The Problem of Induction and Machine Learning

F. Bergadano*
University of Torino
Corso Svizzera 185, Torino, Italy
bergadan@di.unito.it

Abstract

Are we justified in inferring a general rule from observations that frequently confirm it? This is the usual statement of the problem of induction. The present paper argues that this question is relevant for the understanding of Machine Learning, but insufficient. Research in Machine Learning has prompted another, more fundamental question: the number of possible rules grows exponentially with the size of the examples, and many of them are somehow confirmed by the data - how are we to choose effectively some rules that have good chances of being predictive? We analyze if and how this problem is approached in standard accounts of induction and show the difficulties that are present. Finally, we suggest that the Explanation-based Learning approach and related methods of knowledge intensive induction could be a partial solution to some of these problems, and help understanding the question of valid induction from a new perspective.

1 The traditional problem of induction

Induction seems to escape all deductive explanations, because its conclusions cannot be proved to be correct. Worse than this, it is not even possible to prove that they are correct *most of the time*, unless we are ready to accept very elaborate and questionable premises. Many conclusions obtained by an inductive process are totally wrong, although infinitely many examples confirm them. Some actually get worse as more confirming evidence is found. The philosophical literature is full of such examples; for instance, let me paraphrase a bit the well known argument of Goodman [Goodman, 1954]. Suppose we define a learning system of "unexpected value" as a system that performs quite badly until August 30, 1991, and then starts to produce incredibly good results. If one was to believe blindly in the power of induction, then an IJCAI-91 paper describing all kinds of very poor results and emphasizing how badly their system works would thus confirm in many ways that the system is of "unexpected value". The more and the more varied the confirming examples that are possible before the IJCAI conference, the worse the conclusion seems to follow.

In this paper, we analyze the problem of induction in a computational framework, where it is possible to make clear the assumptions that we could rely upon when we (or computers) infer general rules that are justified only by a finite number of confirming examples.

When the scope of the enquiry is so restricted, one of the most authoritative approaches to the problem is statistical estimation, as developed, for example, by Neyman. This theory is very well known, but the following will make our later discussion clearer. Suppose, for example, that we are to estimate the mean μ of a given population, that we know to be normal and with standard deviation σ. Let us observe a sample having mean $\overline{x}$. Then, because of the properties of the normal distribution, we may say that, with probability 0.95

$$\overline{x} - 1.96\sigma n^{-1/2} \leq \mu \leq \overline{x} + 1.96\sigma n^{-1/2} \qquad (1)$$

that is to say, we have estimated the value of the real mean with some precision, and we know we are right with high probability. This is a form of inductive reasoning, since we have inferred a general statement about a large population from the observation of a limited number of facts. But the puzzles seem to have vanished because we now know that we are right *most of the time*. This, as Neyman also believes, follows deductively from the premises. Alas, the premises are quite hard to demonstrate and cannot be taken as intuitively plausible, as they involve knowledge of the form of the distribution and of some of their parameters. It seems that what we are required to know before we perform any experiment is much more than what we are actually able to infer. Moreover, this technique deals with the estimation of parameters in a continuous domain, so that approximation with an interval makes sense; it is not clear how this can be adapted to inductive arguments in general.

These problems have led many researchers and practitioners to adopt the alternative approach of subjective Bayesian inference. In this framework, probabilities of inductive hypotheses are defined as subjective degrees of belief and objective data are used to update their values. All that is needed for the updating is the application of Bayes' theorem. The Bayesians defend themselves from the difficulties of subjectivism by arguing that the precise value of prior probabilities is not important: in many

*I am grateful to Paola Dessi', Stuart Russell and Lorenza Saitta for helpful comments on a draft version

cases even large differences in the prior probabilities lead to the same conclusions if sufficient data are provided (e.g. [Edwards *et al.*, 1963]). We will argue later that in many other cases relevant to induction and AI this is not true. Subjective Bayesian induction, as well as traditional statistical reasoning, is often concerned with the inference of numerical parameters when various kinds of continuity assumptions are acceptable. Such assumptions are quite far from the mark when dealing with logical hypotheses such as the ones we work with in many fields of science and particularly in AI.

An approach that emphasizes the logical and symbolic aspects of inductive reasoning is the framework of inductive logic, as developed, for example, by Carnap and Hintikka. In their views, the probability of some statement is defined on the basis of the ground conjunctive expressions that it implies. In Carnap's theory, for instance, the function $C(\phi,\psi)$ indicates the degree, or the strength of the implication $\phi \rightarrow \psi$, and is defined as follows:

$$C(\phi, \psi) = m(\phi \wedge \psi)/m(\psi) \qquad (2)$$

where m indicates some measure of the extension of a formula, that is of how many and how important are the basic statements covered by the formula. Some such measures allow one to justify inductive inference, in the sense that a positive example e of a general rule ϕ will be such that $C(\phi, K \wedge e) > C(\phi, K)$, where K is our previous knowledge. Unfortunately there are, as Carnap finally came to admit, infinitely many such functions, and there is no objective criterion for preferring one over the other. Moreover, the choice of the confirmation function will have profound effects on what one actually deems to be credible on the basis of the very same evidence.

But before we move on to further analysis and criticism, we must discuss the new challenges and the new aspects of induction that have been brought to our attention by the recent research on Machine Learning.

2 New issues raised by Machine Learning

Let us start with a terminological question. Induction, in Machine Learning, is not only taken as the inference from observations to given general rules. It includes the search for these rules in a large set of possibilities. This is not always so in the philosophical literature; for example Peirce seems to be calling "abduction" the choice of the hypotheses, and "induction" the leap from observing the chosen hypotheses work on the available data to accepting them in general. Also when dealing with scientific reasoning, the works on induction are mainly concerned with the testing of some theory which is already given, and strive to justify the conclusions that are derived from the test. We argue, on the contrary, that it is the search and the choice of a plausible inductive hypothesis that is problematic, more than the inductive leap *per se*.

This issue is emphasized in Machine Learning and AI also because of the great importance given to inductive hypotheses of a logical form. The move from numerical to symbolic AI-oriented learning methods has prompted the problem of the very great number of possible rules. Moreover the rules are quite unrelated and distinct: it is hard to see how one rule can be accepted as an approximation of another. The problem at hand is very different from the one of numerical estimation, where the hypotheses are embedded in a continuous space of ordered values. For example, if the standard deviation is 10 and we estimate the mean μ from a sample of size 25, we will regard as practically equivalent the hypotheses μ=12 and μ=13. On the contrary, even if we were to translate a logical space of hypotheses into a numerical representation, e.g. by means of a Gödel numbering, we could by no means consider hypotheses n and $n+1$ to be "close" in any other sense.

Nevertheless, both because of cognitive and knowledge engineering motivations, we cannot set aside our goal of learning symbolic information. The problem of induction is in this case related to the problem of chosing an inductive hypothesis in a large space of distinct possibilities and this choice leads us to a twofold discussion. On the one hand we are to analyze the predictivity of induction from a new perspective, that turns out to be rather pessimistic with respect to the standard account of statistical estimation. On the other, we are concerned with the computational complexity of inductive inference.

2.1 A new understanding of predictivity

This analysis will be limited to the problem of concept acquisition. In this framework we are given concept examples and counterexamples in the form of ground conjunctive expressions, and we search for propositional or first order descriptions of the concepts. For instance, we may be given an example of a "block's world car" as follows:

$$circle(a) \wedge circle(b) \wedge rectangle(c) \wedge on(c,a) \wedge on(c,b) \qquad (3)$$

and we might learn from the above and other examples the following description of the concept "block's world car":

$$\exists x, y[circle(x) \wedge circle(y) \wedge x \neq y] \qquad (4)$$

Although this description may seem insufficient, it could result from the learning process if it distinguishes well enough the examples from the counterexamples of the concept.

If we were given a concept description such as the above and were asked how predictive it is, i.e., how well it could distinguish independent examples and counterexamples of block's world cars, we could use Bernoulli's limitation[1]:

$$Pr\{|observed_error - true_error| > \epsilon\} < 2e^{\frac{-m\epsilon^2}{2}} \qquad (5)$$

were m is the number of examples and counterexamples that we have seen, the ones the observed_error is computed from. True_error is the recognition error that would be observed on all the possible examples[2]. If

[1] This limitation is found in the original proof of his theorem. The constants have been improved upon in an inequality of Hoeffding [Hoeffding, 1963].

[2] There is a finite number of possible examples representable with conjunctive ground expressions without functions, if the number of available predicates and constants is finite, as it usually is.

the number of examples is sufficiently large, the above limitation states that, with high probability, the performance of the given concept description as measured on the available examples is quite close to the performance that we may expect on new examples.

Unfortunately, this is not what we call induction in Machine Learning. This would just be the testing of a hypothesis supplied by an oracle. Normally, this hypothesis is not given and we must search for the best one in a very large number of possibilities. If we are asked how predictive we expect this best hypothesis to be, we cannot use Bernoulli's limitation, which is valid when the recognition rule is fixed, but we must modify it into the following [Vapnik, 1982; Devroye, 1988] :

$$Pr\{\max_{\Phi} |observed_error - true_error| > \epsilon\} < 2|\Phi|e^{\frac{-m\epsilon^2}{2}} \tag{6}$$

where Φ is the set of hypotheses that are possible. Better limitations may be found, but this is the basic idea. From this we see that a large number of possible rules has very bad effects on predictivity. In other words, the larger the language that we use for learning the concept descriptions, the worse the correspondence between the observed and the unknown. Obviously, there *is* an advantage in having a larger set of hypotheses: the observed error rate is likely to be reduced, because by trying more rules we have more chances of finding one that performs well on the training data. Nevertheless, this advantage may vanish when we move to classify new examples.

When performing induction with a machine, one cannot ignore this fact, which is constantly observed in Machine Learning experiments. As a consequence, much emphasis has been given to "bias" or preference criteria [Mitchell, 1982; Bergadano *et al.*, 1989; Bergadano *et al.*, 1988], for choosing only the hypothesis that are deemed plausible $a - priori$.

2.2 The problem of computational complexity

It is not sufficient that learning be predictive, it must also be performed efficiently. If the time needed for obtaining predictive concept descriptions grows exponentially with the size of the examples (e.g. the number of propositional variables), then induction may turn out to be practically unfeasible. Following the work of Valiant [Valiant, 1984a; Valiant, 1984b] there has been a growing interest on these aspects of induction, giving rise to the new field of computational learning theory.

In that approach, the dimension of the problem depends both on the size of the examples and on the accuracy that is required (i.e. the value of ϵ in the above limitations). It can be shown that if the language used for formulating the inductive hypotheses is not adequately restricted, learning accurate hypotheses is NP-hard. For example, under some general assumptions, disjunctive normal form propositional formulas with at most k terms are "not learnable" [Kearns *et al.*, 1987], in the sense that the computation time grows exponentially with the size of the examples and the required accuracy. In other words, it turns out that even relatively simple description languages cause the search for accurate inductive hypotheses to be practically impossible when the size of the examples is large.

It is indeed very interesting that predictiveness and complexity are so closely interrelated. Here is the tradeoff: if the language for expressing inductive hypotheses is simple and contains few alternatives, then we may not be able to find a description that discriminates the available examples, i.e. we cannot adequately describe the observations; if the language is too complex then it may take too long to find acceptable hypotheses and what we find may turn out to perform badly on new data, i.e. we cannot produce effectively reliable predictions of the unobserved.

3 The above problems as addressed by standard accounts of induction

When learning classification rules, we want to estimate the performance (1-*true_error*) of the hypothesis that we have found to be the best for the given examples. For this purpose it is sufficient to evaluate the probable difference ϵ between the error observed on the data and the error that we want to estimate:

$$\epsilon \ s.t. \ Pr\{\max_{\Phi} |observed_error - true_error| > \epsilon\} < \eta \tag{7}$$

where η is what we deem to be a high probability (e.g. 0.95); ϵ will then be the probable performance loss when moving from past to future examples. Limitation (6) allows us to compute ϵ given η, m (the number of learning examples) and $|\Phi|$.

This analysis does not take into account the nature of the hypothesis space Φ, it only counts the number of its elements. It may be the case that the possible hypotheses, while being many, are very similar, in the sense that they classify the possible examples (about) in the same way. In particular, there may be an infinite number of possible hypotheses, but there are at most 2^m ways of classifying a set of m examples. The work of Vapnik and Chervonenkis [Vapnik, 1982] has extended traditional estimation methods to deal with these problems, and provided limitations such as (6), where the cardinality of the hypothesis space is replaced by a measure of its expressiveness. Even if there is an infinite number of possible hypotheses (e.g. the set of linear discriminants), their expressiveness when classifying m examples may be limited, and often much lower that 2^m. Nevertheless, the bounds that may be found following this approach are quite inadequate for the symbolic languages used in Machine Learning [Pearl, 1979; Buntine, 1989; Bergadano and Saitta, 1989]. Suppose that we fix η and ϵ in the above limitation, and we want to known the minimum number m of examples needed for this level of probable performance. It turns out that this number is much larger than the number of examples that ML programs seem to need.

The subjectivist Bayesians have often criticized the traditional approach to statistical inference and proposed an alternative approach that could alleviate some of the above problems [Lindley, 1990; Howson and Urbach, 1989]. To the concept of a sample space they have

substituted subjective prior probabilities, and have defended themselves from the problems of subjectivism in science by noting that, when sufficient data are available, the posterior probabilities will converge to the same values, even for very different initial priors. From the above discussion, it seems that there is a more fundamental problem in inductive inference, when explained in terms of objective probabilities: the arbitrariness of the hypothesis space. The choice of which hypotheses are possible is as crucial as the definition of the sample space (i.e. which examples are possible). If the hypotheses are unrestricted (or even just too many), what we learn will not be probably predictive, if they are few and badly chosen, we will not even describe effectively the available data, and may expect bad performance on new examples as well. The subjectivist approach would substitute prior probabilities of hypotheses to the concept of a hypothesis space. We would not define a minimum number of examples needed to make learning predictive, but will just update the probabilities as we see new examples, making more probable the hypotheses that are confirmed by the data. It would seem that the problem of the prohibitive number of examples is avoided, enabling a learning system to learn even from limited information.

Unfortunately, in symbolic Machine Learning it is no longer true that the choice of the priors is not critical, and is corrected by sufficient data. This is easily seen in the following example. Suppose we give equal prior probabilities to any hypothesis of a propositional language with n variables; then the best classifier on the available examples will always be just the disjunction of the example descriptions. Needless to say, its predictiveness will be null (unless future examples are perfectly equal to the ones that were seen). By contrast, if we give non-zero prior probabilities only to a small, fixed set of hypotheses, the learned description will not be equivalent to the disjunction of the examples and might be predictive according to whether the priors are well chosen and to how many examples we see. In general, if we have not seen most of the 2^n examples, the two above choices of priors will not converge to the same result.

We then need to face the problem of subjectivism, because the irrelevance of the prior is no longer an excuse. Of course, the choice of a sample space is as arbitrary as the choice of a prior probability [Lindley, 1990]. But there is certainly a difference between "subjective" and "arbitrary". The premises of any inference are, in a way, arbitrary, in the sense that one has to chose them on the basis of no other reason at all. The chain of "why" and "because" has to stop at some point and there are assumptions which either correspond to self-evident facts or which we accept without further enquiry. A subjective probability is not an "arbitrary" assumption in this sense. It is not clear when the assumption is true or false, even if we were to obtain the whole factual information that is needed, i. e. the *meaning* of a subjective probability is not clear. But this is only my (subjective) opinion.

The fact that the choice of the priors is in fact very important is connected to another traditional criticism to subjectivist Bayesian induction, which is usually stated as an unrelated question: hypotheses that are constructed explicitly to fit the data should not be confirmed by the very same data. By contrast, Bayesian updating does not distinguish among hypotheses that were generated at different times, or with different intentions. For example, if one Sunday we perform an experiment and we see that our scientific theory is wrong, we should not just correct it by adding that our laws only work during weekdays. Nevertheless, Bayesian updating will increase the probability of the corrected theory. On the other hand, it is hard to explain why one should prefer theories that where generated before seeing any datum. This is the same as saying that the credibility of scientific hypotheses depends on the order some particular scientist performs his actions. The puzzle can be explained, I think, if we understand that the choice of the priors is actually important. If we give non-zero probability to many hypotheses of high complexity, it may happen that inductive inference comes out with rules such as the Sunday-excluding theory. If, on the contrary, we chose only some hypotheses as possible, and we give them non-zero prior probabilities, then we may not be able to obtain high posterior probabilities for any of them. The problem of defining a suitable hypothesis space is essential when learning symbolic concept descriptions, both in the traditional and in the subjectivist accounts of induction.

4 Is Explanation-based Learning a possible answer?

There is an understanding of the word "learning" that is apparently unrelated to induction. For instance, after our first few tic-tac-toe games, we might have "learned" that the second move in Fig. 1 should never be made because it leads to certain loss.

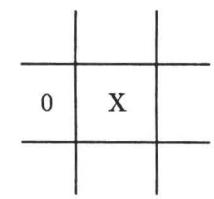

Fig. 1 - Tic-tac-toe example

This is in fact a simple form of Explanation-based Learning (EBL). Here is another example: suppose we are given Peano's arithmetic and the fact "4+3 = 3+4" - then we could "learn" the commutative property of addition. This is in fact a deductive consequence of the theory but "learning" has somehow been guided by the example. As a more general case, suppose we have the following first order theory, defining the predicate P:

$$\forall(x)\,[(R(x) \rightarrow P(x)) \wedge (Q(x) \rightarrow P(x))] \qquad (8)$$

and the ground formula "$Q(a) \wedge P(a)$", which will serve as an example of the concept P. We may now learn (through EBL) a specialized description of P, by sub-

stituting the theory of (8) with

$$\forall(x)\,(Q(x) \rightarrow P(x)) \qquad (9)$$

This example may look trivial, and in fact it is, but I think it shows precisely the basic idea behind EBL. Here the theory (8) plays the role of the rules of tic-tac-toe in the first example and the ground formula "$Q(a) \wedge P(a)$" corresponds to the board situation of Fig. 1. A similar correspondence holds with respect to Peano's arithmetic and the fact "4+3 = 3+4" in the second example. A given theory has a number of deductive consequences (in the extreme case of (8) these consequences are just the two disjunctive components of the theory itself, or their instantiations), we are given some examples or observations about the predicates involved in the theory and we "learn" a subset of the deductive closure of the theory which allows us to explain the observations.

There has been quite a number of interpretations of this form of reasoning from examples to rules with logical constraints that justify the conclusions. Some have argued that EBL is not truly learning, because it is just a form of deduction. Others have stressed, that, although this form of inference is in fact truth-preserving, there is nevertheless some degree of generalization from the available data. For example, in the above tic-tac-toe game, we might have learned not only that this particular move is wrong, but that, in general, any move on a middle-side position is wrong. Similarly, formula (9) is more general than "$Q(a) \wedge P(a)$". Others have compared EBL to lemma generation and to partial evaluation.

These perspectives, though all correct, fail to show that EBL is actually related to induction, and in an important way. Let me put forward an alternative definition: *EBL is inductive inference with a logical definition of the hypothesis space.*

The fact that EBL was not unrelated to induction was noticed early on, when systems started to include statistics of how often a given EBL-generated rule was used; the most frequently used rules were kept in the knowledge base, and the others were discarded. In other words, it was soon understood that, although the learned rule is certainly correct (it follows deductively from the theory), the fact that it will be of some utility on future examples is not guaranteed, and the assumption implies an inductive leap. For instance, it is certain that x+y=y+x, but it is not certain that by adding this rule to Peano's axioms we will prove more efficiently a given theorem. For this reason EBL is often performed with more than one example - for instance the commutative property of addition may be established as another axiom only after seeing more cases were it is useful (e.g. 12+2=2+12 and 2+2*3=2*3+2).

In the usual approach to inductive learning, any hypothesis may be generated, if it explains the data and satisfies general constraints (e.g. being a DNF formula with less than n terms). In EBL, only the hypotheses that may be obtained deductively from the theory are possible. The examples are used to select, among the hypotheses that are possible, the ones that seem to perform well.

If the theory is complete and correct[3], the only risk is that the EBL-generated rules do not represent the most useful part of the theory: these rules are a subset of the deductive closure of the theory, and as such they must continue to be correct, but when they are used alone (without the rest of the theory) they may be incomplete and fail to classify new examples - this will occur more often if the training examples were biased or insufficient.

If the theory is incorrect and/or incomplete[4], then this only means that a 100% correct rule for our learning examples may be excluded from the hypothesis space. This corresponds to the general stochastic classification problem in Pattern Recognition, and is not necessarily a harmful situation if the hypothesis space is well chosen and of a limited size. Restriction to correct hypotheses (the deterministic problem) often leads to larger search spaces and degrading performance [Bergadano and Saitta, 1989][5]. This is apparently contradicted by the fact that in the deterministic case we can obtain formulas such as (6), where ϵ^2 is substituted by ϵ [Vapnik, 1982; Blumer *et al.*, 1987], leading to better limitations of the number of examples needed for a specified performance. Nevertheless, in order to be sure that a 100% correct rule is possible, the hypothesis space has to be larger, and, as a consequence, the difference between the observed error (which is 0) and the true error grows. This is also observed experimentally when pruning decision trees [Quinlan, 1987] or simplifying logical descriptions [Bergadano *et al.*, 1989].

The possibility of describing a hypothesis space in a knowledge-based style is an important advantage of EBL and other forms of declarative bias (e.g. determinations [Russell, 1988])[6]. It allows us to inform our inductive procedures of the basic constraints that are present in a given domain, e.g. the meaning of high level features usually employed by experts, or maybe the results of automated learning on another, similar, domain[7]. We have argued that the definition of a good hypothesis space is responsible for the difficulty of induction and some-

[3] In EBL, a domain theory is said to be complete when it is able to classify any example. The theory is correct if all the classifications that are produced correspond to the a priori classification of the examples.

[4] There has been substantial work on EBL with incorrect and incomplete theories, for example many papers in the 1989 Machine Learning workshop on "Combining Empirical and Explanation-based Learning" are devoted to this problem (Proceedings published by Morgan Kaufmann)

[5] Induction in the deterministic and in the stochastic case have also been distinguished in the philosophical literature, for instance Mill calls the latter "approximate generalization" and Keynes speaks of "universal induction" and "inductive correlation", respectively. Keynes discusses inductive correlation separately, in the context of statistical inference in the last part of his Treatise on Probability. Here we basically view the two cases as instances of the same problem

[6] An alternative perspective that distinguishes EBL from declarative bias is found in [Russell, to appear]

[7] This feature is related to what is called "local induction" in the philosophical literature [Kyburg, 1976], i.e. induction that is based on knowledge that was itself obtained with some form of inductive reasoning.

times the reason of the paradoxes of inductive inference. Therefore, we must devote our efforts to this goal, acquire knowledge about the domain of a given inductive problem, understand the basic constraints, and transform all this into a domain theory, the logical definition of the hypothesis space.

5 Conclusion

Of course, this is a good start, but it is not a solution of the problem. With EBL we have, in some sense, a good programming language for describing the hypothesis space. But how to write a good program? How to know which hypotheses we should regard as possible? The statistical analysis which we have discussed above might give us some advice about how many these hypotheses should be, if we want to avoid total degradation of performance when moving to test examples. But this advice, which has been compared to Occam's razor [Blumer *et al.*, 1987], is so pessimistic, that it is more adequately associated to an axe. According to theoretical analysis, even when quite many examples are available, we should restrict our search to very few hypothesis; as a consequence, we will easily obtain bad performances even on the training data. Maybe, better advice could be obtained by experimental analysis, with cross-validation techniques (such as leave-one-out), in order to evaluate the probable performance loss.

In any case, all these analyses might well tell us how many hypotheses we should have, they will never tell us which. Better understanding of the inductive problem and more research in Machine Learning might allow us to transfer knowledge of relevant features from one domain to the other, to explore the hypothesis space more efficiently, to collect more data in order to save predictiveness even if our previous knowledge is poor. But it will always be important, and sometimes necessary, to describe a hypothesis space which is both expressive and reasonably constrained. It is not surprising, I think, that knowledge of nothing at all does not lead to anything useful, in induction as well as in any other kind of inference. All forms of reasoning need premises that do not follow from anything else. "The starting point of scientific knowledge is not itself scientific knowledge. Therefore, since we possess no other infallible faculty besides scientific knowledge, the source from which such knowledge starts must be intuition", or, I would rephrase, some form of a lucky guess. Good luck!

References

[Bergadano and Saitta, 1989] Bergadano, F. and Saitta, L. 1989. On the Error Probability of Boolean Concept Descriptions. In *Proc. IV European Working Sessions on Learning*, Montpellier, France. Pitman. 25–36.

[Bergadano *et al.*, 1988] Bergadano, F.; Giordana, A.; and Saitta, L. 1988. Automated Concept Acquisition in Noisy Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(4):555–578. New York.

[Bergadano *et al.*, 1989] Bergadano, F.; Matwin, S.; Michalski, R. S.; and Zhang, J. 1989. Learning Flexible Concepts through a Search for Simpler but still Accurate Descriptions. In *Proc. of the second workshop on knowledge acquisition*, Banff, Canada. 4.1–4.10.

[Blumer *et al.*, 1987] Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. 1987. Occam's Razor. *Information Processing Letters* 24:377–380.

[Buntine, 1989] Buntine, W. 1989. A critique of the Valiant Model. In *Proc. of the IJCAI*, Detroit, Mi. 837–842.

[Devroye, 1988] Devroye, L. 1988. Automatic Pattern Recognition: A Study of the Probability of Error. *IEEE Trans. on PAMI* 10(4):530–543.

[Edwards *et al.*, 1963] Edwards, W.; Lindman, H.; and Savage, L. J. 1963. Inference for Psychological Research. *Psychological Review* 70:193–242.

[Goodman, 1954] Goodman, N. 1954. *Fact, Fiction and Forecast.* The Athlone Press, London.

[Hoeffding, 1963] Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58:13–30.

[Howson and Urbach, 1989] Howson, C. and Urbach, P. 1989. *Scientific Reasoning - the Bayesian Approach.* Open Court, Illinois.

[Kearns *et al.*, 1987] Kearns, M.; Li, M.; Pitt, L.; and Valiant, L. 1987. On the learnability of boolean formulae. In *Proc. of the fourth Machine Learning workshop*, Irvine, CA. 285–295.

[Kyburg, 1976] Kyburg, H. 1976. Local and global induction. In Bogdan, , editor 1976, *Local Induction*, Amsterdam.

[Lindley, 1990] Lindley, D. V. 1990. The present position in Bayesian Statistics. *Statistical Science* 5(1):44–65.

[Mitchell, 1982] Mitchell, T. M. 1982. Generalization as Search. *Artificial Intelligence* 18:203–226.

[Pearl, 1979] Pearl, J. 1979. Capacity and error estimates for boolean classifiers with limited complexity. *IEEE Trans. on PAMI* 1(4):350–355.

[Quinlan, 1987] Quinlan, R. 1987. Simplifying Decision Trees. *Int. J. of Man Machine Studies* 27:221–234.

[Russell, 1988] Russell, S. 1988. Tree-structured bias. In *Proc. AAAI Conference.* 641–645.

[Russell, to appear] Russell, S. Inductive learning by machines. *Philosophical Studies.*

[Valiant, 1984a] Valiant, L. G. 1984a. A Theory of the Learnable. *Communications of the ACM* 27(11):1134–1142.

[Valiant, 1984b] Valiant, L. G. 1984b. Deductive learning. *Phil. Trans. Royal society of London* 312:441–446.

[Vapnik, 1982] Vapnik, V. 1982. *Estimation of Dependencies Based on Empirical Data.* Springer Verlag, New York.