Some Philosophical Problems with Formal Learning Theory

Jonathan Amsterdam*
MIT Laboratory for Artificial Intelligence
Cambridge, MA 02139

Abstract

Recent work in formal learning theory has attempted to capture the essentials of the concept-learning task in a formal framework. This paper evaluates the potential contributions of certain kinds of models to the study of learning by exploring the philosophical implications of these models. Some of my remarks bear on mainstream AI learning techniques such as version spaces and explanation-based learning.

1 Introduction

Recently there has been a renewed interest in formal learning theory, due largely to Leslie Valiant's paper "A Theory of the Learnable" [1984]. Since a formal model of learning provides a clear and precise definition of learnability, results in the model could have considerable impact on the study of human and machine learning. It might, for example, indicate limits on what is efficiently learnable by people or computers, much as computability theory has done for computation. It is important, therefore, to assure that the definition of learnability and the assumptions of the model are reasonable. Here I consider several problems, largely philosophical, with the assumptions of the Valiant model and related models of concept learning. I begin by providing a brief overview of the Valiant model for concreteness.

2 The Valiant Model

The following provides only the briefest sketch of the model; for more detail, see [Kearns *et al.*, 1987b; Valiant, 1984].

We assume a space X of examples, with a probability distribution D imposed on X. A concept is a subset of X, a concept representation is a description of a concept, and a concept class is a set of concept representations. The situation modeled is that of a learner trying to acquire a particular concept, called the target concept, drawn from a concept class. The learner can examine examples drawn at random from X according to D; each example is labeled + or − according to whether it is a member of the target concept.

The learner takes as input the size of the target concept representation and two parameters ϵ and δ, with $0 < \epsilon, \delta \leq 1$. The learner must quickly produce a concept representation in the concept class that is close to the target concept with high probability. More precisely, the learner must produce, with probability $1 - \delta$, a concept representation that can disagree with the target concept with probability at most ϵ on examples drawn randomly from X according to D, and it must accomplish this task in time polynomial in the size of the target concept representation, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. If there is an algorithm that can accomplish this task for any target concept in the concept class and any distribution D, then the class is learnable.[1]

One concept class that is learnable is the class of conjunctions of Boolean variables; in this case, X contains vectors of truth-assignments to the variables. Other learnable concept classes include k-CNF and k-DNF [Valiant, 1984], decision lists [Rivest, 1987], and a subclass of linear threshold functions [Littlestone, 1987]. Several classes are also known to be unlearnable, assuming $RP \neq NP$ [Kearns *et al.*, 1987a].

Although the model defines only the learnability of concept classes, the extensions to other senses of 'learn' are obvious. So, if a program outputs a concept that is within ϵ (in the sense defined above) of the target concept, then it is in the spirit of the model to say that the program has learned the concept.

3 Concepts and the Possible

My first criticism applies to probabilistic models of concept learning, like the Valiant model, that interpret the probabilities in a certain way. In his original paper, Valiant says that the probability distribution "is intended to describe the relative frequency with which the... examples... occur in nature" [Valiant, 1984, p. 1136]. But accuracy on only the naturally occurring examples of a concept is rarely sufficient for its acquisition. An agent who succeeded admirably in classifying existing instances but failed on hypothetical ones might be considered to have a good recognition method for the concept, but could not be said to have learned it.

Consider the concept 'bachelor,' which we can assume to be defined as 'unmarried male' for the time being. Assume further what is almost certainly the case, that very few bachelors wear wedding rings. Still, if you are asked "Is an unmarried male wearing a wedding ring a bachelor?" you will reply in the affirmative. Bachelors are just unmarried males; wedding rings don't enter into it.

Now say a robot tries to learn the concept 'bachelor' from natural examples and manages to acquire the concept

*The author was supported by an ONR Fellowship. ARPAnet address: jba@wheaties.ai.mit.edu

[1] This description is a simplification of Valiant's original definition in several ways. The differences play no role in what follows.

'unmarried male who does not wear a wedding ring.' The robot will classify most existing things correctly with this description, and so has learned 'bachelor' as far as the Valiant model is concerned. But *we* would not say the robot has learned the concept 'bachelor,' for it will answer the above question incorrectly, even though it may perform excellently—perhaps better than most people do—in the task of identifying real-world bachelors.

Note that this problem does not arise because the Valiant model allows a small error in the learner's concept. For we can assume that no bachelors currently wear wedding rings, hence that the robot classifies existing bachelors perfectly; but we still would not say it had learned 'bachelor,' because it fails on hypothetical cases.

While the robot's concept is coextensive with 'bachelor'—it picks out the same set of things in the real world—it differs from 'bachelor' in hypothetical worlds, *and this matters to us.* We do not count two concepts as the same if they are merely coextensive. Nor would doing so be a wise idea: we require our concepts to be meaningful in possible situations so we can be understood when discussing these situations. When considering the bachelorhood of a hypothetical unmarried man wearing a wedding ring, we do not want our deliberations affected by whether an unmarried man has ever done such a thing, or even whether the situation is physically possible (perhaps the man in question is a Lilliputian). The point extends beyond daily life into science: to discuss alternative theories meaningfully, there must be some constancy of concepts across the theories, even though at most one theory can be right. For example, if the bachelorhood of Lilliputians ever became a factor in the evaluation of rival physical theories, one would hope that the impossibility of Lilliputians in one theory would not prevent its adherents from understanding the other theory's point of view.

Valiant takes the interpretation of the probability distribution as over only the natural examples to carry some philosophical weight:

> A learnable concept is nothing more than a short program that distinguishes some natural inputs from some others. If such a concept is passed on among a population in a distributed manner, substantial variations in meaning may arise. More importantly, what consensus there is will only be meaningful for natural inputs. The behavior of an individual's program for unnatural inputs has no relevance. Hence thought experiments and logical arguments involving unnatural hypothetical situations may be meaningless activities [Valiant, 1984, p. 1142].

It is certainly true that some situations tug our intuitions both ways: what if your unmarried male friend was actually the product of a synthetic sperm and egg? Philosophers love such borderline cases because they can help to tease apart the many threads that run through even the most basic of our concepts, like 'person.' To call these thought-experiments meaningless because they fail to conform to a model which cannot capture even the clear-cut cases seems to be getting things the wrong way round.

4 The Classical Theory

In this section, I consider a more far-reaching argument that implicates many concept-learning schemes, not just formal models but also more traditional AI approaches such as version spaces [Mitchell, 1981]. The criticism claims that concepts cannot be defined in the so-called 'classical' way, by necessary and sufficient conditions for membership. Usually this is intended as a psychological criticism [Lakoff, 1987; Schank *et al.*, 1986; Smith and Medin, 1981], but an ontological argument can be made as well [Lakoff, 1987, ch. 12]. The psychological argument claims that humans do not employ concepts defined by necessary and sufficient conditions; the ontological argument claims, roughly, that no such concepts exist in the world.

4.1 Psychological Critique

The relevant aspect of the psychological argument is the claim that human concept representations are not *bivalent*—they do not classify every object as either inside or outside the concept. A body of pyschological evidence that has been accumulating for fifteen years indicates that human concepts cannot be described bivalently [Smith and Medin, 1981]. The evidence includes demonstrations of typicality effects, and inconsistencies across and within subjects on classification tasks. The results seem to show that concept membership is a matter of degree: some birds, like robins, are better examples of the concept 'bird' than others, like penguins.

These results call into question the the psychological plausibility of models that use bivalent concept representations. To account for them, researchers have proposed concept representations based on *prototypes,* highly typical concept members; these representations allow for degrees of membership that vary with proximity to the prototype.

This critique is not immediately conclusive against models, like Valiant's, that permit a wide range of concept classes. It is true that Boolean functions, the class of representations used most often in literature on the Valiant model, are bivalent. But the model is compatible with other representations. In fact, both linear threshold functions and hyperspheres in Euclidean space could form the basis for psychological models of concepts based on prototypes [Smith and Medin, 1981]. (Interestingly, the class of hyperspheres is learnable in the Valiant sense [Amsterdam, 1988], as is a restricted class of linear threshold functions [Littlestone, 1987].) Both representations divide naturally into two components: a method to compute a graded (i.e. real-valued) measure of inclusion in the concept, and a threshold which uses the computed measure to determine whether the example is a member of the concept. For linear threshold functions, the sum of weighted attribute values provides the measure of inclusion; if the sum exceeds a threshold, the example is considered to be in the concept. For hyperspheres, the Euclidean metric provides the measure of inclusion, and the hypersphere's radius the threshold.

It is plausible to view the inclusion measure as the actual, graded, concept definition, and the threshold as merely an artifact forced by the Valiant model's requirement of bivalence, akin to experimenters' insistence that their subjects answer concept inclusion questions with a 'yes' or 'no'. Still, the Valiant model is inconsistent with the fact that the same subject will sometimes give conflicting answers to the same question about concept inclusion [McCloskey and Glucksberg, 1978], because in the model a given concept will always classify an example the same way.

Some solace may be found in [Armstrong *et al.*, 1983]. They show that typicality effects occur even for obviously bivalent concepts like 'even number'; in fact, subjects will rate some numbers 'more even' than others after explicitly stating that membership in 'even number' does not admit of degree. This result suggests that typicality data can shed little light on concepts' membership criteria. Armstrong *et. al.* propose a picture of human concepts that consists of a bivalent *core,* which determines membership, associated with a collection of heuristic identification procedures whose interactions give rise to typicality effects. Bivalent concept-learning models could be seen as models of concept core acquisition. Two problems inhere in this suggestion: first, it may be that many concepts are coreless, hence excluded from the purview of bivalent models; and second, many models, Valiant's included, are concerned with the process of identifying examples, and this is not the role of the core but of the identification procedures. So although the issue is far from settled, it would seem that bivalent concept-learning models are not psychologically plausible.[2]

4.2 Ontological Critique

One could claim in response that the Valiant model can still be used to obtain correct, albeit psychologically implausible, definitions of many concepts. So even though the definitions of 'bird' that are produced do not fit the psychological data, they nonetheless classify birds well.

While immune from the psychological critique, this response assumes that there is *some* description that provides necessary and sufficient conditions for birdhood, because the Valiant model depends on there being some target concept that classifies the examples seen by the learner.

The ontological critique questions this assumption. It claims that for most, if not all, empirical concepts, there is no necessary-and-sufficient description that is couched in terms of reasonable input features. An example should clarify this. Say we wish to separate seedless grapes from others at a grape processing plant. Now we could define 'seedless grape' as 'grape without seeds,' but we wish to perform the classification without actually cutting the grapes open and looking for the seeds. The features we might use are overt sensory ones like color and shape as well as more esoteric ones like genetic analysis of the vine of origin and X-ray pictures of the grapes' insides. It would be quite amazing if some function of these features provided necessary and sufficient conditions for 'seedless grape.' The presence of the gene responsible for seedlessness will not do, because there are probably numerous environmental factors that would interfere with its expression, or conversely, that would give rise to seedless grapes on a genetically seeded vine. The X-ray picture fails because other structures in the grape (possibly introduced unnaturally) might cast identical X-ray shadows, and because the concept 'X-ray shadow of a grape seed' is itself subject to the same criticism, where the features are the X-ray image pixels.

The point is not that we can never be absolutely certain of our categorizations; this is true and uninteresting. Rather, even if we were certain about the DNA sequence of the vine of origin and the intensity of every pixel in the X-ray image, we still would not be able to define 'seedless grape' from these features. The point is also not the mundane (though very important) one that we rarely know all the relevant features, but rather that there is no finite set of relevant features; cosmic rays, cropdusting, and schoolboy pranks may all play a part.

This is a severe philosophical criticism of the Valiant model, for even though the model allows the learner to merely approximate the target concept, it still assumes that there is a target concept defined by a representation in the concept class. For most cases in which the input features are not trivially definitive of the concept, this assumption appears untenable.

5 Learning Language

Let us turn from these general considerations for a moment to consider a particular example of concept learning, in order to bring out an interesting way in which many formal models fail. Learning the grammar of a natural language is an appropriate choice, for it is a major problem in cognitive science and has inspired several formal models [Gold, 1967; Wexler and Culicover, 1980].

There are numerous problems with formal approaches to language acquisition, many of them discussed elsewhere [Chomsky, 1986]. I wish to raise only two. First, the class of human languages might be finite, a conclusion suggested by the parameter theory of language developed by Chomsky [1981], in which natural-language grammars are distinguished only by the settings of a few switches. Finite concept classes are trivially learnable in models based on asymptotic behavior, including the Valiant model with its requirement of polynomial-time learnability. This triviality is misleading, however, because language learning poses some serious problems. In particular, it involves the simultaneous acquisition of multiple, interdefined concepts. The ramifications of this observation are, I think, most far-reaching. They call into question the appropriateness not only of the Valiant model, but of much of the concept-learning paradigm as practiced in AI.

The problem can be seen most clearly by viewing parameter-style language acquistion in the Valiant model's terms, ignoring the finiteness problem for the moment. If we think of the parameters as features, the learning problem is then very simple and straightforward: just set the parameters to match the features in the examples. Hence there would seem to be little to say about learnability. But the problem is that the features are not given in the examples; most of the parameters that have been conjectured are fairly abstract, employing terms such as 'phrase,' 'head'

[2] See [Lakoff, 1987; Smith and Medin, 1981] for discussion at great length. It should be realized that many of the traditional arguments against the classical view really attack the much weaker position that concept definitions are *conjunctions* of features.

and so on. And even concepts like 'noun' and 'word' are not part of the learner's (that is, the young child's) raw input. That input consists primarily of streams of sounds and, in cases where the learner's perceptual machinery can perceive the content of an utterance, simple 'meanings.' For example, if 'Mary hits John' is uttered while the action of Mary hitting John is visible to the learner, then it is assumed that the learner can decompose the event into two objects (Mary and John) and an action (hitting); this decomposed event would be the 'meaning' of the sentence.

Because the learner's input is so far removed from the concepts in which the innate grammar is couched, it would seem that language learning is quite difficult. How is it accomplished?

It appears that we would have trouble even in phrasing the question in terms of models of single-concept learning, since language acquisition is patently a multi-concept problem. But single-concept learning models could provide an answer to this question. One might claim that the lowest-level concepts, like 'word,' are learned directly from the input. These concepts then become features themselves, and another round of concept-learning occurs which results in the acquisition of concepts at the next level, such as 'noun'.[3]

This reductionistic account is simple and obvious, but it cannot be right. There is no way to define concepts like 'word' and 'noun' from the features available to the language learner. Consider 'noun'. The nounhood of some words might be deduced from the input; for instance, that 'John' is a noun might be determined from the fact that it often occurs in conjunction with a particular part of 'meanings'—namely the object John himself. Indeed, this is probably how acquisition of nouns and verbs is started. But it cannot be the whole story, for some nouns, like 'ride' and 'strength,' do not correspond to physical objects and so cannot be assumed to be available in the child's perceptual input. Likewise, some verbs, like 'resemble,' do not correspond to actions.

According to one theory of language acquisition, children tentatively classify some words as nouns and verbs because they seem to correspond to objects and actions present in 'meanings.' This enables the parsing of very simple sentences. They then use this knowledge to set parameters,[4] which in turn allows them to parse more complex sentences. Information from such parses can then serve to classify other words, like 'ride' and 'resemble,' that do not occur in 'meanings'. This picture of language acquisition is known as 'semantic bootstrapping' [Pinker, 1984].

Note that this process is quite different from the hierarchical learning procedure that seems natural for single-concept learning models. Concepts are not acquired hierarchically, but rather piecemeal; and partially learned concepts can help the learning process by driving theory (here, grammar) construction.

Language is not the only domain that exhibits this bootstrapping pattern. Recent work has shown that children may acquire common-sense biological knowledge in the same manner [Carey, 1985]. And, most crucially, science in general works this way, the formation of new concepts suggesting further experiments and making additional data available. On this view, concept acquisition is theory formation and revision. Concepts are not composed, layer by layer, from more primitive, already acquired concepts; instead, the whole cluster of concepts forms a complexly interacting web with no clear levels. The task of acquiring a single concept is at best an idealization, for in learning a new concept we will almost certainly alter others, so that our beliefs are as consistent, coherent and accurate as we can make them [Quine, 1971].

This observation extends the ontological critique, which held that concepts are not definable from the input. It shows that most concepts are not even definable in terms of other concepts; the relationships between concepts in a theory, and between theory and data, are not relationships of definition. We cannot define 'electron' in terms of observable properties, nor even in terms of other concepts of physics; but 'electron' is a part of the web, connected to 'particle,' 'quark' and 'positron' by links of implication, links that are defeasible in the face of overwhelming data or pragmatic considerations. If the point is obscure in physics, consider 'seedless grape' again. I said earlier that we can define 'seedless grape' as 'grape without seeds.' While this may be appropriate for our classification task, the connections among 'grape,' 'seed' and 'seedless grape' are in fact much more subtle and subject to considerations not only of practical utility, but of scientific parsimony as well. 'Seedless grape' might be a genetic term to a molecular biologist, or a species designation to an evolutionary biologist; and in some overarching, unified biology (should such a thing exist), it might have a different character entirely. Because of mutation, genetic tampering or cross-breeding, a seedless grape with seeds might not be a contradiction. The standard picture of single-concept learning that operates against a fixed background of theory and data cannot account for these facts.

Although some work has addressed these issues, especially work on discovery systems [Haase, 1986; Lenat, 1983], the problems are enormous and largely unexplored. How is the web of interconnecting concepts structured? How is blame apportioned when contradictions are discovered? When should contradictions be ignored or papered over rather than repaired? And how can experiments best be designed to resolve contradictions and distinguish rival explanations?

If much of our learning is best characterized as a process of theory formation and revision, where the multiple concepts of a theory interact with each other, with pragmatic constraints, and with the data in complex ways, then the single-concept learning model that has been with us for some time may be a poor model for all but a few learning tasks. And if, as its ubiquitous use in science suggests, this complex process is either inevitable, or superior to a hierarchical one for acquiring knowledge, then it would seem that single-concept learning may be a practical tool of only limited use.

6 Conclusion

I have considered several ways in which certain learning models fail to capture important learning phenomena. The

[3] Strictly speaking, we are defining the concepts 'word-in-L,' 'noun-in-L' etc. where L is the language being learned.

[4] Or, in other theories of grammar, to acquire rules or constraints. The story is not tied to the parameter theory.

requirement for mere extensional equivalence of concepts cannot account for performance on hypothetical situations; the assumption that concepts can be defined is on shaky ground; and single-concept learning is likely an unusual special case of the theory-construction process.

Some of my criticisms implicate a broad class of work in machine learning. The idea that interesting concepts can be characterized by definitions is presupposed not only by most formal models, but also by techniques like version spaces [Mitchell, 1981] and explanation-based generalization as described in [Mitchell *et al.*, 1986]. Happily, many machine learning researchers are moving away from this conception. It is important to realize that not all the power of probabilistic models like Valiant's is lost in this retrenchment: we can use statistical techniques to verify that a hypothesis is accurate with high probability without making any assumptions about the definability of a target concept [Etzioni, 1988]. What needs to be given up, or at least considerably diluted, is the idea of completeness—that a learning algorithm will always (or almost always) produce an accurate hypothesis. Such results invariably assume the existence of definable concepts.

The attention granted single-concept learning has resulted in some useful techniques, but the paradigm does not scale up cleanly to multiple concepts; rather, it is a special case, probably a rare and unrepresentative one. Single-concept learning does go on in practice, at least to a first approximation, and its techniques may serve as useful modules in larger learning systems; but the more fundamental and interesting problems center around the interaction of many concepts in the course of theory construction.

Acknowledgements

I thank David Chapman, William Gasarch, Melanie Mitchell, Ron Rivest, Orca Starbuck, Patrick Winston and especially Oren Etzioni for helpful comments and discussion.

References

[Amsterdam, 1988] Jonathan Amsterdam. *The Valiant Learning Model: Extensions and Assessment.* Master's thesis, MIT, January 1988.

[Armstrong *et al.*, 1983] S. L. Armstrong, L. R. Gleitman, and Henry Gleitman. What some concepts might not be. *Cognition*, 13:263–308, 1983.

[Carey, 1985] Susan Carey. *Conceptual Change in Childhood.* MIT Press, Cambridge, Mass., 1985.

[Chomsky, 1981] N. Chomsky. *Lectures on Government and Binding.* Foris, Dordrecht, Holland, 1981.

[Chomsky, 1986] N. Chomsky. *Knowledge of Language: Its Nature, Origin and Use.* Praeger, New York, 1986.

[Etzioni, 1988] O. Etzioni. Hypothesis filtering: a practical approach to reliable learning. In *Proceedings of the 5th International Workshop on Machine Learning*, Morgan Kaufmann, 1988.

[Gold, 1967] E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[Haase, 1986] K. W. Haase. Cyrano: A Thoughtful Reimplementation of Eurisko. In *Proceedings of ECAI-86*, 1986.

[Kearns *et al.*, 1987a] Michael Kearns, Ming Li, Leonard Pitt, and Leslie Valiant. On the learnability of Boolean formulae. In *Proceedings of the 19th Symposium on the Theory of Computing*, pages 285–295, ACM, 1987.

[Kearns *et al.*, 1987b] Michael Kearns, Ming Li, Leonard Pitt, and Leslie Valiant. Recent results on Boolean concept learning. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 337–352, Morgan Kaufmann, 1987.

[Lakoff, 1987] George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press, Chicago, 1987.

[Lenat, 1983] D. B. Lenat. Eurisko: A program which learns new heuristics and domain concepts. *Artificial Intelligence*, 21, 1983.

[Littlestone, 1987] N. Littlestone. Learning quickly when irrelevant attributes abound. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 68–77, IEEE, October 1987.

[McCloskey and Glucksberg, 1978] M. McCloskey and S. Glucksberg. Natural categories: well defined or fuzzy sets? *Memory & Cognition*, 6(4):462–472, 1978.

[Mitchell, 1981] T. Mitchell. Generalization as search. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, Tioga, Palo Alto, CA, 1981.

[Mitchell *et al.*, 1986] T. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80, 1986.

[Pinker, 1984] Steven Pinker. *Language Learnability and Language Development.* Harvard University Press, Cambridge, Mass., 1984.

[Quine, 1971] W. V. O. Quine. Two dogmas of empiricism. In Jay F. Rosenberg and Charles Travis, editors, *Readings in the Philosophy of Language*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

[Rivest, 1987] R. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, November 1987.

[Schank *et al.*, 1986] R. C. Schank, G. C. Collins, and L. E. Hunter. Transcending inductive category formation in learning. *Behavioral and Brain Sciences*, 9:639–686, 1986.

[Smith and Medin, 1981] E. E. Smith and D. L. Medin. *Categories and Concepts.* Harvard University Press, Cambridge, Mass., 1981.

[Valiant, 1984] L. Valiant. A theory of the learnable. *Communications Of The ACM*, 27(11):1134–1142, November 1984.

[Wexler and Culicover, 1980] K. Wexler and P. W. Culicover. *Formal Principles of Language Acquisition.* MIT Press, Cambridge, 1980.