Journal of Mathematical Psychology 76 (2017) 13–24

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Identification of probabilities

Paul M.B. Vitányi [a,b], Nick Chater [c,*]

[a] National Research Institute for Mathematics and Computer Science, CWI, Science Park 123, 1098 XG, Amsterdam, The Netherlands
[b] University of Amsterdam, The Netherlands
[c] Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, CV4 7AL, UK

HIGHLIGHTS

- A fundamental problem of Bayesian inference is solvable in a number of contexts.
- Computability assumptions turn out crucially to simplify the learning problem.
- Exceptions can be learned from positive data, a long-standing puzzle in language acquisition.
- Data alone is often sufficient to learn an underlying model in perception.

ARTICLE INFO

Article history:
Received 1 August 2015
Received in revised form
13 October 2016

Keywords:
Learning
Bayesian brain, identification
Computable probability
Markov chain
Computable measure
Typicality
Strong law of large numbers
Martin-Löf randomness
Kolmogorov complexity

ABSTRACT

Within psychology, neuroscience and artificial intelligence, there has been increasing interest in the proposal that the brain builds probabilistic models of sensory and linguistic input: that is, to infer a probabilistic model from a sample. The practical problems of such inference are substantial: the brain has limited data and restricted computational resources. But there is a more fundamental question: is the problem of inferring a probabilistic model from a sample possible even in principle? We explore this question and find some surprisingly positive and general results. First, for a broad class of probability distributions characterized by computability restrictions, we specify a learning algorithm that will almost surely identify a probability distribution in the limit given a finite i.i.d. sample of sufficient but unknown length. This is similarly shown to hold for sequences generated by a broad class of Markov chains, subject to computability assumptions. The technical tool is the strong law of large numbers. Second, for a large class of dependent sequences, we specify an algorithm which identifies in the limit a computable measure for which the sequence is typical, in the sense of Martin-Löf (there may be more than one such measure). The technical tool is the theory of Kolmogorov complexity. We analyze the associated predictions in both cases. We also briefly consider special cases, including language learning, and wider theoretical implications for psychology.

1. Introduction

Bayesian models in psychology and neuroscience postulate that the brain learns a generative probabilistic model of a set of perceptual or linguistic data (Chater, Tenenbaum, & Yuille, 2006; Oaksford & Chater, 2007; Pouget, Beck, Ma, & Latham, 2013; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Learning is therefore often viewed as an inverse problem. Some aspect of the world is presumed to contain a probabilistic model, from which data is sampled; the brain receives a sample of such data, e.g., at its sensory surfaces, and has the task of inferring the probabilistic model. That is, the brain has to infer an underlying probability distribution, from a sample from that distribution.

This theoretical viewpoint is implicit in a wide range of Bayesian models in cognitive science, which capture experimental data across many domains, from perception, to categorization, language, motor control, and reasoning (e.g., Chater & Oaksford, 2008). It is, moreover, embodied in a wide range of computational models of unsupervised learning in machine learning, computational linguistics, computer vision (e.g., Ackley, Hinton, & Sejnowski, 1985; Manning & Klein, 2003; Yuille & Kersten, 2006). Finally, the view that the brain recovers probabilistic models from sensory data is both theoretically prevalent and has received considerable empirical support in neuroscience (Knill & Pouget, 2004).

* Corresponding author.
E-mail addresses: paulv@cwi.nl (P.M.B. Vitányi), Nick.Chater@wbs.ac.uk (N. Chater).

http://dx.doi.org/10.1016/j.jmp.2016.11.004

The idea that the brain may be able to recover a probabilistic process from a sample of data from that process is an attractive one. For example, a recovered probabilistic model might potentially be used to explain past input or to predict new input. Moreover, sampling new data from the recovered probabilistic model could be used in the generation of new data from that probabilistic process, for creating mental images (Shepard, 1984) or producing language (Chater & Vitányi, 2007). Thus, from a Bayesian standpoint, one should expect that the ability to perceive should go alongside the ability to create mental images; and the ability to understand language should go alongside the ability to produce language. Thus, the Bayesian approach is part of the broader psychological tradition of analysis-by-synthesis, for which there is considerable behavioral and neuroscientific evidence in perceptual and linguistic domains (Pickering & Garrod, 2013; Yuille & Kersten, 2006).

Yet, despite its many attractions, the proposal that the brain recovers probabilistic processes from samples of data faces both practical and theoretical challenges. The practical challenges include the fact that the available data may be limited (e.g., children learn the probabilistic model of highly complex language using only millions of words). Moreover, the brain faces severe computational constraints: even the limited amount of data encountered will be encoded imperfectly and may rapidly be lost (Christiansen & Chater, 2016; Haber, 1983). The brain has limited processing resources to search and test the vast space of possible probabilistic models that might generate the data available.

In this paper we explore the conditions under which exactly inferring a probabilistic process from a stream of data is possible even in principle, with no restrictions on computational resources like time or storage or availability of data. If it turns out that there is no algorithm that can learn a probabilistic structure from sensory or linguistic experience when no computational or data restrictions are imposed, then this negative result will still hold when more realistic settings are examined.

Our analysis differs from previous approaches to these issues by assuming that the probabilistic process to be inferred is, in a way that will be made precise later, computable. Roughly speaking, the assumption is that the data to be analyzed is generated by a process that can be modeled by a computer (e.g., a Turing machine or a conventional digital computer) combined with a source of randomness (for example, a fair coin that can generate a limitless stream of random 0s and 1s that could be fed into the computer). There are three reasons to suppose that this focus on computable processes is interesting and not overly restrictive. First, some influential theorists have argued that all physical processes are computable in this, or stricter, senses (e.g., Deutsch, 1985). Second, most cognitive scientists assume that the brain is restricted to computable processes, and hence can only *represent* computable processes (e.g., Rescorla, 2015). According to this assumption, if it turns out that some aspects of the physical world are uncomputable, these will trivially be unlearnable simply because they cannot be represented; and, conversely, all aspects of learning of relevance to psychology, i.e., all aspects of the world that the brain can successfully learn, will be within the scope of our analysis. Third, all existing models of learning in psychology, statistics and machine learning are computable (and, indeed, are actually implemented on digital computers) and fall within the scope of the present results.

1.1. Background: pessimism about learnability

Within philosophy of science, cognitive science, and formal learning theory, a variety of considerations appear to suggest that negative results are likely. For example, in the philosophy of science it is often observed that theory is underdetermined by data (Duhem, 1914–1954; Quine, 1951): that is, an infinite number of theories is compatible with any finite amount of data, however large. After all, these theories can all agree on any finite data set, but diverge concerning any of the infinitely large set of possible data that has yet to be encountered. This might appear to rule out identifying the correct theory—and hence, *a fortiori* identify a correct probability distribution.

Cognitive science inherits such considerations, to the extent that the learning problems faced by the brain are analogous to those of inferring scientific theories (e.g., Gopnik, Meltzoff, & Kuhl, 1999). But cognitive scientists have also amplified these concerns, particularly in the context of language acquisition. Consider, for example, the problem of acquiring language from positive evidence alone, i.e., from hearing sentences of the language, but with no feedback concerning whether the learner's own utterances are grammatical or not (so-called negative evidence). It is often assumed that this is, to a good approximation, the situation faced by the child. This is because some and perhaps all children receive little useful feedback on their own utterances and ignore such feedback even when it is given (Bowerman, 1988). Yet, even without negative evidence, children nonetheless learn their native language successfully. For example, an important textbook on language acquisition (Crain & Lillo-Martin, 1999) repeatedly emphasizes that the child cannot learn *restrictions* on grammatical rules from experience—and that these must therefore somehow arise from innate constraints. For example, the English sentences *which team do you want to beat*, *which team do you wanna beat*, and *which team do you want to win*, would seem naturally to imply that **which team do you wanna win* is also a grammatical sentence. As indicated by the asterisk, however, this sentence is typically rejected as ungrammatical by native speakers. According to classical linguistic theory (e.g., Chomsky, 1982), the contraction to *wanna* is not possible because it is blocked by a "gap" indicating a missing subject—a constraint that has sometimes been presumed to follow from an innate universal grammar (Chomsky, 1980).

The problem with learning purely from positive evidence is that an overgeneral hypothesis, which does not include such restrictions, will be consistent with new data; given that languages are shot through with exceptions and restrictions of all kinds, this appears to provide a powerful motivation for linguistic nativism (Chomsky, 1980). But this line of argument cannot be quite right, because many exceptions are entirely capricious and could not possibly follow from innate linguistic principles. For example, the grammatical acceptability of *I like singing*, *I like to sing*, and *I enjoy singing* would seem to imply, wrongly, the acceptability of **I enjoy to sing*. But the difference between the distributional behavior of the verbs *like* and *enjoy* cannot stem from any innate grammatical principles. The fact that children are able to learn restrictions of this type, and the fact that they are so ubiquitous throughout language, has even led some scholars to speak of the *logical* problem of language acquisition (Baker & McCarthy, 1981; Hornstein & Lightfoot, 1981).

Similarly, in learning the meaning of words, it is not clear how, without negative evidence, the child can successfully retreat for overgeneralization. If the child initially proposes that, for example, *dog* refers to any animal, or that *mummy* refers to any adult female, then further examples will not falsify this conjecture. In word learning and categorization, and in language acquisition, researchers have suggested that one potential justification for overturning an overgeneral hypothesis is that absence-of-evidence can sometimes be evidence-of-absence (Hahn & Oaksford, 2008; Hsu, Horng, Griffiths, & Chater, 2016). That is, a child might take the absence of people using the word *dog* when referring to cats or mice; and the absence of *Mummy* being used to refer to other female friends or family members might lead to the child to be in doubt concerning their liberal use of these terms. But, of course, this line