Is there any Need to Mention Induction?

Chris Thornton
COGS/Informatics
University of Sussex
Brighton
BN1 9QH
UK
c.thornton@sussex.ac.uk

Abstract

Induction is the process by which seen data becomes the basis for prediction of unseen data. There has long been a desire to explain the procedure in a context-free way. But Hume's circularity problem and the no-free-lunch theorems both seem to suggest the logical impossibility of any context-free mechanism. Machine Learning takes the position that no such mechanism exists. But an alternative comes from Epistemology. Popper's falsificationist theory holds that there is a general mechanism, but that it does not *perform* induction. Inductive effects arise implicitly, through pursuit of a non-inductive goal. Less plausibly, the mechanism is taken to be uninformed exploration of hypotheses. But as the paper shows, Popper's solution can be reworked using information theory. Increasing the informational efficiency with which representations predict seen data can be shown to produce inductive effects. With representation optimization taking the place of hypothesis-search in the argument, it then becomes possible to explain induction in a context-free way.

Keywords: problem of induction; no free lunch (NFL); cognitive informatics; theoretical cognitive science

Introduction

Induction has traditionally been equated with the way scientists use observations to form predictions. But a broader view is now taken. Induction is understood to be any process by which seen data comes to be used for prediction of unseen data. On this basis, the procedure is understood to play a key role not only in science, but also in cognition. Embedded inductive functionality is also observed in sensory, perceptual and adaptive phenomena (Smith, 2000). We even see inductive effects in the behaviour of certain forms of plant life. The behaviour of young sunflowers, for example, exhibits heliotropism (sun-tracking), involving implicit prediction of the sun's trajectory across the sky.

But recognizing that inductive functionality can be expressed in a number of different ways does not significantly improve the prospects for explaining how it works. Understanding induction to involve use of seen data for predicting unseen data, we are led to envisage a mechanism that exploits the latter being in some way *uniform* with the former. Unfortunately, any attempt to explain the process in these terms runs into difficulties. First, there is Hume's 'circularity' problem. Any understanding about uniformity between seen and unseen data must itself be inductively derived. An argument which references that understanding is necessarily circular. As Hume noted, such explanations [1] end up 'going in a circle, and taking that for granted, which is the very point in question' (Hume, 1988/1748, p. 80).

In the modern era, the no-free-lunch (NFL) theorems (Wolpert and Macready, 1997; Wolpert, 1996), and conservation law of generalization (Schaffer, 1994) have added fuel to the fire, making it doubtful that *any* context-free principle could exist. Any such principle should yield induction in all possible scenarios, it is observed. But any principle tested in all possible scenarios is found to perform at the level of random guessing *on average*. A completely general principle of induction would then seem to have no value for induction *in general*. Again, the implication seems to be that no such principle can exist.

As a result of such arguments, it has become widely accepted that there can be no context-free basis for preferring one inductive model over another (Langley, 1996; Mitchell, 1997). The outcome has been readily accommodated in machine learning. The fact that multiple models can be derived in different ways from given data (Michie *et al.* 1994) is seen to be the reason why each 'learning algorithm has a different inductive bias, makes different assumptions, and defines a different objective function' (Alpaydin, 2010, p. 309). Neither has the problem held back theoretical work. The uniformity assumption that Hume sees as problematic becomes a methodological *requirement* in computational learning theory (Bishop, 2007). This allows induction to be treated as the problem of sampling in a known (e.g., IID) distribution (Mackay, 2003).

In other areas, more difficulty is encountered. Consider the effect on Epistemology. The lack of a universal principle for induction means there can be no assumption-free basis for inductively-derived knowledge. All of science then seems reduced to the status of guesswork. Russell appraises the situation in vivid terms. Our inability to identify any general principle of induction suggests 'there is no intellectual difference between sanity and insanity', and that scientists are on an equal footing with 'the lunatic who believes that he is a poached egg' (Russell, 1946, p. 673).

[1] Hume referred to predictions of cause/effect relationships rather than induction in general.

The difficulties are not merely epistemological in nature, however. The idea that induction performed in a context-free way is necessarily *unprincipled* leads to counter-intuitive conclusions. Any agent (or embedded module) embarking on the process must then be making some kind of random choice between context-sensitive approaches. This seems at odds with what is observed in the natural world.

Such are some of the problems stemming from the much-debated *problem of induction* (Sloman and Lagnado, 2005). Can any viable solution for the problem be worked out? Prominent among the proposals put forward is Popper's *falsification* framework (Popper, 1959; Popper, 1979). This solve's Hume's circularity problem by showing how quasi-inductive effects can arise implicitly, through application of a procedure that makes no assumptions about uniformity. The vehicle proposed is *falsification*: systematic elimination of refuted hypotheses. This is put forward as a fully context-free method for achieving (implicitly) inductive results without requiring any (explicitly) inductive step to be taken. There is then 'no need even to *mention* induction' (Popper, 1979, p. 315) in explaining how it works.

A difficulty for Popper's proposal is that hypothesis falsification is not seen to be a plausible vehicle for induction in general. It does not reflect the real practices of science (Kuhn, 1962; Lakatos, 1970), is unworkable where the number of hypotheses is large (Hempel, 1945; Churchland, 1986; Duhem, 1954/1914; Putnam, 1974; Quine, 1953)[2] and seems highly inappropriate as a description of the behaviour of more primitive forms of agency. On the other hand, there seems nothing wrong with Popper's strategy. If science can be shown to be applying a certain procedure that yields inductive success without a uniformity assumption being made, Hume's circularity is eliminated. The problem is that falsification doesn't quite fit the bill. Is there a way to reconstruct the argument using some other vehicle?

Turning to the machine learning literature for ideas about what this vehicle could be, we find a promising candidate in the form of data compression. Identification of compression as a vehicle for induction has long been a key part of thinking on learning. Through the work of researchers such as Solomonoff (1964), Watanabe (1969), Wolff (1980), Rissanen (1978), Chater (2003) and many others, the idea has been developed into a major paradigm of the field. Given the general workability of compression as a vehicle for induction, can we strengthen Popper's proposal by replacing falsification with compression?

Unfortunately, this move still leaves a residue of descriptive implausibility. Taking inductive behaviour to entail data compression is more general than taking it to entail hypothesis elimination, but not much. Referencing the principle of Occam's Razor, we might argue that data compression is what scientists are really doing when they believe they're doing induction. But even if this proposal is accepted, the assumed complexity of information processing seems incompatible with observations about ways more primitive agencies behave. Ultimately, the proposal seems to break down. The idea that sunflowers produce inductive effects by means of data compression, for example, seems outlandish.

As the present paper argues, there is a version of this argument that can be made to work, however. Instead of taking the inductive vehicle to be compression, we can take it to be *representation optimization*. Deploying concepts of information theory (Shannon, 1948; Shannon and Weaver, 1949), inductive effects can be shown to arise when there is any increase in the informational efficiency with which representations predict seen data. Representation optimization can then be viewed as a principled, well-motivated but non-inductive procedure that yields inductive effects implicitly. As such, it is able to take the place of falsification in Popper's argument.

The workability of representation optimization as a vehicle for induction can be demonstrated in two ways: either directly, using illustrations, or indirectly, by showing the process to be an implicit compression method. The improvement in descriptive plausibility is more readily apparent. Agents are no longer envisaged to be engaging in complex forms of information processing. Rather, they are seen to gravitate towards more efficient representation of seen data. In then becomes possible to give an account of inductive behaviour that more successfully generalizes the activities of scientists with more primitive forms of agency.

The paper sets out the proposal in more detail. The following section sets out the information-theoretic model through which representation optimization is shown to produce inductive effects implicitly. Following that, there is a section presenting illustrative examples. The paper then concludes with a brief discussion of implications.

Efficient reconstruction of symbolic data

The proposal is that increasing the informational efficiency with which representations predict seen data produces inductive (and compressive) effects implicitly. In order to demonstrate the effect, some basic definitions are needed. In what follows, D will represent a particular set of symbolic data. D is assumed to comprise constructs whose constituents are symbols drawn from an alphabet of n elements. Letting $|D|$ denote the total number of symbols in D, we can obtain the total information content using Shannon's logarithmic measure. Given a symbol with n possible values expresses $\log n$

[2]Modern Bayesian approaches to inductive confirmation follow the practice of Machine Learning in use of closed-world assumptions (Earman, 1992; Horwich, 1982; Howson and Urbach, 1989).

bits of information,[3] the total content of D is then

$$I(D) = |D| \log n \tag{1}$$

It is assumed that constituent symbols in constructs of D can be indexed, and that where two or more constructs have the same structure, the combination of those constructs can be referenced explicitly. Such combinations are named *unions*. If x represents a union, $|x|$ denotes the number of symbols it utilizes, and x_i is the set representing the choice of symbols for the i'th element of the (common) structure.

D' is then used to denote a *reconstruction* of D. This is defined to be a modification of D, in which some constructs are replaced with symbols representing unions. Replacement is deemed feasible just in case the construct is *within* the represented union. Where replacements introduce choice (multiple symbols for the same constituent) there is a well-defined loss of information. The loss resulting from a replacement by union x is

$$H(x) = \sum_i \log |x_i| \tag{2}$$

Equivalently, the loss may be defined as the log of the combinatorial product of x's choices:

$$H(x) = \log \prod_i |x_i| \tag{3}$$

The total information lost in a reconstruction is thus the sum of information losses of its constituent symbols:

$$H(D') = \sum_i H(D'_i) \tag{4}$$

Here, $H(D'_i)$ is zero if D'_i is an original symbol, and the information loss of the represented union otherwise.

Where replacement of constructs has the effect of reducing the total number of symbols in use, the symbol cost of the reconstruction must be less than $|D|$. The actual value can be calculated as the number of symbols used in the reconstruction itself, added to the total number of symbols used in referenced constructs. This cost is denoted $c(D')$:

$$c(D') = |D'| + \sum_{x \in D'} |x| \tag{5}$$

Here, $x \in D'$ enumerates the set of unions referenced by D'.

Combining the reconstruction loss with the reconstruction cost, it is then possible to define the informational *efficiency* of a reconstruction. This is the mean information content of symbols, i.e., the net information content divided by symbol usage:

$$\bar{I}(D') = \frac{I(D) - H(D')}{c(D')} \tag{6}$$

The informationally *optimal* reconstruction of D is then that reconstruction that maximizes mean information. Termed the *first refinement* of D,[4] this is denoted $r(D)$:

$$r(D) = \operatorname{argmax} \bar{I}(D') \tag{7}$$

A constraint on this is that the mean information of $r(D)$ can be no less than that of D itself. Were this to be the case, D would be its *own* optimal reconstruction by definition. Given $r(D) \neq D$, it must be the case that

$$\bar{I}(r(D)) > \bar{I}(D)$$

which further dictates that

$$c(r(D)) < |D|. \tag{8}$$

Increasing the mean content of symbols above the level they have in D itself must involve reducing their number. Any reconstruction of D must therefore use a lesser number of symbols than D itself. Forming a more efficient reconstruction of D thus necessarily produces the effect of compressing D, as we would expect.

Inductive properties of reconstructions

Within the analysis, data D and all its reconstructions are ways of predicting D under different levels of information loss. All represent the same content. Deriving a reconstruction of D that has higher efficiency than D itself, is thus the act of *increasing* informational efficiency in representation of D's content. By Eq. 8, this must have the effect of reducing the number of symbols in use. This can only be achieved through replacement of two or more constructs with a union. The informational cost of this replacement then depends on the degree to which the replaced constructs *differ* in their constituent symbols. The greater the similarity between constituent symbols in replaced constructs, the lower the information cost, and the greater the efficiency of the resulting representation.

Unions can thus be viewed as implicit generalizations, whose informational value increases with the constituent similarity of the constructs they replace. The informational efficiency of a representation is increased through the introduction of what are, in effect, 'similarity exploiting' generalizations. Putting it another way, increasing the efficiency of a representation has the effect of identifying (more) effective ways of exploiting commonalities.

[3] Logarithms are taken to base 2 in all cases.

[4] Not covered in this paper is an extension of the model to deal with recursive enhancement, a regime that typically produces a series of refinements.

We then begin to see how increasing informational efficiency produces implicit inductive effects. Any reconstruction embodies a certain number of unions, and every union specifies a choice of symbols for each of its constituents. The reconstruction represents the content of the original D with a certain loss of information. At the same level of loss, however, it represents the content of any *other* body of data whose constructs conform to the embodied symbol choices. This can be formulated as a rule:

$$D'_i \models D_j \text{ if } \exists\ D'_j : D'_j = D'_i \tag{9}$$

This asserts that reconstruction D'_i generalizes data D_j just in case there is a reconstruction of D_j which is identical to a reconstruction of D_i. The predictive properties of a reconstruction may be formalized in the same way. D' generalizes and thus (implicitly) predicts all bodies of data in the set

$$\{\ d \mid D' \models d\ \} \tag{10}$$

Increasing the efficiency of a representation generates reconstructions that implicitly predict unseen bodies of data. Such predictions will be valid just to the extent that unseen data exhibit constructs that are similarly structured. Representation enhancement thus implicitly predicts unseen data that exhibit structural uniformity with D.

Illustrative example

The scenario long favoured for examining induction is the case of 'white swans'. In this example, observations of white swans lead to the conclusion that 'all swans are white'. To examine the way in which this conclusion might arise implicitly from representation enhancement, we envisage data in the form of attribute vectors:

large	white	flying	swan
large	white	swimming	swan
small	white	flying	swan
medium	white	swimming	swan
small	white	swimming	swan

Each vector is placed on a separate line here. In left-to-right order, the attributes are size, color, behavior and type. The attribute of color is always 'white', while other attributes vary, reflecting the observed regularity that all observed swans are white. Taking each attribute to have four possible values, the information content of each original symbol is 2.0 bits. This yields a total information content of 40.0 bits for the data.

Taking constructs to be complete attribute vectors, any union must combine two or more vectors. Constituents of unions are thus choices of symbols for the four attributes. On this basis, we might build a reconstruction as follows.

-2.0	\$0 = small/large white flying/swimming swan
-2.0	\$0
-2.0	\$0
	medium white swimming swan
-2.0	\$0
-8.0	(40.0-8.0)/12 = 2.67 bits per symbol

The reconstruction is set out schematically. The first five lines represent the reconstruction itself, with the vertical ordering corresponding to the listing of the data. Where a replacement is made, we see the relevant information loss on the left, followed by the symbol generated (\$0 being the only one used here) to represent the construct. In the final line of the listing, we see the calculation of mean symbol information. Deducting the aggregate information loss of 8.0 bits from the original content of 40.0 bits, and dividing by the 12 symbols in use, we obtain a mean of 2.67 bits. The reconstruction increases informational efficiency (mean symbol content) by 0.67 bits.

A still more efficient reconstruction can be obtained, however. Consider

-2.58	\$1 = medium/large/small white flying/swimming swan
-2.58	\$1
-2.58	\$1
-2.58	\$1
-2.58	\$1
-12.9	(40.0-12.92)/9 = 3.01 bits per symbol

This yields a mean of 3.01 bits, with the introduced symbol (\$1 here) representing the union 'medium/large/small white flying/swimming swan'.

In both reconstructions, we see effects of implicit induction. The content of the original data is represented in terms of unions that create choice about (i.e., generalize) properties of size and behavior. The more efficient of the two embodies the expected generalization 'all swans are white'. This is predictive in the sense of predicting the observation 'medium white flying swan', a case not contained in the original set.

Illustrating the same effect in a slightly more complex way is the example below. This takes the previous data and mixes in additional vectors representing observations of black ravens.

medium	black	flying	raven
large	white	flying	swan
small	white	flying	swan
medium	black	perching	raven
small	white	perching	swan
large	black	flying	raven
large	white	perching	swan
medium	white	perching	swan
small	black	flying	raven

An efficient reconstruction in this case is

-2.58	$2 = medium/small/large black perching/flying raven
-2.58	$3 = medium/large/small white perching/flying swan
-2.58	$3
-2.58	$2
-2.58	$3
-2.58	$2
-2.58	$3
-2.58	$3
-2.58	$2
-23.26	(72.0-23.26)/17 = 2.87 bits per symbol

Here, the original content comes to be represented in terms of two unions, one of which captures the general properties of white swans, and the other of which captures the general properties of black ravens. Implicitly, the result embodies two inductive generalizations: 'all swans are white' and 'all ravens are black'. This generalizes the vectors 'medium white flying raven' and 'small black perching raven', neither of which are in the original data. Again, we see the effect of *implicit* inductive prediction.

Discussion

While the difficulty of establishing a universal basis for induction has plagued Epistemology for centuries, it is now also a source of counter-intuitive conclusions regarding induction, broadly defined. Referencing the NFL results particularly, we have to assume context-free inductive behaviour necessarily commences with a random selection of a context-sensitive bias. Referencing Hume's problem of circularity, we have to assume that a universal principle — were it to exist — could not be given any coherent definition.

Popper gets around Hume's problem of circularity by assuming induction is achieved through a non-inductive vehicle. But the implausibility of the vehicle as a general description of inductive behaviour poses a difficulty. In the present proposal, this is resolved by replacing falsification with representation optimization. As an implicit form of data compression, this has reasonable inductive credentials. As a vehicle for implicit induction, it more easily accommodates the full range of processes and behaviours we recognize to be involved. A non-circular account is then forthcoming for a context-free inductive methodology that has the potential to accommodate all levels of functionality, including the behaviours of scientists and sunflowers.

The original Popperian version of the explanation refers to scientists, and our observations of how they seem to use induction to predict future (or otherwise unseen) data. Popper resolves the worry that this process can have no non-circular explanation by arguing it is not really happening. Scientists are really engaged in systematic elimination of refuted hypothesis. The inductive element is pure interpretation. Induction is explained by showing it to be an interpretation of a process that is principled, advantageous but strictly non-inductive.

In the proposed revision, the scope of the explanation is broadened but the final effect remains the same. We now focus on inductive functionality broadly defined, taking into account the possibility of it being expressed in different forms of agency, and at multiple levels of organization. The difficulty of explaining how such processes can predict unseen data is again resolved by showing that prediction is not really what is happening. Such processes are adopting efficient representations of seen data, no more. The inductive effects that loom large in our interpretation are then recognized to be implicit. Provided the world exhibits uniformity they will be predictively successful, however, giving the impression of effective induction.

The strategy of explaining induction by 'not mentioning it' thus yields a reasonable accommodation of the philosophical problem of induction. Rather than being a flaw in our understanding, it begins to seem more of an artefactual difficulty, resulting from imposition of anthropocentric interpretations. The salience that the notion of *prediction* has within human concerns may be the origin of a tendency to frame interpretations of adaptive functionality around this particular concept. Adopting a more neutral position then has the effect of addressing some of the difficulties arising.

One other feature of the proposal is worth highlighting. It is an implication of the framework that induction involves compression by definition. The process is then understood to be feasible only with data which can be compressed. On the Kolmogorov criterion (Li and Vitányi, 1997), purely random data cannot be compressed, meaning such cases are implicitly ruled out. The proposed method is definitionally incapable of addressing them. The NFL objection, which relates to average performance in all possible scenarios, is then avoided. Induction under the present proposal is understood to be universal, but not in the sense of accommodating random data.

References

Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd Edition). Cambridge, Massachusetts: MIT Press.

Bishop, C. (2007). *Pattern Recognition and Machine Learning*. Springer.

Chater, N. and Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science. *Trends in Cognitive Sciences*, *7* (pp. 19-22).

Churchland, P. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.

Duhem, P. (1954/1914). *The Aim and Structure of Physical Theory* (Original work published 1914). Princeton, NJ: Princeton University Press.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Hempel, C. (1945). Studies in the logic of confirmation (i.). *Mind*, *54* (pp. 1-26).

Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge University Press.

Howson, C. and Urbach, P. (1989). *Scientific Reasoning: The Bayesian Approach*. Chicago, IL, US: Open Court Publishing Co.

Hume, D. (1988/1748). *An Enquiry concerning Human Understanding*. La Salle, Illinois: Open Court.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos and A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91-196). Cambridge, England: Cambridge University Press.

Langley, P. (1996). *Elements of Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.

Li, M. and Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications: Second Edition*. New York: Springer-Verlag.

Mackay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Michie, D., Speigelhalter, D. and Taylor, C. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Popper, K. (1979). *Objective Knowledge: An Evolutionary Approach* (Revised Edition). Oxford: Clarendon Press.

Putnam, H. (1974). The "corroboration" of theories. In P.A. Schilpp (Ed.), *The philosophy of Karl Popper* (Vol. I) (pp. 221-240). LaSalle, IL: Open Court.

Quine, W. (1953). Two dogmas of empiricism. In W.V.O. Quine (Ed.), *From a Logical Point of View* (pp. 20-46). Cambridge, MA: Harvard University Press.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, *14* (pp. 465-471).

Russell, B. (1946). *History of Western Philosophy*. London: George Allen & Unwin.

Schaffer, C. (1994). Conservation law for generalization performance. *Proceedings of the International Conference on Machine Learning* (pp. 259-265). July 10th-13th, Rutgers University, New Brunswick, New Jersey.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27* (pp. 379-423 and 623-656).

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.

Sloman, S. and Lagnado, D. (2005). In K.J. Holyoak and R.G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 95-116). Cambridge, UK: Cambridge University Press.

Smith, C. (2000). *Biology of Sensory Systems*. New York: John Wiley & Sons, Ltd.

Solomonoff, R. (1964). A formal theory of inductive inference, part i. *Information and Control*, *7*, No. 1 (pp. 1-22).

Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York: Wiley.

Wolff, J. (1980). Data compression, generalisation and overgeneralisation in an evolving theory of language development. *Proceedings of the AISB-80 conference on Artificial Intelligence*. Amsterdam.

Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, No. 7 (pp. 1341-1390).

Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computing*, *1* (pp. 67-82).