A crash course in Bayesian statistics

Kenneth Tay and Stephen Bates

August 8, 2019

1 Introduction to Bayesian statistics

Let's say we have some parameter θ which we wish to estimate, and that we have data X to help us with this.

In the frequentist paradigm, we assume that θ is *fixed*, and that given θ, our data X has some distribution $p(X \mid \theta)$. As a function of θ (X fixed), $p(X \mid \theta)$ is called the likelihood function. One way to estimate θ here is to maximize $p(X \mid \theta)$ w.r.t. θ, giving the MLE.

In the **Bayesian paradigm**, we assume instead that θ is randomly drawn from some **prior distribution** $p(\theta)$. The prior distribution is a reflection of what we believe about θ before we see the data X. Once we see the data X, we update our beliefs in the distribution of θ. This is done using Bayes rule:

$$p(\theta \mid X) = \frac{p(\theta)p(X \mid \theta)}{\int p(\theta)p(X \mid \theta)d\theta}. \tag{1}$$

The quantity on the LHS above is called the **posterior distribution**, and this is the fundamental object in Bayesian statistics. With the posterior distribution, we can compute any statistic that we want, and we are frequently interested in

- the posterior mean, $\mathbb{E}[\theta \mid X]$, as an estimator for θ
- the posterior mode, $\arg\max_{\theta} p(\theta \mid X)$, as an estimator for θ
- the posterior variance, to asses uncertainty and give a credible interval for θ
- quantiles of the posterior distribution, to form a credible interval

Recall that a *credible interval* is a interval I such that $P(\theta \in I \mid X) = 1 - \alpha$ for pre-determined α. We can form a credible interval with quantiles of the posterior distribution, but this may be hard to compute, and becomes hard to do the analagous operation in high-dimensions. Instead, we can form an approximate credible interval by using a Gaussian approximation to the posterior.

2 Hierarchical models

Let's consider a concrete example[1]: estimating the probability θ of a disease in a population of interest. If we have a sample of n subjects and if X denotes the number of subjects in the sample who have the disease, then $X \mid \theta \sim \text{Binom}(n, \theta)$. If we pick a fixed prior $\theta \sim \text{Beta}(\alpha, \beta)$ (α and β fixed), then the posterior distribution is given by $\theta \mid X \sim \text{Beta}(\alpha + X, \beta + n - X)$. In this context, θ is the parameter of interest; α and β are called **hyperparameters**.

The above works if we know what α and β are; in practice we don't. One approach is to set α and β so that $\text{Beta}(\alpha, \beta)$ is as "non-informative" as possible (e.g. $\text{Beta}(1, 1) = U[0, 1]$). If we have other samples on hand, we could use them to estimate α and β. Assuming we have J samples of size $n_1, \ldots, n_J$, we consider the following model:

$$\begin{aligned} \theta_j &\overset{iid}{\sim} \text{Beta}(\alpha, \beta), \quad j = 1, \ldots, J, \\ X_j \mid \theta_j &\overset{ind}{\sim} \text{Binom}(n_j, \theta_j), \quad j = 1, \ldots, J. \end{aligned}$$

We can use the observed mean and standard deviation of the $\frac{X_j}{n_j}$'s to estimate α and β. This approach is known as the **empirical Bayes** approach.

Hierarchical modeling is yet a different approach. Instead of assuming that α and β are fixed, we give them their own prior distributions (which we call **hyperpriors**). Usually, we try to make these hyperprior distributions as "non-informative" as possible to reflect our ignorance of the unknown hyperparameters. In our example, we may want the beta distribution's mean $\frac{\alpha}{\alpha+\beta}$ to be uniformly distributed on $[0, 1]$ and $\frac{1}{\sqrt{\alpha+\beta}}$, an approximation of the standard deviation, to be uniformly distributed on $[0, \infty)$. The hierarchical model defined by this is as follows::

$$\begin{aligned} \frac{\alpha}{\alpha + \beta} \sim U[0, 1], \qquad \frac{1}{\sqrt{\alpha + \beta}} &\sim U[0, \infty), \\ \theta_j &\overset{iid}{\sim} \text{Beta}(\alpha, \beta), \quad j = 1, \ldots, J, \\ X_j \mid \theta_j &\overset{ind}{\sim} \text{Binom}(n_j, \theta_j), \quad j = 1, \ldots, J. \end{aligned}$$

[1] This is a slight variation of the rat tumor example in Gelman et al.'s *Bayesian Data Analysis, 3rd ed.*, pp102-111.

As with any Bayesian set-up, there is a lot of disagreement on what the appropriate prior/hyperprior should be.

The standard Bayesian machinery goes through with hierarchical models, albeit in a more complicated form. We usually compute 3 things:

1. The *joint posterior distribution* of all the parameters:
$$p(\theta, \alpha, \beta \mid X) \propto p(\alpha, \beta) \cdot p(\theta \mid \alpha, \beta) \cdot p(X \mid \theta, \alpha, \beta).$$
2. The *conditional posterior distribution* for the parameters given the hyperparameters: $p(\theta \mid \alpha, \beta, X)$.
3. The *marginal posterior distribution* for the hyperparameters: $p(\alpha, \beta \mid X) = \int p(\theta, \alpha, \beta \mid y) d\theta$.

We then draw samples from the posterior distribution in the following way:

1. Draw hyperparameters α, β from marginal posterior $p(\alpha, \beta \mid y)$.
2. Draw parameter θ from the conditional posterior distribution, $p(\theta \mid \alpha, \beta, X)$, given the drawn value of (α, β).

3 Sampling in Bayesian models

In section 1, we saw that the posterior distribution is the key to Bayesian inference. Let π denote the posterior distribution $p(\theta \mid X)$. If we can get S samples θ^s from π, then for any function h we can estimate

$$\mathbb{E}[h(\theta) \mid X = x] = \int h(\theta)\pi(\theta)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^s). \tag{2}$$

Thus, we want to find good algorithms for sampling from π. In some cases, the integral in the denominator in (1) cannot be computed tractably; when this happens, we need algorithms that can sample from an unnormalized density π_u.

Here are some sampling methods that can be used:

- Acceptance-rejection sampling[2],

[2]See Section 4.7 of http://statweb.stanford.edu/~owen/mc/Ch-nonunifrng.pdf for details.

- Importance sampling[3],
- Markov chain Monte Carlo (MCMC) methods (e.g. Gibbs sampling or the Metropolis–Hastings algorithm),

3.1 Gibbs sampling

Let's say we want to draw samples $X \sim \pi$, where $X \in \mathbb{R}^d$. In some cases, while we cannot sample from π directly, we may have access to the **full conditional distributions**, i.e.

$$X_j \mid X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d, \qquad j = 1, \ldots, d.$$

Gibbs sampling gives us a way to use these full conditional distributions to get samples from π. This can be very attractive when the full conditional distributions have closed forms. The algorithm is as follows:

1. Initialize $x_0 \in \mathbb{R}^d$.
2. For $s = 1, \ldots, S$:
   (a) Let j be the index in $\{1, \ldots, d\}$ such that $j = s \mod d$.
   (b) Draw z from the full conditional distribution $X_j \mid X_1 = x_{s-1,1}, \ldots, X_{j-1} = x_{s-1,j-1}, X_{j+1} = x_{s-1,j+1}, \ldots, X_d = x_{s-1,d}$.
   (c) Set
   $$x_{s,i} = \begin{cases} z & \text{if } i = j, \\ x_{s-1,i} & \text{if } i \neq j. \end{cases}$$

The $x_0, x_1, \ldots, x_S$ constitute our sample from π. Note that the x_i's can be very correlated with each other: adjacent x_i are equal in all but one coordinate.

The algorithm as described above is called the **systematic scan Gibbs sampler**, as we sweep through the d coordinates systematically. The **random scan Gibbs sampler** replaces step 2(a) with $j \sim \text{Unif}\{1, \ldots, d\}$.

In this simple form, Gibbs sampling is a special case of the Metropolis–Hastings algorithm.

[3] See http://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf for details.

3.2 Metropolis–Hastings

Metropolis–Hastings is the most widely used MCMC method, because it applies to a wide variety of cases. As above, suppose we wish to draw samples $X \sim \pi$ where $X \in \mathbb{R}^d$. For each $w \in \mathbb{R}^d$, let $g(\cdot \mid w)$ be some density that is easy to sample from. This is specified by the user, and this can be anything, provided the support of g contains the support of π.

1. Initialize $x_0 \in \mathbb{R}^d$.
2. For $s = 1, \ldots, S$:
   (a) Draw $W \sim g(\cdot \mid x_{s-1})$
   (b) With probability $\min\left(1, \frac{\pi(W)g(x_{s-1}|W)}{\pi(x_s)g(W|x_{s-1})}\right)$, set $x_s = W$. Otherwise, set $x_s = x_{s-1}$.

Again, $x_0, x_1, \ldots, x_S$ form an approximate sample from π, and as always, these points may be highly correlated. A good MCMC sample will have S large enough such that there are many "indpendent" samples, and precisely quantifying when this will happen is the subject of a lot of research in Bayesian statistics.