Is there any Need to Mention Induction?

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

February 9, 2011

General plan

General plan

- Introduce the problem of induction

General plan

- Introduce the problem of induction
- Mention poached eggs and black ravens

General plan

- Introduce the problem of induction
- Mention poached eggs and black ravens
- Talk about Popper's falsificationism

General plan

- Introduce the problem of induction
- Mention poached eggs and black ravens
- Talk about Popper's falsificationism
- Pros and cons of the idea

General plan

- Introduce the problem of induction
- Mention poached eggs and black ravens
- Talk about Popper's falsificationism
- Pros and cons of the idea
- Information-theoretic revamp

General plan

- Introduce the problem of induction
- Mention poached eggs and black ravens
- Talk about Popper's falsificationism
- Pros and cons of the idea
- Information-theoretic revamp
- Mention Grue

General plan

- Introduce the problem of induction
- Mention poached eggs and black ravens
- Talk about Popper's falsificationism
- Pros and cons of the idea
- Information-theoretic revamp
- Mention Grue

What's the problem?

We have the impression that we *learn* things.

We feel the things we learn constitute knowledge (e.g., science)

We'd like to know how new things are learned.

We'd like to know the sense in which they constitute (scientific) knowledge.

But rather weirdly there's no easy solution........

What's the problem?

We have the impression that we *learn* things.

We feel the things we learn constitute knowledge (e.g., science)

We'd like to know how new things are learned.

We'd like to know the sense in which they constitute (scientific) knowledge.

But rather weirdly there's no easy solution........

- *Circularity problem*: theories about how we learn things seem inevitably to apply some assumption based on previously-learned knowledge.

What's the problem?

We have the impression that we *learn* things.

We feel the things we learn constitute knowledge (e.g., science)

We'd like to know how new things are learned.

We'd like to know the sense in which they constitute (scientific) knowledge.

But rather weirdly there's no easy solution........

- *Circularity problem*: theories about how we learn things seem inevitably to apply some assumption based on previously-learned knowledge.
- *No-free-lunch problem*: theories about how we learn things that don't rely on previously-learned knowledge seem not to work — they produce 'knowledge' that's no more use than random guessing.

What's the problem?

We have the impression that we *learn* things.

We feel the things we learn constitute knowledge (e.g., science)

We'd like to know how new things are learned.

We'd like to know the sense in which they constitute (scientific) knowledge.

But rather weirdly there's no easy solution........

- *Circularity problem*: theories about how we learn things seem inevitably to apply some assumption based on previously-learned knowledge.
- *No-free-lunch problem*: theories about how we learn things that don't rely on previously-learned knowledge seem not to work — they produce 'knowledge' that's no more use than random guessing.

All swans are white?

The lack of a solution becomes apparent every time an inductively derived hypothesis turns out to be wrong.

Many hundreds of observations of white swans seemed to support the hypothesis ‘all swans are white’.

Then, Capt. Parker observed black swans in the New World....

Hume on the circularity problem

Descriptions of inductive principles end up ‘going in a circle, and taking that for granted, which is the very point in question’ (Hume, Enquiry, Salle Court, 1748, p. 80).

Wolpert on the NFL problem

Wolpert on the NFL problem

- No-free-lunch theorems: (Wolpert, Existence, 1996; Wolpert, Lack, 1996)

Wolpert on the NFL problem

- No-free-lunch theorems: (Wolpert, Existence, 1996; Wolpert, Lack, 1996)

See also

Wolpert on the NFL problem

- No-free-lunch theorems: (Wolpert, Existence, 1996; Wolpert, Lack, 1996)

See also

- Conservation law of generalization: (Schaffer, Conservation, 1994)

Wolpert on the NFL problem

- No-free-lunch theorems: (Wolpert, Existence, 1996; Wolpert, Lack, 1996)

See also

- Conservation law of generalization: (Schaffer, Conservation, 1994)
- Universal induction is a 'contradiction in terms' (Mitchell, 1997)

Wolpert on the NFL problem

- No-free-lunch theorems: (Wolpert, Existence, 1996; Wolpert, Lack, 1996)

See also

- Conservation law of generalization: (Schaffer, Conservation, 1994)
- Universal induction is a 'contradiction in terms' (Mitchell, 1997)

Facing up to the implications

The lack of a solution means

Facing up to the implications

The lack of a solution means

- We can't say we really *know* anything.

Facing up to the implications

The lack of a solution means

- We can't say we really *know* anything.
- Science is just a bunch of hunches.

Facing up to the implications

The lack of a solution means

- We can't say we really *know* anything.
- Science is just a bunch of hunches.
- Cognitive science is doomed.

Facing up to the implications

The lack of a solution means

- We can't say we really *know* anything.
- Science is just a bunch of hunches.
- Cognitive science is doomed.

Russell's poached egg

For Russell, 'there is [then] no intellectual difference between sanity and insanity'

Scientists are on an equal footing with 'the lunatic who believes that he is a poached egg.' (Russell, 1946, p. 673)

Keep calm and carry on?

Favourite strategies:

Keep calm and carry on?

Favourite strategies:

- Assume nature follows some general 'law of uniformity'. Unseen data then resemble seen data. Similarity-based induction justified.

Keep calm and carry on?

Favourite strategies:

- Assume nature follows some general 'law of uniformity'. Unseen data then resemble seen data. Similarity-based induction justified.
- Introduce a closed-world assumption. This reduces the task of induction to the task of sampling.

Keep calm and carry on?

Favourite strategies:

- Assume nature follows some general ‘law of uniformity’. Unseen data then resemble seen data. Similarity-based induction justified.
- Introduce a closed-world assumption. This reduces the task of induction to the task of sampling.

New formalisms often seem to offer a way out.

Contemporary epistemologists focus on ways to use Bayesian estimation.

Early 20th C (pre-computer era) epistemologists focused on ways to use FOL as a way of building up confirmatory evidence.

This does not solve either of the main problems, and also throws up the ‘black ravens’ paradox.

All ravens are black?

If we view induction as the process of accumulating confirmation through FOL-based processes, we run into the problem that negations seem to count as confirming evidence.

So observations of non-black non-ravens seem to be evidence in favour of 'all ravens are black'.

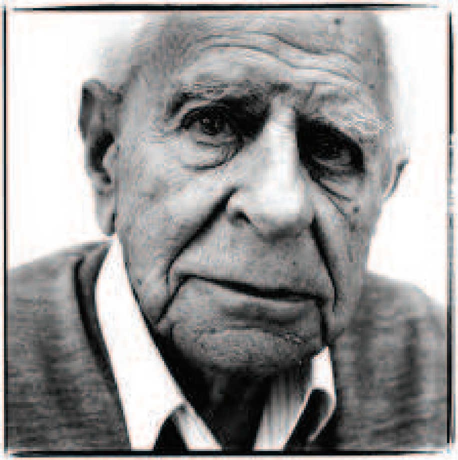
Observations of pink cows etc. become confirmatory evidence in favour of the hypothesis that all ravens are black.

The Machine Learning solution

Machine Learning plays it safe by introducing a closed-world assumption.
rThis is the IID criterion.

Unseen data assumed to be 'identically and independently distributed' with seen data.

Popper



Popper's approach

Popper's approach is completely different.

He says, there is 'no need even to *mention* induction' (Popper, 1959, p. 315).

Induction can arise indirectly out of a process that proceeds according to a *deductive* principle.

Less plausibly, Popper proposes that process to be uninformed exploration of hypotheses

Hypotheses found to be false are eliminated, producing some kind of progression in the inductive direction.

Does it work?

Popper's falsification must go in the right general direction.

But is there much chance of it getting anywhere?

Most see the procedure as implausibly 'blind', both descriptively and prescriptively.

The process does not reflect practices of induction (Kuhn, 1962; Lakatos, 1970)

Its unworkable where the number of hypotheses is large (Hempel, 1945; Churchland, 1986; Duhem, 1914/1954 Putnam, 1974; Quine, 1953)

Get rid of the bathwater but keep the baby

The falsification procedure may be implausible.

But the solution strategy is very interesting.

If a process proceeding non-inductively principle can produce inductive effects *indirectly*, the problem of circularity goes away.

There is then some hope of getting inductively-derived knowledge onto a principled footing.

Bring in information theory

Rework Popper's solution using information theory (Shannon, 1948; Shannon and Weaver 1949).

Show that inductive effects emerge implicitly when the informational *efficiency* of a representation is increased.

Induction can then be viewed as the indirect consequence of enhancing representation of *seen* data.

This produces the same effect as falsification, but without use of 'blind' search.

The methodology is informed by the general principles of information theory.

Formalization

D represents a particular set of symbolic data.

No constraints other than that D contains constructs whose constituents are symbols drawn from an alphabet of n elements.

Letting $|D|$ denote the total number of symbols in D, the total information content of D is then

$$I(D) = |D| \log n$$

Assume constituent symbols in constructs of D can be indexed

Where two or more constructs have the same structure, the *combination* of those constructs can be referenced explicitly.

These combinations are named *unions*.

Notation for unions

If x represents a union,

Notation for unions

If x represents a union,

- $|x|$ denotes the number of symbols it utilizes,

Notation for unions

If x represents a union,

- $|x|$ denotes the number of symbols it utilizes,
- x_i is the set representing the choice of symbols for the i'th element of the (common) structure.

Notation for unions

If x represents a union,

- $|x|$ denotes the number of symbols it utilizes,
- x_i is the set representing the choice of symbols for the i'th element of the (common) structure.

D' then denotes a *reconstruction* of D.

This is a modification of D, in which some constructs are replaced with symbols representing unions.

Replacement is feasible if the construct is *within* the represented union.

Information loss

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

$$H(x) = \sum_i \log |x_i|$$

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

$$H(D') = \sum_i H(D'_i)$$

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

Symbol cost

Where replacement of constructs reduces the number of symbols in use, the symbol cost of a reconstruction is less than $|D|$.

It is the number of symbols used in the reconstruction itself plus the number used in referenced constructs.

This is

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

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

Informational efficiency (mean symbol inf)

Combining reconstruction loss with the reconstruction cost, we can define the informational *efficiency* of a reconstruction

This is the net information content divided by the symbol usage:

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

The informationally *optimal* reconstruction of D is then that reconstruction that maximizes mean information.

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

Inductive effects

Increasing efficiency reduces the number of symbols (because it must involve replacing two or more constructs with a union).

Inductive effects

Increasing efficiency reduces the number of symbols (because it must involve replacing two or more constructs with a union).

- The informational cost of this depends on the degree to which the replaced constructs *differ* in their constituent symbols.

Inductive effects

Increasing efficiency reduces the number of symbols (because it must involve replacing two or more constructs with a union).

- The informational cost of this 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.

Inductive effects

Increasing efficiency reduces the number of symbols (because it must involve replacing two or more constructs with a union).

- The informational cost of this 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.
- More efficiency means more use of unions capturing more similar constructs.

Inductive effects

Increasing efficiency reduces the number of symbols (because it must involve replacing two or more constructs with a union).

- The informational cost of this 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.
- More efficiency means more use of unions capturing more similar constructs.

Increasing the efficiency of a representation thus implicitly produces generalizations, and these are based on patterns of similarity.

Generalizations may be genuinely predictive.

White swans example

Let the data be

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

with each value giving 2.0 bits of inf (i.e., four choices).

A reasonable reconstruction

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

Better still...

But we can achieve 3.01 bits per symbol with

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

$1 implicitly predicts 'medium white flying swan'

White swans with 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

Reconstruction

```
-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
 ---
(72.0-23.26)/17 = 2.87 bits per symbol
```

The optimal reconstruction here implicitly identifies two categories.

Is this a solution?

To solve the problem of induction, you need a general principle for predicting unseen (from seen) data that makes no assumptions about a relationship between the two.

To most people, this seems obviously *impossible*.

Popper says never mind, because the problem doesn't really exist.

Induction is just the interpretation we put on the process of hypothesis falsification.

The problem viewed as an artefact

In the inf. theory revamp, induction becomes the interpretation we put on the process of representation optimization.

This has the advantage of making the process an informed rather than a blind search.

If this process can reproduce all behaviours that we see as exhibiting prediction, then the problem of induction is *eliminated*, as Popper proposes.

The problem is understood to be just an artefact of our anthropocentric conceptualization.

What about the NFL problem?

The NFL result says that any general inductive principle must average performance at the level of random guessing when tested *exhaustively*.

This seems to rule out the possibility of completely general inductive principles, including representation optimization.

But representation optimization isn't claimed to produce above average performance *exhaustively*.

We cannot produce more efficient representation for completely random data.

No inductive effects are expected in such cases.

What about Grue?

Say we keep observing emeralds are green. The expected response would be to induce the hypothesis 'all emeralds are green'.

What if someone comes up with a crazy colour term, that accommodates these observations?

Will it be as good to induce 'all emeralds are <crazy colour term>'?

Goodman proposed 'grue', defining it to be true of an object if it is green and examined before some time t, or blue and examined after t.

Is 'all emeralds are grue' as good as 'all emeralds are green'?

If not, why not? This is Goodman's 'new riddle of induction'.

The Popperian elimination doesn't address this, even when revamped.

References

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.

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

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.

Mitchell, T. (1997). MACHINE LEARNING. McGraw-Hill.

Popper, K. (1959). THE LOGIC OF SCIENTIFIC DISCOVERY. London: Hutchinson.

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.

More refs

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.

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.

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

Wolpert, D. (1996). The existence of a priori distinctions between learning algorithms. NEURAL COMPUTATION, 8, No. 7.