An Introduction To Bayesian Models of Cognitive Development

Rishabh Samra
ee17s300

Abstract

This study presents an introduction to Bayesian Inference as it is used in probabilistic models of cognitive development. The goal is to provide what sorts of problems and data the framework is most relevant for. It also introduces the framework of Hierarchical Bayesian Models where knowledge is represented at multiple levels of abstraction. Also, some important issues were discussed that often arise when evaluating Bayesian Models in cognitive science.

Keywords: Bayesian Models, Cognitive Development, Inductive Constraint

1 Introduction

One of the central questions of cognitive development is how we learn so much from apparently limited evidence and arrive at the generalized conclusion. Probabilistic models provide a computational framework for exploring how a learner might make these inductive leaps explaining them as forms of bayesian inference. This study considers three inductive problems a learner faces:

- Inductive generalization from examples, with a focus on learning the referents of words for object categories.
- Acquiring inductive constraints, tuning and shaping prior knowledge from experience, with a focus on learning to learn categories.
- Learning inductive frameworks, constructing or selecting appropriate hypothesis space for inductive generalization.

1.1 Bayesian Inference

The most basic question is how to update belief based on limited data we observe. A central assumption is that degrees of belief can be represented as probabilities: that our conviction in some hypothesis h can be expressed as a real number ranging from 0 to 1. The framework also assumes that learners represent probability distributions and that they use these probabilities to represent uncertainty in inference which turns maths of probability theory into the engine of inference.

Computing degrees of belief as probabilities depend on two components. One is called the prior probability and is denoted by $P(h_i)$ which captures how much we believe in h_i before observing the data d. The other is called the likelihood and is denoted by $P(d \mid h_i)$which captures the probability with which we would expect to observe the data d if h_i were true. The product of the two and divided by evidence gives the posterior probability of h_i, given via Bayes' rule:

$$P(h_i \mid d) = \frac{P(d \mid h_i)P(h_i)}{P(d)} \tag{1}$$

Consider the example of a person having a cough visiting a doctor. The Bayesian inference can be done as follows:

- The doctor would enquire about the problem a patient faces (helps in the construction of the possible hypothesis set).
- Then he will enquire about the locality he belongs to(to calculate the prior probability of each hypothesis to be true)
- Then he will calculate the likelihood of symptoms for each possible hypothesis.
- Using Bayes' formula, he can calculate the best hypothesis supporting the evidence. This will help him in finding the best cure for the disease.

For example, consider a person having a cough. So there are three possible hypothesis in this case: $h_{cold}; h_{lung-cancer}; h_{heartburn}$ with respective prior probabilities 0.5, 0.1, 0.4(chance of lung cancer is rare given no information about cough). Since we know likelihood of cough is more due to lung-cancer, P(d| h_{cold}) = 0.8 P(d| $h_{lung-cancer}$) = 0.9 P(d| $h_{heartburn}$) = 0.1. Using the Bayes' formula, the posterior probability of each hypothesis can be calculated.

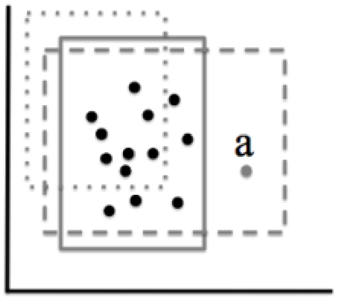

Figure 1: Example data and Hypothesis

In figure 1, dots represent individual data points generated independently from some process that is depicted in terms of a region or subset of space. The data are consistent with h_{solid} and h_{dashed}, but not h_{dotted}, since some of the points are not within the dotted rectangle. So likelihood of data to be generated from h_{solid} and h_{dashed} is greater than zero, but likelihood of data to be generated from h_{dotted} is 0. Bayesian inference can also yield predictions about unobserved data. One would only observe new data at position a if h_{dashed} is correct since P(a| h_{solid})= 0, but P(a| h_{dashed}) > 0

1.2 Inductive Generalization

It refers to the ability to infer beyond available data and arrive at a generalized conclusion. To validate the conclusion, additional assumptions that are required to be made are known as inductive constraints. For example, let us say we are given a bag of marbles. We sample marbles from the bag 3 times with replacement and observed that each time the marbles were of the same color. Then we can conclude that the marble that we draw next time will be of the same color. The assumption that all the marbles in the bag are uniform in color is an Inductive Constraint.

2 A Case Study: Learning names for object categories

Consider the task a child faces in learning the categories of the object which explains how the Bayesian analysis of inductive generalization is applied to cognitive development. Even a single word dog can refer to a large number of hypotheses like a living being, animal, mammal, labrador or four-legged creature, etc.

Now a child should select the most suitable hypothesis that defines the problem. They are said to possess the whole object bias which rules out part features like a four-legged creature. Children are said to possess strong prior knowledge about what sort of word meanings are natural which constrains the possible set of hypotheses and helps them acquire the meaning of words even in the presence of fewer examples.

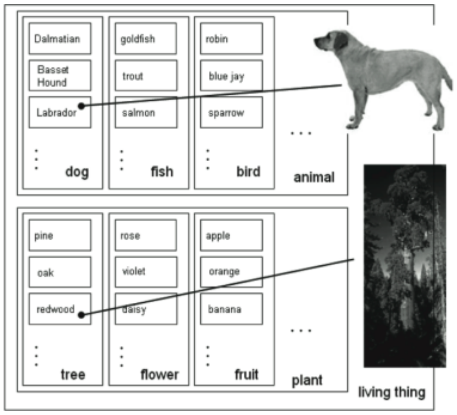

Figure 2: Hierarchical taxonomy for object categories

If we assume that multiple data points are generated independently from the concept, then the likelihood of h with n consistent examples is inversely proportional to the size of h, raised to the n^{th} power. This is known as the size principle. For example, consider the task of drawing marbles from two bags with replacement, one bag having red and green marbles, other having red, green and yellow marbles. If we want to find the probability of drawing marbles from the bag in the following sequence:[red, green, red, green], then the probability will be $1/2^4$ for the first bag and $1/3^4$ for the second bag.

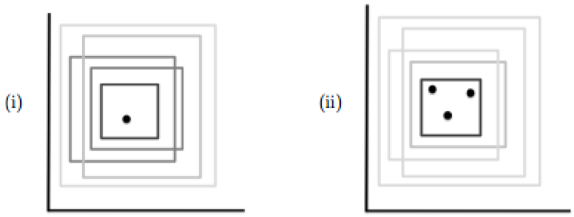

Figure 3: (i) Hypothesis space is depicted as a dot diagram, higher probability hypotheses are darker rectangles.With one data point, many hypotheses have some support. (ii) With three examples, the most restrictive hypothesis is strongly favored.

3 Inductive constraints

Inductive constraints are the additional set of assumptions that are required to validate the conclusion in case of inductive generalization and acquiring inductive constraints is said to enhance learning which is a learned overhypothesis.

3.1 Example of learning Overhypothesis

In the figure below we have distance a-b = a-c. So given the point a, we can't distinguish which point b or c will be similar which is just based on the first-order knowledge. So additional explanation would be required to find a similar point. For example, we would consider the region of hypothesis space that's oriented along y-axis which makes b having a higher preference over c. This is the Second-order knowledge or overhypothesis(l<w) that helps to find a solution. Knowledge at the higher level imposes weaker constraints as compared to the knowledge at the lower level. For example in the above case, l<w is the second-order knowledge, even if a learner doesn't know whether l is 10 units long and w is 20 units it is obvious that l is smaller than w.

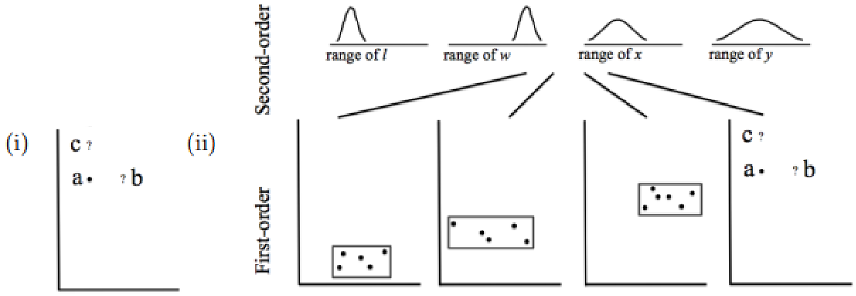

Figure 4: Learning higher-order information. (i) Given point a, one cannot identify whether b or c is more likely. (ii) Given additional data, a model that could learn higher-order information about hypotheses might favor regions that tend to be long, thin rectangles oriented along the y axis

3.2 Hierarchical Bayesian Framework

The notion of generalized knowledge is captured in the framework of the Hierarchical Bayesian Model where knowledge is represented at multiple levels of abstraction. As we move from the bottom level to top-level, we get more generalized information and the upper level imposes constraints on the lower level. Hierarchical Bayesian Models support both top-down and bottom-up inference [3].

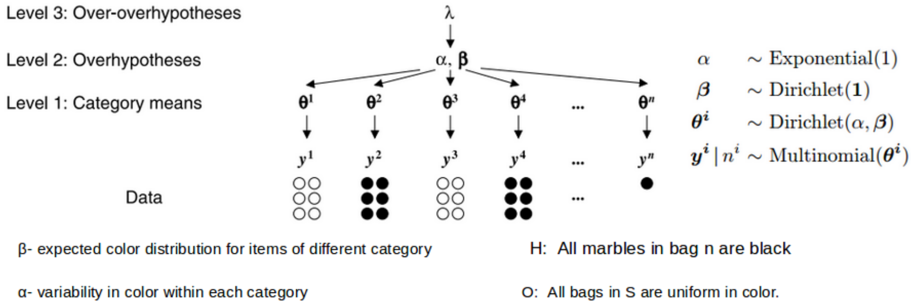
Level 3: Over-overhypotheses
Level 2: Overhypotheses
Level 1: Category means
Data
λ
α, β
θ¹ θ² θ³ θ⁴ ... θⁿ
y¹ y² y³ y⁴ ... yⁿ
$\alpha \sim$ Exponential(1)
$\boldsymbol{\beta} \sim$ Dirichlet(**1**)
$\boldsymbol{\theta^i} \sim$ Dirichlet($\alpha, \boldsymbol{\beta}$)
$\boldsymbol{y^i} \mid n^i \sim$ Multinomial($\boldsymbol{\theta^i}$)
β- expected color distribution for items of different category
α- variability in color within each category
H: All marbles in bag n are black
O: All bags in S are uniform in color.

Figure 5: A Hierarchical Bayesian Model

Figure illustrates an example of bottom-up inference where input is the observation of different bags of marble(y^i) and we are interested to predict the color of next marble to be drawn from bag n.The first step is to identify level 1 knowledge(θ^i) which is the distribution of colors in the bag which in turn is specified by level 2 knowledge(α, β) for each bag [4].

In the above example, modeler specifies a hyperparameter,λ which captures prior knowledge of α and β. We assume that the marbles responsible for the observations in y^i are drawn independently at random from the i^{th} bag, and the color of each depends on the color distribution θ^i of that bag. The vector θ^i are drawn from a Dirichlet distribution parameterized by a scalar α and a vector β.The larger the value of α, the more likely that color distribution for any given bag will be close to the vector β. When α is small, each individual bag is likely to be near-uniform in color and β will determine the relative proportion of ‘mostly black’ and ‘mostly white’ bags.

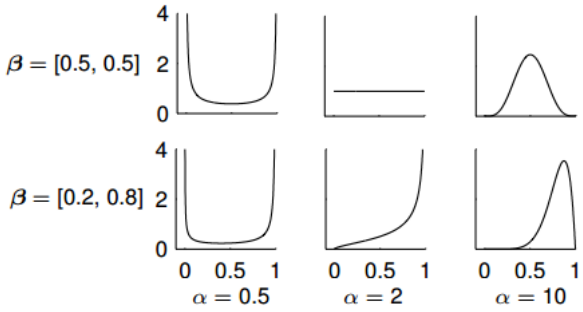

Figure 6: 2-D dirichlet distribution that serves as prior on θ

To fit the model to data, we assume that counts y are observed for one or more bags and use a Markov chain Monte Carlo (MCMC) scheme to draw a sample from p(α, β |y). Predictions about the color distribution of a new, sparsely observed bag(θ^{new}) can be computed by calculating the mean prediction made by all pairs (α, β) in the MCMC sample.

4 Developing Inductive Framework

Children are said to possess biological, psychological and causal knowledge at an early age and they update their knowledge as they gather more data. The Bayesian framework also supports the same notion of updating model as more and more data is observed and there is a trade-off between the simplicity of the model and the goodness of fit. For a simple model A, less number of parameters are required to define the region of space and is having more prior probability than a complex model C.

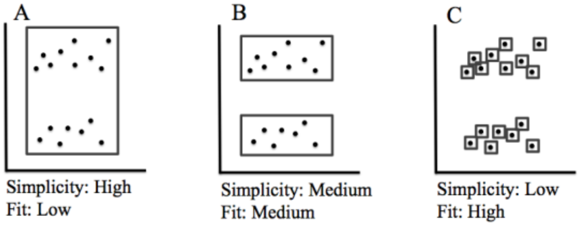

Figure 7: Hypothesis A is too simple, fitting the observed data poorly; C fits closely but is too complex; while B is just right."

Initially, with less amount of data, the simpler model is preferred. As more data is observed, more flexibility is required and we prefer a more complex model. Finally, a model that is closest to the generative process is said to be optimal and is preferred. Occam's Razor prefers the simplest model among all the preferred model that satisfies our requirement.

5 Discussion:

5.1 Optimality: What does it mean?

Bayesian probability theory is the set of unique, consistent rules for conducting plausible inference [2]. It is an extension of deductive logic to the case where propositions have a degree of truth or falsity. Bayesian probability theory is considered optimal in the sense that non-bayesian reasoning is less efficient then Bayesian reasoning. But before deciding if it can be considered as an appropriate tool for modeling cognition, it is required to study what optimal thinking corresponds to as human thinking is not optimal and is subjected to emotions, heuristics, and biases.

We make several decisions and often these decisions must be made in the face of uncertainty where the consequences or outcomes of our decisions are unclear. As a way of navigating this uncertainty humans rely upon heuristic rules that allow us to simplify complex tasks and make decisions efficiently. These heuristics are quite useful but sometimes they induce biases and leads to systematic errors [6]. Following three heuristics were examined:

1. **Representativeness:** In answering questions like what is the probability of A belonging to class B, people rely on the representativeness heuristic in which probability is evaluated by the degree to which A is a representative of B. For example, consider a man who is very shy and withdrawn, invariably helpful but with little interest in people or the world of reality. A meek and tidy soul, he needs order and structure, and a passion for detail. There are possibilities of the man being either a farmer, a pilot, a librarian or a physician. Based on the given description of him, it can be inferred that he is a librarian. But there are many more farmers than a librarian. This causes the error because representativeness is not influenced by several factors that should affect judgement of probability like they are insensitive to prior probability, insensitive to sample size, etc.

2. **Availability**: There are situations in which people assess the frequency of the class by the ease with which instances can be brought to mind. For example, one may assess the risk of heart attack among middle-aged people by recalling such occurrences among one's acquaintances. Availability is a useful clue for assessing frequency or probability but reliance on availability leads to predictable biases like biases due to the retrievability of instances, biases due to the effectiveness of a search set, etc.

3. **Anchoring:** Anchoring occurs when a starting point is given to a subject and estimates rely too heavily on it.It was illustrated using a study with two groups of high school students where each group was told to find the product of a sequence of numbers in a few seconds. While group 1 was asked to estimate $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$, group 2 was asked to estimate $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$. Because the result of the first few steps of multiplication is higher in descending sequence, former expression was judged larger than the latter.

Being able to understand what optimal reasoning is, is also useful for ideal learnability analysis. If some knowledge could not be learned by an optimal learner presented with the type of data children receive, it is safe to conclude that actual children could not learn it either or some of the assumptions in the model is inaccurate. So not all Bayesian models operate on the computational level and not all Bayesian models strive to capture optimal inference [7].

5.2 Biological Plausibility

Because cognitive scientists are ultimately interested in understanding human cognition, and human cognition is ultimately implemented in the brain, our computational-level explanations be realizable on the neurological level. Connectionist networks contain many interconnected neurons that communicate with each other by sending activation or inhibition through their connections and knowledge is represented in the distributed fashion over connections. As a result, representations degrade with neural damage and reasoning is fuzzy. Bayesian models may appear implausible from the neurological perspective because brains are neither transparent nor representation is explicitly defined.

5.3 Limitations Of Bayesian Models

- Many problems in cognitive science are not cast as inductive problems.Many scientists are concerned with understanding how different characteristics(IQ and attention) are related to each other.
- Bayesian models cannot explain the behavior of the brain if it emerges due to certain architecture of the brain.
- In the brain, knowledge is represented in the distributed fashion over connections. So reasoning appears to be fuzzy so the Brain does not always deal with probability theory.
- Sometimes when we have a large set of hypotheses, the computation of posterior is intractable and we have to approximate the distribution.

6 Conclusion

Bayesian models offer explanatory insights into many aspects of human cognition and development. The framework is valuable for defining optimal standard of inference and for exploring trade-off between simplicity and goodness of fit that must guide any learner's generalizations from observed data. Its representation flexibility makes it applicable for wide range of learning problems and its transparency makes it easy to be clear about what assumptions are being made, what is being learned and why learning works.

7 References

1. Griffiths T L, Chater N., Kemp C., Perfors A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases

2. Jaynes, E. (2003). Probability theory: The logic of science.

3. Kemp C (2008). The acquisition of Inductive Constraints

4. Kemp C., Perfors A. & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models

5. Sanborn, A., Griffiths, T. L. & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning.

6. Tversky A & Kahneman D. (1974). Judgment under uncertainty: Heuristics and biases.

7. John Kruschke. Doing Bayesian Data Analysis: A Tutorial Introduction with R