Meaning as Inverse Interpretation

Linas Vepstas

24 February 2019

Abstract

The extraction of meaning from observed reality can be thought of as an inversion of the process of assigning meaning to axiomatic mathematical systems of formal mathematical logic. This essay formulates the analogy and explores it a little.

Introduction

In mathematical logic,[3] a collection of axioms, and the language that they form, can be studied independently of the interpretation of their meaning. This statement can be made precise: one defines what a "language" is, and separately, one provides a map from that language to a set. That map is called the "interpretation"; when it obeys a certain set of requirements regarding consistency, it can be said to provide the "meaning" of the axioms.

In exploring the real world - in the activity of performing science, and of discovery, one is faced with the inverse task: one has a collection of evidence, and from this evidence, one attempts to tease out a minimal explanation of the world, as a collection of rules, axioms and theories.

This essay provides a short definition of the formal mathematical concept of an "interpretation", illustrates why it provides "meaning", and then examines how one can find "meaning" in the evidence obtained from real life. Effectively, the discovery of meaning is the inversion of the process of assigning it: given some evidence collected from real life, we may not know the meaning, but we can seek to find it by discovering the underlying structure, and coupling the two with an interpretation.

This essay could devolve into a philosophical tract; rather, it attempts to be just concrete enough, just grounded and practical enough that one can imagine creating software that actually performs the described process.

Meaning

This subsection gives a quasi-formal definition of an interpretation. It is quite brief, providing just enough detail to get the general idea across, but playing fast and loose with important theorems. The reader is urged to consult standard texts for a full and complete definition.[3, 4, 1, 2]

One defines the concept of "terms" and "relations". A "term" is either some constant c or some variable x or some (uninterpreted) expression $f(t_1, \cdots, t_n)$ where the t_k are also terms. In some given, fixed system, there may be zero, one or more constants, and there may be zero or more function-expressions taking some number of arguments. A relation is an expression $R(t_1, \cdots, t_n)$ that can be either true, or false.

One standard example of terms and relations is arithmetic: it has two function-expressions (or operations): "plus", and "times", both taking two arguments, and two relations: "equals" and "greater-than", and two constants: "zero" and "one". Another standard example is set theory, which has no operations, one constant "the empty set", and two relations: "equals" and "is member of". Examples can be found everywhere in mathematics: groups, rings, fields, algebras and so on.

Formulas are obtained by combining relations with logical connectives: "and", "or", "not", "for all" and "there exists". This collection of terms, relations and logical connectives, and their various combinations forms a "language": a collection of symbols arranged in various ways with respect to one-another. But what is the "meaning" of that language? What do all of the symbols mean?

The canonical answer is that an interpretation provides "meaning" to the above formal symbolic expressions. It makes a language concrete by assigning real-world items to each of the symbols, each of the expressions obtained by combining these various elements. It connects the abstract to the real.

An interpretation is a function ϕ that maps constants, terms and relations to elements of some set M. For example, the set M could consist of an apple, pear and banana, and of all of the different ways that they might be arranged on a table-top. In this example, one might consider the axioms that define the permutation group S_3, which contains abstract expressions such as a^2b and e and ba and so on. What do a^2b and e and ba mean? Acting on the set of arrangements of an apple, pear and banana on a tabletop, it becomes clear: the element e does nothing at all, the element b swaps a pair of fruit, and the element a exchanges all three pieces of fruit. The equality relation expression as well: the world in which we applied only e comes out just the same as the world where we applied b^2. The "meaning" of the permutation group S_3 is the re-arrangement of three items.

Interpretation

A quick sketch of the more formal definition of an interpretation is in order. An interpretation of a constant c is some element $\phi(c)$ of the set M. (In real life, this "element" might be a single object, a movement or action, a color or attribute; nothing is presupposed). In casual speech, the constants c are the "names" of real-life things $\phi(c)$. We call something c whenever we see, in the real world, an object $\phi(c)$.

The interpretation of an n-ary term $f(t_1, \cdots, t_n)$ is a function $\phi(f) : M \times \cdots \times M \to M$, that is, it is some function defined on the Cartesian product of n copies of M, returning a value in M. Here, by "function" we mean the usual thing from school: a function that has a domain and a range: takes objects as inputs and generates an output. Its specific and concrete, whereas the uninterpreted term $f(t_1, \cdots, t_n)$ had no such constraints: it wasn't a function at all, just a (meaningless) arrangement of the the typographical symbols $f, t, 1, n, (,)$ on the written page. That the typography of it just

happens to resemble that of a function is already a powerful nod to the importance of notation and visual resemblance to understanding... and also to misunderstanding.

The interpretation of a relation $R(t_1, \cdots, t_n)$ is a subset $\phi(R)$ of $M \times \cdots \times M$. That is, the n-tuple $\langle \phi(t_1), \cdots, \phi(t_n) \rangle \in \phi(R)$ if and only if $R(t_1, \cdots, t_n)$ is true. This is just the usual correspondence between set-membership, and indicator functions. So a relation R is some intensive predicate; it's interpretation is some extensive set of objects. It is easiest to consider a unary predicate first, i.e. $n = 1$ so that $R(t)$ is either true or false for any given t. Given the set $\phi(R) = \{\text{cat}, \text{dog}, \text{mouse}\}$ we might conclude that $R(t)$ is true whenever $\phi(t)$ is a mammal. (or perhaps if $\phi(t)$ is furry...)

Terms can also be variables. The interpretation of variables is more abstract and difficult to understand; it is picked up again in a later section. It is, however, the central theme to this essay: the interpretation of variables is how one actually "learns".

Metaphysics

The above is a broader, and looser sketch of an "interpretation" than what is canonically found in books on mathematics. There's nothing particularly wrong with the formulas provided; rather, its the examples. Fruit and animals are never used as examples; whereas integers and sets are. The goal here is different: it is to illustrate that the formal, narrow mathematical concept of "interpretation" has a broader application than just mathematical logic.

But, I can almost hear you object: if the goal is to make analogies between real life and mathematics, there are many ways that this can be done. So, the interpretation ϕ behaves a lot like a homomorphism, why not talk about homomorphisms, instead? But also, ϕ behaves a lot like an action, so why not talk about actions? The answer is perhaps a let-down: the interpretation ϕ is a lot closer to what we really want, than a homomorphism or action is.

What do we really want? We want a correspondence between an internal mode of thought and perceived external experience, coupled in such a way that the internal system adequately reflects or describes the external world: predicting it, as it were, allowing the external world to be manipulated based on deductions performed on the internal system.

But why mathematical logic? Here the answer is stronger. Mathematical logic is a theory of symbols, and how they work together. Much of (but not all of) human endeavor involves a relationship with the symbolic world: for communication, for science, for deducing evidence in murder-mystery thrillers. The symbolic domain is part of human existence, and vitally important for science; its not going away. Mathematical logic has a well-articulated, deep and broad theory, with a giant treasure-chest of useful results that can be used (here, I am thinking broadly: proof theory, model theory, category theory and type theory all being sub-disciplines of mathematical logic). Many of these results are described as algorithms, with various degrees of efficiency: parsers and solvers coming in a rainbow of different forms.

Perhaps there is a better way of thinking about and talking about the correspondence between abstract understanding and real-world evidence. For right now, though, the concept of "interpretation" seems adequate. In particular, the previous section is

trying to set up the next section: a claim that one can invert this process, and discover interpretations, and meaning, algorithmically, automatically, by machine.

Science

One naive way of understanding science is that it is the process of obtaining formal, axiomatic systems that have interpretations in physical reality. To oversimplify: one observes how gravity makes things fall down, and then deduces $\vec{F} = m\vec{a}$. What is the meaning of $\vec{F} = m\vec{a}$? It means "things fall down". How do we know that is what it means? We know in two different ways. First, it has a very large collection of interpretations: one where m is an apple, another where m is a canon-ball, and yet another where its a planet. The second way is that $\vec{F} = m\vec{a}$ can be combined with the logical connectives "and", "or", "for all", "there exists", as well as other formulas an symbol manipulation techniques to obtain a very large collection of related formulas, each of which also has an interpretation that is consistent with logical reasoning applied to the formulas. That is, the "weight of evidence" is multi-faceted: not just that there is a large body of evidence, but that this evidence can be reverse-mapped to the theory, and to the symbolic articulations of the theory.

The claim of this essay is that the formal understanding of what an interpretation is, together with the informal explanation provided above, is just enough to be able to design and implement an actual, functional machine that extracts "meaning" from "observation". In short, "machine learning". But this first requires an examination of how variables are interpreted.

The Interpretation of Variables

In general, terms can be of the form $f(x,y,\cdots,z)$ containing variables $x,y,\cdots,z$ and likewise relations $R(x,y,\cdots,z)$. These two have to be given an interpretation: what do they mean? In colloquial terms, its not hard: one imagines that $x,y,\cdots,z$ stands for the set of all of those things that are appropriate for the term, or make the relation true. It takes some work to convert this informal understanding into a formal statement.

One interprets not a single variable at a time, but all of them, at once, assigning to each variable x an element M. The assignment is effectively "random". In itself, it has no particular meaning. One defines the space $\overline{M}$ as the space of all possible variable assignments. Thus, a single point $\xi \in \overline{M}$ corresponds to an assignment of elements in M for $x,y,\cdots,z$. In formulas, $\phi(x)(\xi) = \xi(x) \in M$, so that $\phi(x) : \overline{M} \to M$ is a function that maps a point in $\overline{M}$ to an element of M, and (by currying) $\phi : V \times \overline{M} \to M$ takes a variable, and some point in $\overline{M}$, and produces and element in the "real world" M. That is, the interpretation of a variable is a function ranging over variable assignments.

This is a bit painfully abstract, but it allows for a natural way of examining statements such as "if x is furry, then x is an animal". We can try to test if this statement is generically true by replacing x with an instance of every possible object in the real world. If a formula $p(x)$ remains true as $\xi \in \overline{M}$ is varied, i.e. if $\phi(p)(\phi(x)(\xi)) = \phi(p)(\xi(x))$ remains true for all $\xi \in \overline{M}$, then one says that $p(x)$ is a ϕ-true statement.

The goal of learning is to discover ϕ-true statements.[1]

Revising the example might make this clearer. Consider instead the proposition $p(x) =$"if x is c_f then x is a c_a". The symbols c_f and c_a are (uninterpreted) constants. What might they mean? Well, if $\phi(c_f) =$ "furry" and $\phi(c_a) =$ "animal" then we are on to something, because when we interpret $\phi(x)$ as a the set of all possible things that could be x, we can then examine if $p(x)$ is true (or not) (or rather, more precisely, if $\phi(p)(\phi(x))$ is true or not, as it varies over all of $\overline{M}$).

This now provides a criterion for determining whether an interpretation is real-world-reasonable, or not. So, $\phi(c_f) =$ "furry" and $\phi(c_a) =$ "animal" seems to work, but $\phi(c_f) =$ "slimey" and $\phi(c_a) =$ "wooden" does not.

Machine Learning

The point of this machinery is that it provides just enough specificity to imagine that it could be implemented as a machine learning algorithm. There is no actual machine learning algorithm presented; nor is there any claim that it might be efficient, fast or practical. You are welcome to tell me that doing the above will take longer than the age of the universe when implemented on an IBM-386 PC. Rather, it sketches what one must attempt to do, if one wishes to develop a symbolic understanding of reality.

Nothing presented her so far is in any way astounding, taken from the point of view of machine learning. A standard route in machine learning is to formulate some hypothesis $R(x,y,\cdots,z)$, and then crawl over a dataset of $x,y,\cdots,z$ to see if it holds. This is typically CPU-intensive, and most hypothesis do not hold. The holy grail of machine learning is to be able to generate accurate hypothesis from the get-go, and then to evaluate them quickly.

But what it does show is that it is possible, at least in principle, to formulate the hypothesis "if x is c_f then x is a c_a" together with an interpretation $\phi(c_f) =$ "furry" and $\phi(c_a) =$ "animal" and then run around, look at things, and see if it is true. If one can confirm this hypothesis, and a handful of others, then one is well on the way of developing an understanding of the real world, in such a way that one can navigate in it, without getting killed.

Conclusion

The point here, so far, is perhaps a more philosophical one, after all: a defense of the idea that learning is all about the extraction of patterns from the observation of nature. But there is more: the extraction can be done in such a way that a symbolic structure emerges. The symbols can be arranged into terms and relations, in a concrete enough that they can be compounded with logic connectives to obtain formulas, and

[1] In mathematics, when an interpretation ϕ of a collection of formulas $\mathscr{E}$ renders all of the formulas ϕ-true, then ϕ is called a "model" of $\mathscr{E}$. This is perversely backwards from the meaning of the word "model" in engineering and process control. In the latter case, one has the evidence $\phi(\mathscr{E})$ of real life, but doesn't really know what either ϕ or $\mathscr{E}$ are. When some ϕ and $\mathscr{E}$ are found that approximate reality, these are called a "model" of reality. Examples include a thermostat, which attempts to model the temperature inside a house. It cannot ever know the true temperature (drafty windows, cold basements), but it can form some reasonable model of what this might be, and then act to turn on the heat (or cooling). Likewise, the casual use of the word "interpretation" can mean exactly the opposite of the defintioin given here.

that those formulas can then be used to make predictions about reality. We've got tools and theories and systems that know how to manipulate symbols, and to do so reliably: again: the human race is deeply entangled with the processing of symbols. The point here is that symbols can be extracted from observations of nature; once they are, we are home-free.

References

[1] Franz Baader and Tobias Nipkow. *Term Rewriting and All That*. Cambridge University Press, 1998.

[2] Wilfrid Hodges. *A Shorter Model Theory*. Cambridge University Press, 1997.

[3] Yu. I. Manin. *A Course in Mathematical Logic*. Springer-Verlag, 1977.

[4] Dirk van Dalen. *Logic and Structure*. Springer-Verlag, fourth edition edition, 2004.