Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 40 (2016) 99–113

Cognitive Systems Research

www.elsevier.com/locate/cogsys

Neural implementation of probabilistic models of cognition

Action editor: Peter Erdi

Milad Kharratzadeh [a,*], Thomas Shultz [b]

[a] Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada
[b] Department of Psychology & School of Computer Science, McGill University, Montreal, Quebec, Canada

Received 26 October 2015; received in revised form 3 March 2016; accepted 12 April 2016
Available online 28 April 2016

Abstract

Bayesian models of cognition hypothesize that human brains make sense of data by representing probability distributions and applying Bayes' rule to find the best explanation for available data. Understanding the neural mechanisms underlying probabilistic models remains important because Bayesian models provide a computational framework, rather than specifying mechanistic processes. Here, we propose a deterministic neural-network model which estimates and represents probability distributions from observable events—a phenomenon related to the concept of probability matching. Our model learns to represent probabilities without receiving any representation of them from the external world, but rather by experiencing the occurrence patterns of individual events. Our neural implementation of probability matching is paired with a neural module applying Bayes' rule, forming a comprehensive neural scheme to simulate human Bayesian learning and inference. Our model also provides novel explanations of base-rate neglect, a notable deviation from Bayes.

Keywords: Neural networks; Probability matching; Bayesian models; Base-rate neglect

1. Introduction

Bayesian models are now prominent across a wide range of problems in cognitive science including inductive learning (Tenenbaum, Kemp, & Shafto, 2006), language acquisition (Chater & Manning, 2006), and vision (Yuille & Kersten, 2006). These models characterize a rational solution to problems in cognition and perception in which inferences about different hypotheses are made with limited data under uncertainty. In Bayesian models, beliefs are represented by probability distributions and are updated by Bayesian inference as additional data become available. For example, the baseline probability of having cancer is lower than that of having a cold or heartburn. Coughing is more likely caused by cancer or cold than by heartburn. Thus, the most probable diagnosis for coughing is a cold, because having a cold has a high probability both before and after the coughing is observed. Bayesian models of cognition state that humans make inferences in a similar fashion. More formally, these models hypothesize that humans make sense of data by representing probability distributions and applying Bayes' rule to find the best explanation for available data.

Forming internal representations of probabilities of different hypotheses (as a measure of belief) is one of the most important components of several explanatory frameworks. For example, in decision theory, many experiments show that participants select alternatives proportional to their frequency of occurrence. This means that in many scenarios, instead of maximizing their utility by always choosing

* Corresponding author.
E-mail addresses: milad.kharratzadeh@mail.mcgill.ca (M. Kharratzadeh), thomas.shultz@mcgill.ca (T. Shultz).

http://dx.doi.org/10.1016/j.cogsys.2016.04.002
1389-0417/

the alternative with the higher chance of reward, they match the underlying probabilities of different alternatives. For a review, see Vulkan (2000).

There are several challenges for Bayesian models of cognition as suggested by recent critiques (Bowers & Davis, 2012; Eberhardt & Danks, 2011; Jones & Love, 2011; Marcus & Davis, 2013). First, these models mainly operate at Marr's computational level (Marr, 1982), with no account of the mechanisms underlying behavior. That is, they are not concerned with how people actually learn and represent the underlying probabilities. Jones and Love (2011, p. 175) characterize this neglect of mechanism as "the most radical aspect of Bayesian Fundamentalism". Second, in current Bayesian models, it is typical for cognitive structures and hypotheses to be designed by researchers, and for Bayes' rule to select the best hypothesis or structure to explain the available evidence (Shultz, 2007). Such models often do not typically explain or provide insight into the origin of such hypotheses and structures. Bayesian models are under-constrained in the sense that they predict various outcomes depending on assumed priors and likelihoods (Bowers & Davis, 2012). Finally, it is shown that people can be rather poor Bayesians and deviate from the optimal Bayes' rule due to biases such as base-rate neglect, the representativeness heuristic, and confusion about the direction of conditional probabilities (Eberhardt & Danks, 2011; Eddy, 1982; Kahneman & Tversky, 1996; Marcus & Davis, 2013).

In this paper, we address some of these challenges by providing a psychologically plausible neural framework to explain probabilistic models of cognition at Marr's implementation level. As the main component of our framework, we study how deterministic neural networks can learn to represent probability distributions; these distributions can serve later as priors or likelihoods in a Bayesian framework. We consider deterministic networks because from a modeling perspective, it is important to see whether randomness and probabilistic representations can emerge as a property of a population of deterministic units rather than a built-in property of individual stochastic units. For our framework to be psychologically plausible it requires two important properties: (i) it needs to learn the underlying distributions from observable inputs (e.g., binary inputs indicating whether an event occurred or not) and (ii) it needs to adapt to the complexity of the distributions or changes in the probabilities. We discuss these aspects in more details later.

The question of how people perform Bayesian computations (including probability representations) can be answered at two levels (Marr, 1982). First, it can be explained at the level of psychological processes, showing that Bayesian computations can be carried out by modules similar to the ones used in other psychological process models (Kruschke, 2006). Second, probabilistic computations can also be treated at a neural level, explaining how these computations could be performed by a population of connected neurons (Ma, Beck, Latham, & Pouget, 2006). Our artificial neural network framework combines these two approaches. It provides a neurally–based model of probabilistic learning and inference that can be used to simulate and explain a variety of psychological phenomena.

We use this comprehensive modular neural implementation of Bayesian learning and inference to explain some of the well-known deviations from Bayes' rule, such as base-rate neglect, in a neurally plausible fashion. In sum, by providing a psychologically plausible implementation-level explanation of probabilistic models of cognition, we integrate some seemingly contradictory accounts within a unified framework.

The paper is organized as follows. First, we review necessary background material and introduce the problem's setup and notation. Then, we introduce our proposed framework for realizing probability matching with neural networks. Next, we present empirical results and discuss some relevant phenomena often observed in human and animal learning. Finally, we propose a neural implementation of Bayesian learning and inference, and show that base-rate neglect can be implemented by a weight-disruption mechanism.

2. Learning probability distributions via deterministic units

2.1. Problem setup

The first goal of this paper is to introduce networks that learn probability distributions from realistic inputs. We consider the general case of multivariate probability distributions defined over $q \geqslant 1$ different random variables, $\mathbf{X} = (X_1, X_2, \ldots, X_q)$. We represent the value of the density function by $p(\mathbf{X}|\Theta)$, where Θ represents the functional form and parameters of the distribution. We assume that Θ is unknown in advance and thus would need to be learned. As shown in Fig. 1, the neural network learning this multivariate distribution has q input units corresponding to $(X_1, X_2, \ldots, X_q)$ and one output unit corresponding to $p(\mathbf{X}|\Theta)$.

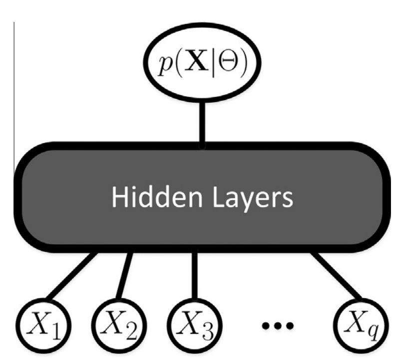

Fig. 1. The basic structure of the network learning a q-dimensional probability distributions. Both structural details and connection weights in the hidden layers are learned.