, right or wrong.* Cambridge University.

M. P. Georgeff and C. S. Wallace (1984). A general selection criterion for inductive inference. *European Conf. on Artificial Intelligence,* 473-482.

Giere, R. (1973). History and philosophy of science: Marriage of convenience or intimate relationship? *British Journal for the Philosophy of Science, 24,* 282-297.

Glymour, C., and G. Cooper (eds.) (1999). *Computation, causation, and discovery.* MIT Press.

Hacking, I. (1983). *Representing and intervening.* Cambridge University.

Hempel, C. (1958). The theoretician's dillemma. In H. Feigl, M. Scriven and G. Maaxwell (eds.) *Minnesota Studies in the Philosophy of Science.* University of Minnesota Press.

Howson, C. and P. Urbach (1993). *Scientific reasoning: The Bayesian approach,* second edition. Open Court.

Hume, D. (1739/1888). *A treatise of human nature,* edited by L.A. Selby-Bigge. Oxford: Clarendon.

Humphreys, P., and D. Freedman (1996). The grand leap, *British Journal for Philosophy of Science, 47,* 113-123.

Korb, K. (1992). *A Bayesian platform for automating scientific induction.* PhD dissertation, Indiana University.

K B Korb (1995). Inductive learning and defeasible inference, *Journal of Experimental and Theoretical Artificial Intelligence, 7,* 291-324.

K B Korb (1996). Symbolicism and connectionism: AI back at a join point, *Proceedings of the Conference, ISIS'96 Information, Statistics and Induction in Science,* 247-257. World Scientific.

Korb, K.B., and C.S. Wallace (1997). In search of the philosopher's stone: Remarks on Humphreys and Freedman's critique of causal discovery. *British Journal for Philosophy of Science, 48,* 543-553.

Pat Langley, Herbert A. Simon, Gary L. Bradshow, Jan M. Zytkow (1987). *Scientific discovery: Computational explorations of the creative processes.* MIT Press.

Laudan, L. (1987). Progress or rationality? The prospects for normative naturalism. *American Philosophical Quarterly, 24,* 19-31.

McCarthy, J. (1968). Programs with common sense. In M. Minsky (ed.) *Semantic Information Processing*, pp. 403-418. Cambridge: MIT.

McDermott, D. (1987). A critique of pure reason, *Computational Intelligence, 3,* 151-160.

Mascaro, S., Korb, K.B. and Nicholson, A.E. (2000). *Suicide as an evolutionarily stable strategy,* working paper.

Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays,* edited by R. B. Braithwaite. New York: Humanities Press.

Reichenbach, H. (1949). *The theory of probability,* second edition. Berkeley: Univ of California.

Schaffer, C. (1994). A conservation law for generalization performance. In *Proceedings of the 11th International Conference on Machine Learning,* pp. 259-265.

Simon, H. (1983). Why should machines learn? In R.S. Michalski, J.G. Carbonell and T.M. Mitchell (eds.) *Machine learning,* pp. 25-37.

Slezak, P. (1989). Scientific discovery by computer as refutation of the strong programme, *Social Studies of Science, 19,* pp. 563-600.

Sneed, J. (1979). *The logical structure of mathematical physics.* D. Reidel.

R. Solomonoff (1964). A formal theory of inductive inference, I and II, *Information and Control, 7,* 1-22 and 224-254.

Strawson, P. (1999). Dissolving the problem of induction. In Louis P. Pojman (ed.) *The theory of knowledge,* second edition (pp. 502-6).

Thagard, P. (1988). *Computational philosophy of science.* Cambridge: MIT.

Wolpert, D. H., and Macready, W. G. (1995). *No free lunch theorems for search.* Santa Fe Institute Technical Report. 95-02-010.