B

Bayesian Learning

Bradley C. Love[1], Matt Jones[2]
[1]Department of Psychology, The University of Texas at Austin, Austin, TX, USA
[2]University of Colorado, Boulder, CO, USA

Synonyms

Bayesian model; Normative; Probabilistic approaches; Rational

Theoretical Background

Bayesian methods have undergone tremendous progress in recent years, due largely to mathematical advances in probability and estimation theory (Chater et al. 2006). These advances have allowed theorists to express and derive predictions from far more sophisticated models than previously possible. These models have generated a good deal of excitement for at least two reasons. First, they offer a new interpretation of the goals of cognitive systems, in terms of inductive probabilistic inference, which has revived attempts at rational explanation of human behavior (Oaksford and Chater 2007). Second, Bayesian models may have the potential to explain some of the most complex aspects of human cognition, such as language acquisition or reasoning under uncertainty, where structured information and incomplete knowledge combine in a way that has defied previous approaches (e.g., Kemp and Tenenbaum 2008).

Constructing a Bayesian model involves two steps. The first step is to specify the set of possibilities for the state of the world, which is referred to as the hypothesis space. Each hypothesis can be thought of as a prediction by the subject about what future sensory information will be encountered. However, the term hypothesis should not be confused with its more traditional usage in psychology, connoting explicit testing of rules or other symbolically represented propositions. In the context of Bayesian modeling, hypotheses need have nothing to do with explicit reasoning, and indeed the Bayesian framework makes no commitment whatsoever on this issue. For example, in Bayesian models of visual processing, hypotheses can correspond to extremely low-level information, such as the presence of elementary visual features (contours, etc.) at various locations in the visual field (Geisler et al. 2001). There is also no commitment regarding where the hypotheses come from. Hypotheses could represent innate biases or knowledge, or they could have been learned previously by the individual. Thus, the framework has no position on nativist–empiricist debates. Furthermore, hypotheses representing very different types of information (e.g., a contour in a particular location, whether or not the image reminds you of your mother, whether the image is symmetrical, whether it spells a particular word, etc.) are all lumped together in a common hypothesis space and treated equally by the model. Thus, there is no distinction between different types of representations or knowledge systems within the brain. In general, a hypothesis is nothing more than a probability distribution. This distribution, referred to as the *likelihood function*, simply specifies how likely each possible pattern of observations is according to the hypothesis in question.

The second step in constructing a Bayesian model is to specify how strongly the subject believes in each hypothesis before observing data. This initial belief is expressed as a probability distribution over the hypothesis space, and is referred to as the *prior*. The prior can be thought of as an initial bias in favor of some hypotheses over others, in that it contributes extra "votes" (as elaborated below) that are independent of any actual data. This decisional bias allows the model's predictions to be shifted in arbitrary directions regardless of the data. As we discuss below, the prior can be a strong point of the model if it is derived independently, from empirical statistics of real environments. However, more commonly, the prior is chosen ad hoc, providing substantial unconstrained flexibility to models that are advocated as rational and assumption-free.

Together, the hypotheses and the prior fully determine a Bayesian model. The model's goal is to decide how strongly to believe in each hypothesis after data have been observed. This final belief is again expressed as a probability distribution over the hypothesis space and is referred to as the *posterior*. The statistical identity known

Norbert Seel (ed.), *Encyclopedia of the Sciences of Learning*, DOI 10.1007/978-1-4419-1428-6,

as Bayes' Rule is used to combine the prior with the observed data to compute the posterior. Bayes' Rule can be expressed in many ways, but here we explain how it can be viewed as a simple vote-counting model. Specifically, Bayesian inference is equivalent to tracking evidence for each hypothesis, or votes for how strongly to believe in each hypothesis. The prior provides the initial evidence counts, E_{prior}, which are essentially made-up votes that give some hypotheses a head start over others, before observing any actual data. When data are observed, each observation adds to the existing evidence according to how consistent it is with each hypothesis. The evidence contributed for a hypothesis that predicted the observation will be greater than the evidence for a hypothesis under which the observation was unlikely. The evidence contributed by the *i*th observation, E_{data_i}, is simply added to the existing evidence to update each hypothesis' count. Therefore the final evidence, $E_{\text{posterior}}$, is nothing more than a sum of the votes from all of the observations, plus the initial votes from the prior. (Formally, $E_{\text{posterior}}$ equals the logarithm of the posterior distribution, E_{prior} is the logarithm of the prior, and $E_{data}(H)$ is the logarithm of the likelihood of the data under hypothesis *H*. The model's prediction for the probability that hypothesis *H* is correct, after data have been observed, is proportional to exp $[E_{\text{posterior}}(H)]$).

$$E_{\text{posterior}}(H) = E_{\text{prior}}(H) + \sum_i E_{data_i}(H) \qquad (1)$$

This sum is computed for every hypothesis, *H*, in the hypothesis space. The vote totals determine how strongly the model believes in each hypothesis in the end. Thus, any Bayesian model can be viewed as tracking evidence for each hypothesis, with initial evidence coming from the prior and additional evidence coming from each new observation. At its core, this is all there is to Bayesian modeling.

To illustrate these two steps and how inference proceeds in a Bayesian model, consider the problem of determining whether a fan entering a football stadium is rooting for the University of Southern California (USC) Trojans or the University of Texas (UT) Longhorns based on three simple questions: (1) Do you live by the ocean? (2) Do you own a cowboy hat? (3) Do you like Mexican food? The first step is to specify the space of possibilities (i.e., hypothesis space). In this case, the hypothesis space consists of two possibilities: being a fan of either USC or UT. Both of these hypotheses entail probabilities for the data we could observe, for example, $\text{P(ocean}|\text{USC)} = .8$ and $\text{P(ocean}|\text{UT)} = .3$. Once these probabilities are given, the two hypotheses are fully specified. The second step is to specify the prior. In many applications, there is no principled way of doing this, but in this example, the prior corresponds to the probability that a randomly selected person will be a USC or a UT fan, that is, one's best guess as to the overall proportion of USC and UT fans in attendance.

With the model now specified, inference proceeds by starting with the prior and accumulating evidence as new data are observed. For example, if the football game is being played in Los Angeles, one might expect that most people are USC fans, and hence the prior would provide an initial evidence count in favor of USC. If our target person responded that he lives near the ocean, this observation would add further evidence for USC. The magnitudes of these evidence values will depend on the specific numbers assumed for the prior and for the likelihood function for each hypothesis, but all that the model does is take the evidence values and add them up. Each new observation adds to the balance of evidence among the hypotheses, strengthening those that predicted it relative to those under which it was unlikely.

There are several ways in which real applications of Bayesian modeling become more complex than the simple example above. However, these all have to do with the complexity of the hypothesis space rather than the Bayesian framework itself. For example, many models have a hierarchical structure, in which hypotheses are essentially grouped into higher-level *overhypotheses*. Overhypotheses are generally more abstract and require more observations to discriminate among; thus hierarchical models are useful for modeling learning or change over developmental timescales (e.g., Kemp et al. 2007). However, each overhypothesis is just a weighted sum of elementary hypotheses, and inference among overhypotheses comes down to exactly the same vote-counting scheme as described above. As a second example, many models assume special mathematical functions for the prior, such as conjugate priors, that simplify the computations involved in updating evidence. However, such assumptions are generally made solely for the convenience of the modeler, rather than for any psychological reason related to the likely initial bias of a human subject. Finally, for models with especially complex hypothesis spaces, computing exact predictions often becomes computationally intractable. In these cases, sophisticated approximation schemes are used, such as Markov-chain Monte Carlo (MCMC) or particle filtering (i.e., sequential Monte Carlo). These algorithms yield good estimates of the model's true predictions while requiring far less computational effort. However, once again they are used for the convenience of the modeler and usually are not meant as proposals for how human subjects might solve the same computational problems.

To summarize: Hypotheses are probability distributions and have no necessary connection to explicit reasoning. The model's predictions depend on the initial biases on the hypotheses (i.e., the prior). The heart of Bayesian inference – combining the prior with observed data to reach a final prediction – is formally equivalent to a simple vote-counting scheme. Learning and one-off decision-making both follow this scheme, and are identical except for timescale and specificity of hypotheses. Most of the elaborate mathematics that often arises in Bayesian models comes from the complexity of their hypothesis sets or the tricks used to derive tractable predictions, which generally have little to do with the psychological claims of the researchers. Bayesian inference itself, aside from its assumption of optimality and close relation to vote-counting models, does not make psychological claims in recards to representational format, encoding, retrieval, attention, etc. However, the flexibility and power of the Bayesian framework has allowed researchers to model complex learning and decision-making behaviors that have proven intractable or unwieldly under other formulations.

Important Scientific Research and Open Questions

The restriction to computational-level accounts (cf. Marr 1982) severely limits contact with process-level theory and data. Rational approaches attempt to explain *why* cognition produces the patterns of behavior that it does, but they offer no insight into *how* cognition is carried out. Second, in general, there are multiple rational theories of any given task, corresponding to different assumptions about the environment and the learner's goals. Consequently, there is insufficient acknowledgement of these assumptions and their critical roles in determining model predictions. It is extremely rare to find a comparison among alternative Bayesian models of the same task to determine which is most consistent with empirical data. Likewise, there is little recognition when the critical assumptions of a Bayesian model logically overlap closely with those of other theories. These challenges are currently being addressed by members of the Bayesian community. The end goal is to integrate Bayesian approaches with what we know about the mental processes that support learning and decision making (Jones and Love 2011).

Cross-References

- ► Concept Learning
- ► Human Causal Learning
- ► Human Cognition and Learning
- ► Human Learning
- ► Learning Algorithms
- ► Mathematical Models/Theories of Learning
- ► Metatheories of Learning
- ► Normative Reasoning and Learning

References

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287–291.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*, 711–724.

Jones, M., & Love, B. C. (in press, 2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*, 10687–10692.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*, 307–321.

Marr, D. (1982). *Vision.* San Francisco: W.H. Freeman.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford: Oxford University Press.