Bayesian Modelling for Machine Learning

Paul Rippon
The University of Newcastle, Australia

Kerrie Mengersen
The University of Newcastle, Australia

INTRODUCTION

Learning algorithms are central to pattern recognition, artificial intelligence, machine learning, data mining, and statistical learning. The term often implies analysis of large and complex data sets with minimal human intervention. Bayesian learning has been variously described as a method of updating opinion based on new experience, updating parameters of a process model based on data, modelling and analysis of complex phenomena using multiple sources of information, posterior probabilistic expectation, and so on. In all of these guises, it has exploded in popularity over recent years.

General texts on Bayesian statistics include Bernardo and Smith (1994), Gelman, Carlin, Stern, and Rubin (1995), and Lee (1997). Texts that derive more from the information science discipline, such as Mitchell (1997) and Sarker, Abbass, and Newton (2002), also include sections on Bayesian learning.

Given recent advances and the intuitive appeal of the methodology, Bayesian learning is poised to become one of the dominant platforms for modelling and analysis in the 21st century. This article provides an overview of Bayesian learning in this context.

BACKGROUND

Bayesian Modelling

Bayesian learning aims to provide information about unknown characteristics of a population (such as a mean and/or a variance) or about relationships between characteristics (for example, via a regression equation or a neural network). We often have a set of alternative models or hypotheses, H_1, H_2,…, H_m, that could describe these unknowns, such as possible values for the unknown mean or alternative neural network representations. The Bayesian approach allows prior beliefs about these models to be updated in the light of new data. The fundamental enabling mechanism is Bayes' rule:

$$p(H_i \mid D) = \frac{p(D \mid H_i)\, p(H_i)}{p(D)}, \qquad (1)$$

which states that the *posterior* probability $p(H_i|D)$ of a particular model, H_i, conditional on data, D, is proportional to the probability $p(D|H_i)$ of the data, given the model multiplied by the *prior* probability $p(H_i)$ of the model. The denominator $p(D)$, a normalizing constant designed to make the posterior probability sum or integrate to one, can be termed the probability of the data and is expressed as

$$p(D) = \sum_{i=1}^{m} p(H_i)\, p(D \mid H_i).$$

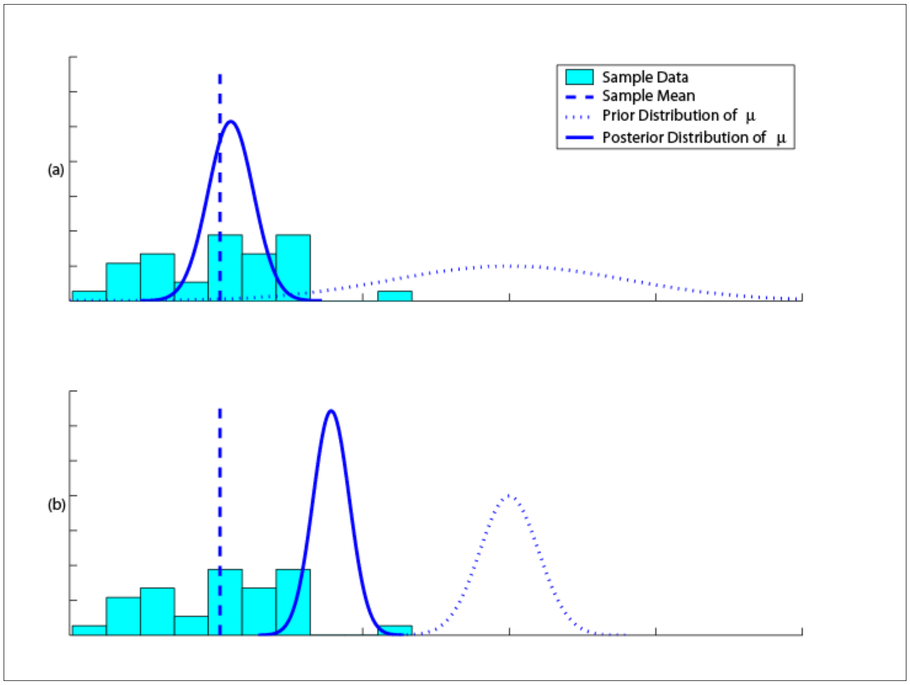
The number of plausible models might be infinite, for example, when the different models are represented by unknown values of a continuously distributed population mean. In this case, probability distributions become densities and the summation in $p(D)$ is replaced by an integral. In either case, it is this denominator, $p(D)$, that is often intractable. This motivates the development of numerical methods such as Markov chain Monte Carlo, described in the next section.

As a simple example, consider sampling n data points y_1, y_2,…, y_n from a population of normally distributed measurements in order to estimate an unknown mean, μ, and assume that the population variance, σ^2, is known. Thus, H is the set of all possible values that μ may take. The sample mean, $\bar{y}$, represents the information contained in the data so that $p(D|H)=p(\bar{y}|\mu)=N(\mu,\sigma^2/n)$.

In practice, we often have some prior knowledge about μ, such as, "μ is known from experience to be around a value μ_0." We might express this prior knowledge as $\mu \sim N(\mu_0, \tau_0^2)$, where τ_0^2 represents the uncertainty around the best guess, μ_0. Now, according to Bayes' rule:

$$p(H \mid D) = p(\mu \mid \bar{y}) \propto p(\bar{y} \mid \mu)\, p(\mu) = N(\mu_n, \tau_n^2) \qquad (2)$$

Figure 1.



with

$$\mu_n = \left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2} \right) \tau_n^2 \quad \text{and} \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

The posterior distribution can be considered as a merging of the prior opinion about μ and the data. Figure 1 illustrates this updating of opinion about μ for different priors.

In 1a, the prior knowledge about μ is fairly vague. Thus, the posterior distribution for μ is dominated by the data. In 1b, the prior knowledge about μ is more precise and has more influence on the posterior distribution.

Bayesian Computation

Continuing this example, suppose more realistically that σ^2 is also unknown. A distribution that reflects theoretical properties of a variance is the inverse Gamma distribution, so we might take $\sigma^2 \sim IG(a_0, b_0)$, where a_0 and b_0 are chosen to reflect our prior knowledge. Application of Bayes' rule results in a joint posterior distribution of μ and σ^2 that is nonstandard and multidimensional, making analytical solutions difficult.

A popular numerical solution is Markov chain Monte Carlo (MCMC). MCMC algorithms allow simulation from a Markov chain whose stationary distribution is $p(H|D)$. If it is not easy to simulate directly from $p(H|D)$, values can be proposed from some easily simulated distribution (such as uniform or normal) and accepted or rejected according to a rule that ensures that the final set of accepted values are from the target posterior distribution. If $p(H|D)$ is high dimensional, it can often be decomposed into a series of lower dimensional, conditional distributions, and (possibly different) MCMC algorithms can iterate around these, eventually forming a sample from the joint distribution (Besag, 1974).

For this example problem, a basic MCMC algorithm would be as follows.

- Choose initial values μ_1, σ_1^2
- Repeat for i=2:k for k large
  - Randomly draw σ_i from $\mu \mid \sigma_{i-1}^2, \bar{y}$ given in Equation (2).
  - Randomly draw σ_i^2 from the conditional posterior distribution $\sigma^2 \mid \mu_i, \bar{y} \sim IG(a_n, b_n)$, where $a_n = a_0 + \frac{1}{2}$ and $b_n = b_0 + \frac{n}{2}(\bar{y} - \mu_i)^2$
- Discard the first part of the above chain as "burn-in," in which the Markov chain is still approaching the target posterior distribution from the (perhaps unlikely) initial values.
- The remaining iterates, μ_i and σ_i^2, represent a large sample from the target posterior distribution. Graphs