CMPE 540 TERM PAPER, FALL, 2015

BAYESIAN APPROACHES IN HUMAN COGNITION

January 17, 2016

Çağatay Yıldız
Boğaziçi University,
Computer Engineering Department

Contents

1 Introduction

For decades, researchers have been working on building human-like artificial intelligence systems. Reverse engineering human learning is one of the methods that help in this search and it turned out that many aspects of cognition depend heavily on the probabilistic representation of uncertainty. As a result, Bayesian approaches have become the mainstream approach in explaining our cognitive abilities.

In this review, we try to approach human cognition from a Bayesian standpoint. It seems plausible to first compare traditional machine learning methods with Bayesian methods. In Section 3, we discuss why applying probabilistic models in AI tasks and understanding human cognition is appropriate. We then give a brief history of Bayesian models of cognition and investigate how abstract knowledge can be represented in a Bayesian setting.

2 Two Perspectives in Machine Learning

For decades, researchers have been working on developing algorithms that learn from the data and make predictions accordingly. The umbrella name for these methods is “machine learning algorithms”. Thanks to the production of powerful computers with relatively reduced prices and the growing interest in the field as a result of its practical applications, a large number of machine learning algorithms have been developed in the last two decades. Logistic regression, support vector machines, decision trees and deep neural networks are just some of them.

Especially in the era of big data, neural networks are applied to a lot of challenging problems such as speech recognition, image classification, natural language processing and so forth. Despite all its successes, neural networks -and all of the traditional approaches- has a big limitation: lack of adaptability. There is usually not much to modify in these approaches, except for playing with parameters. Similarly, domain-specific knowledge is usually excluded, which may not be desired depending on the nature of the problem. For such problems, Bayesian view of machine learning provides an alternative: Bayesian setting allows the development of a broad range of models and problem-specific knowledge can easily be combined with advanced inference algorithms.

Furthermore, traditional approaches suffer from handling uncertainty, which is a key ingredient of many learning and intelligence tasks. There can be many sources of uncertainty: The data may be noisy, model parameters may be probabilistic and at the highest level, there might be uncertainty with the algorithm to choose ([1]). But in traditional machine learning methods, model parameters are point values and the goal is to find the best parameter settings by an optimization technique. Thus, there is no mechanism to represent uncertainty.

Bayesian machine learning uses probability theory to express all sorts of uncertainty in the model([2]). For each observation, current distribution over the set of variables are treated as *prior* knowledge, which is updated by the observation and *posterior* distribution is calculated. This posterior distribution becomes the prior for the next observation. Note that this procedure, or *inference*, replaces parameter optimization mentioned in previous techniques.

3 Why Bayesian Standpoint in Human Cognition is Appropriate

3.1 Belief Representation

One central question in artificial intelligence is how to represent the degree of belief of an AI system about the world. In a Bayesian setting, this can be done by considering probabilities as beliefs. Frequentist view, nevertheless, considers this as misinterpretation of probability theory: There is no connection between probabilities calculated and the real world, as they claim. Cox stated long before machine learning *starts* ([3]) that a system must be consistent with the deductive logic and in correspondence with commonsense reasoning if it is to be used for plausible reasoning. He also showed that such a system is equivalent of what probability theory claims([4]). Staying within the scope of this paper, we can simply conclude that Cox axioms justify Bayesian claim.

3.2 Simple Design and Interpretability

In probabilistic machine learning, one may come up with a large number of graphical models for a single problem. As mentioned above, this helps model to incorporate domain-specific knowledge. What's more, the graphical model can be tuned. For example, if an agent that can learn animal names is to be designed, a tree-like directed graphical model may easily work out. But if the task is to draw causal relations among a set of data, it may be better to have a graph structure that represents two types of variables and links among those.

In relation to the the point above, relatively smaller building blocks (typically made of a few variables) can come together to form more complex Bayesian models. The interpretation of such models then boils down to understanding what each sub-block represents and the relationship among those. This is certainly much easier than interpreting a coupled non-linear system. Take recurrent neural networks, for example. While what hidden units in a neural network learn is still unclear, combining weights in hidden unit at different time steps has no formal explanation, at least for today.

3.3 Better Fit and Parametrization

Another advantage of Bayesian approaches over others is that the former does not require a lot of data to make inference whereas parameter optimization based methods (and maximum likelihood-based methods, in general) tend to over-fit the data. In Bayesian settings, even if the number of free parameters exceeds the number of data points, there would be no problem because the *effective* number of parameters adapts automatically to the data set ([5]). We will see in the next sections that a number of human cognition tasks involve the ability to generalize from sparse data, and thus requires Bayesian perspective.

When designing a learning system, we would naturally like it to be as flexible as possible. One way of bringing flexibility is to use non-parametric methods. Many non-parametric techniques increase the complexity of predictions as more and more data become available, which is yet another feature of many AI problems. Being a non-parametric method, Chinese restaurant process explains how children discover words in unsegmented speech and learn morphological rules. Similarly, the Indian buffet process explains the construction of new perceptual features in object categorization ([6]).

3.4 Marr's Tri-Level Hypothesis

David Marr's distinction between three levels of analysis [7] also leads to the fact that probabilistic representations in cognitive science are plausible. Marr proposes that any information processing system should be understood in three distinct levels: the **computational level**, which defines the nature of the system and problems it solves; the **algorithmic level**, how the system does (whatever it does), the processes and representations it uses; and the **implementational level**, the physical realization of the system.

As mentioned before, cognition involves uncertainty and thus, the problem can be tackled by probabilistic approaches. At the algorithmic level, it has been suggested that cognitive processes are performed by a number of heuristics. However, in some cases, human cognition is considered to be much more general and flexible than what heuristics can provide and currently, this flexibility is best explained by probabilistic approaches (see Section 4 for many examples). Finally, findings in computational neuroscience indeed suggest that the brain represents sensory information probabilistically. In Knill and Pouget's work ([8]), it is showed that cue integration takes place in a Bayes-optimal fashion, no matter which cues are integrated (visual-auditory, sight-touch or sight-sound). They also note that *"behavioral tests have confirmed that motor plans take into account the uncertainty in motor outputs"*, which implies that brain takes uncertainty into account in the sensory input and the motor output.

4 A Brief History of Probabilistic Models of Cognition

In previous section, we have seen that probability theory has a wide range of practices in human cognition. However, not so much time has passed since probability was applied to cognitive sciences. In his inspiring article ([9]), Chater argues three possible reasons. First, scientists in this field were too busy debating "symbolic rule-based processing vs connectionist networks" approaches, which we will come later on. That, in turn, puts the discussion of how to implement a probabilistic framework to the background. What's more, the uncertainty was already represented (to some extent) by non-probabilistic techniques such as default logics, non-monotonic logics and various heuristic methods. Finally, the power of probabilistic methods has not been noticed and it was believed that such methods were quite restrictive when it comes to represent complex cognitive processes.

Initial attempts to build probabilistic models of cognition date back to Pearl's work on Bayesian networks ([10]). He proposed methods to combine the needs of a learning system with the probability theory, which eventually triggered developments in the artificial intelligence systems, machine learning, decision support systems and so on. In parallel with that, Kahneman and Tversky showed that human cognition may be not be rational, optimal and operate probabilistically. ([11]). This belief is rooted to the fact that people are usually terrible at calculations involving probabilities. But being more general, it was realized all sorts of mathematical analysis push people to their limits, it is not just probabilistic calculations. Also, probabilistic models turn out to be best applied to cognitive processes that are well-optimized, not those that depend on individual differences. (However, it is ironic that people's judgment in cases that involve chance are typically poorly explained by such models)

Vision is one the fields that is easily explicable from a rational, probabilistic view. Maloney and Zhang showed that ([12]) human performance in visual tasks are compatible with the ideal performance calculated using Bayesian decision theory. Markov random fields and Gibbs samplers also give explanation to the Gestalt laws of perceptual organization ([9]). In addition, Ernst and Banks made use of the fact that probabilistic models allow to couple different sensory inputs: They built a model that integrates visual and haptic information in a maximum-likelihood based model and this model turned out to behave very similarly in visual-haptic tasks as humans ([9]).

Causal learning is another field in which probabilistic modeling is heavily used. Tenenbaum and Griffiths applied Bayesian structure learning methods and Bayesian model selection to explain human judgments ([13]). Similarly, the influential paper by Gopnik ([14]) proposes that children form *causal maps* to represent the abstract and learned representations of the causal relations among events. Directed acyclic graphs are used to represent relations and in such a setting, the new causal maps formed by 2-to-4 year-old children are consistent with the Bayesian network formalism. More about this topic is discussed in the next section.

5 Bayesian Abstract Knowledge Representation

A wonderful characteristic of cognition is that humans are able to learn using very little amounts of data. For instance, children at age 2 can learn new words just by seeing very few examples. Because they are also able to use the newly acquired words, we deduce that they are capable of making generalizations and construct abstractions. But how does this abstract knowledge guide learning? What is the form of the knowledge and how is it acquired? ([6]) These three questions are the main focus of this section.

In literature, two approaches regarding the origins of knowledge have been commonly discussed: **nativism** and **associationism** ([6], [15]). In the first approach, on the acquisition of new data, learner develops a number of hypotheses, combines those with its well-structured knowledge base and then goes with the hypothesis that is consistent with the prior knowledge. Thus, learning boils down to eliminating incorrect hypotheses about whatever is learned and there is no role of probability or statistics in learning. In associationism, on the other hand, the input data and the output of learning mechanism are appropriately designed but not structured to form a knowledge base. Learning takes place by the application of advanced statistical methods and what's learned in this approach is just the weights in the layers of "hidden" units.

A relatively newer approach is based on the thesis that "children's learning mechanisms are analogous to scientific theory-formation" ([14]). Such a learning mechanism can be modeled by a causal Bayesian network, which allows abstract and structured representations given the data, as Gopnik *et al.* claims. Their experiments on 4 year-old children confirm that: Children use Bayesian analysis to combine prior probability with the conditional probability of the events. In the rest of this article, we elaborate the Bayesian perspective on the abstract knowledge representation that leads to learning and inference.

5.1 How Abstract Knowledge Guides Learning

As noted previously, Bayesian models are at their best when there are not much data or when the data are noisy, which is because they typically handle uncertainty very well. At the heart, such models are based on a very simple but powerful rule:

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_{h' \in H} p(D|h')p(h')} \propto p(D|h)p(h) \tag{1}$$

Here, D corresponds to the real world data, H is a set of hypotheses that may explain the data and h is the candidate hypothesis. Then, $p(h|D)$ represents how likely the hypothesis h, given the data. The abstract knowledge here is actually built-in this equation: The set of all hypotheses, their prior probabilities, the likelihood of each hypothesis producing given event and the equation itself form the knowledge.

A simple example of abstract knowledge encoding is given in concept learning problem ([6]): In an experiment setting, the task is to find the category that items belong to. For example, when children are given three examples of different types of horses, they usually come up with "horses"(h_1) as the hypothesis, not "all horses except "Clydesdales"(h_2) or "all animals"(h_3). Studies show that a Bayesian models over a tree-structured domain representation produces appropriate output. In this representation, all animals are contained in the leaves of a tree and branches are produced according to biological features, which at the end yields a more detailed and deep variant of animal taxonomy charts introduced in high-school biology courses (Note that selecting a category corresponds to picking a branch). If priors are defined as a function of branching factor, which reflects how distinctive the category is, and the likelihoods assume that examples are drawn randomly from the set of all animals under this category, which favors restricted categories, then the posterior probabilities turned out to be very close to the answers given by a human. Getting back to horse example, prior would favor h_1 and h_3 and likelihood would favor h_1 and h_2. Thus, h_1 is produced as the output.

5.2 The Form of Abstract Knowledge

In Bayesian approach to cognition, graphical models are used to represent the underlying procedure generating data. These models also correspond to the form of abstract knowledge. Tree-structured domain representation is, for example, just one of many models. Depending on the problem, one may prefer rings or chains over trees. In hierarchical Bayesian models, model itself learns the best structure, given the data.

Each different form of knowledge representation makes certain assumptions on the prior distribution of variables and thus, imposes different constraints. What makes a model better than others is the fitness of constraints to the data, not the size of the model. In concept learning problem, there are 2^n different subsets and each of them may be a hypothesis. But, this essentially puts no constraint on the data and such a grouping makes no sense, if one can call it "grouping". A better categorization schema was presented in the previous subsection, which contains just $n-1$ categories, if words are contained in leaves and each branch corresponds to a category ([6]).

5.3 The Origins of Abstract Knowledge

The last question we ask is how humans learn how to learn. Considering concept learning problem again, how does a child learn to use a tree structure for representation? Nativists shut down the discussion by stating that different kinds of cognitive models must be innate. For connectionists, such structures can be represented by the weights of a neural network; however, this can be at best a type of approximation whereas experiments show that people do know different representations explicitly.

Nowadays, researchers attempt to explain above phenomena by hierarchical Bayesian models(HBM). In HBM's, there are a number of levels of hidden variables. Each level, being the posterior distribution of the previous level, is the prior on the states in the next level([16]). Because such a model allows defining hypothesis spaces over hypothesis spaces, we expect the model to be quite general and strong enough to represent many types of problems.

Tenenbaum and Kemp indeed showed that ([6]) HBM's can discover the correct forms of structure. In a medical diagnosis problem with 6 diseases and 10 symptoms, they built a bipartite graph that successfully weights the causal links between symptoms and diseases. For the same problem, a two-level HBM, which is not explicitly modeled for medical diagnosis, is able to perfectly recover the true causal network by 1000 samples (patients). An additional level turns out to serve as a prior on different causal networks and thus constrains learning. Such a schema managed to form two classes of variables and prior favoring links from only one class to another, by just having 80 observations. What HBM does, getting the big picture first and weighting the links afterwards, is distinctively human mode of learning.

5.4 Open Questions

Although Bayesian models are powerful in explaining human cognition, we still do not know how everything gets started and to what extent our cognitive abilities are innate. Side questions related to human cognition also remain unanswered: How do humans understand false beliefs, what are the sources of individual differences and how do we constrain learning in such a flexible structure? Finally, the biggest question is how we can represent Bayesian models in neural circuits.

6 Conclusion

We have seen that Bayesian approaches allow us to build rich models by combining complex data representations with advanced statistical tools. Such models also handle uncertainty very well and therefore, do not need huge amounts of data for learning - just like human beings. These are the main reasons why many aspects of human cognition such as cue integration, causal learning, abstract knowledge formation and so on can be modeled from a Bayesian viewpoint.

References

[1] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[2] Christopher M Bishop. Model-based machine learning. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(1984):20120222, 2013.

[3] RT Cox. Probability, frequency and reasonable expectation. *Readings in uncertain reasoning*, pages 353–365, 1990.

[4] Kevin S Van Horn. Constructing a logic of plausible inference: a guide to coxâĂŹs theorem. *International Journal of Approximate Reasoning*, 34(1):3 – 24, 2003.

[5] Christopher M Bishop. *Pattern recognition and machine learning*, volume 1. springer, 2006.

[6] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[7] David Marr. *Vision*. W. H. Freeman and Company, 1982.

[8] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[9] Nick Chater, Joshua B Tenenbaum, and Alan Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, 10(7):287–291, 2006.

[10] Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. *Synthese-Dordrecht*, 104(1):161, 1995.

[11] Daniel Kahneman and Amos Tversky. Choices, values, and frames. *American psychologist*, 39(4):341, 1984.

[12] Laurence T. Maloney and Hang Zhang. Decision-theoretic models of visual perception and action. *Vision Research*, 50(23):2362 – 2374, 2010. Vision Research Reviews.

[13] Thomas L Griffiths and Joshua B Tenenbaum. Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384, 2005.

[14] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[15] Fei Xu and Joshua B Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

[16] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.