STA 141C - Big Data & High Performance Statistical Computing

Week 9-1: Bayesian method and Markov chain Monte Carlo

Instructor: Bo Y.-C. Ning

May 24, 2022

Announcement

- Group presentation next week

Today

- Bayesian method and Markov chain Monte Carlo

References

- [video]

Frequentism and Bayesianism

- Before we get started, let's watch this [video] by Jake vanderPlas, now works at Google.
- As Jake mentioned in the video, comparing to the frequentist problem, the Bayesian problem starts with an important additional ingredient, which is the prior distribution.
- The key feature of a Bayesian analysis is that a Bayesian update her belief of evidence from the prior distribution after seeing the data
- Bayesian rule provides a mathematical formula to integrate the prior distribution and the likelihood function:

$$P(\theta \mid X) \propto \frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta)d\theta}$$

  - $f(X \mid \theta)$: likelihood
  - $\pi(\theta)$: prior distribution
  - $p(X) = \int f(X \mid \theta)\pi(\theta)d\theta$: evidence

How to choose the prior?

There are many ways to choose a prior, some popular ones are

- Elicitation from experts

  This is perhaps the most nature way, you choose the prior distribution through consulting experts' opinions. See the paper [here].

- Conjugate priors

  Before the modern computation era, inference of Bayesian methods often rely on the use of "convenient" priors. Those priors often lead to a closed form for the posterior, which can be easily used for computation. For example, suppose independent $x_1, \ldots, x_n \sim N(\theta, 1)$, a nature prior for θ is the normal distribution. As a result, the posterior $\pi(\theta \mid x_1, \ldots, x_n)$ also follows a normal distribution. There are many more conjugate prior: see [here].

  A major criticism from non-Bayesian is the use of those "convenient priors", as they are often unrealistic.

How to choose the prior? (cont'd)

- Non-informative prior

  Since conjugate priors often considered as unrealistic representations of prior uncertainty but are convenient to work with, the non-informative prior can "rescue" the unrealistic side and also keep the convenient side.

  Since in many cases, a real prior is hard to obtain, the strategy behind the "non-informative prior" is to choose a prior such that it will "influences" the posterior as little as possible. In other words, a Bayesian tries to be objectively "non-informative" about the parameter to be estimated. James O. Berger is one of the key authorities in this field. See his article [here]

  For example, to estimate the mean when likelihood is normal, as we have no information about θ, one can choose the prior $\pi(\theta) \propto 1$. Then the posterior mean of θ equals to the frequentist estimator (namely MLE).

  There are many more non-informative priors, another example is the Jeffreys' prior. However, this course is not about Bayesian analysis, so I am not going to the details of this part.

Frequentist analysis of Bayesian posteriors

- On the rise of nonparametric and high-dimensional statistical inference, finding a non-informative prior is much harder, as the model is infinite-dimensional.
- Choosing a suitable prior in general is even more difficult. In particular, David Freedman argued that except for a relatively small collection of pairs, posterior inconsistent is a common phenomenon!
- While this result cautions about the naive use of Bayesian methods, it does not mean that Bayesian methods are useless. Indeed, a pragmatic Bayesian's only aim may be to find some prior that is consistent at various parameter values.
- Frequentist analysis of Bayesian posteriors is the field which studies Bayesian procedure from the frequentist point of view (e.g., consistency, comparing confidence sets with credible sets). Representative scholars include Aad van der Vaart, Subhashis Ghoshal, Ismaël Castillo, Judith Rousseau, (theory), Christian Robert, David Dunson, Edward I. George, (methodology) and many many others.

Hierarchical Bayes models

The advantage of adopting a Bayesian method is that it provides a simple solution to solve more complicated problems.

Consider the model, $x_i \sim N(\theta, \sigma^2)$, x_is are independent. From Bayes rule, the posterior of the unknown parameters is

$$p(\theta, \sigma^2|X) = \frac{f(X|\theta, \sigma^2)\pi(\theta, \sigma^2)}{p(X)}.$$

The probability rule tells us $\pi(\theta, \sigma^2) = \pi(\theta|\sigma^2)\pi(\sigma^2)$. This leads us to consider putting the prior: $\theta|\sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim$ inverseGamma(α, β).

For more complicated model, one can follow a similar rule to construct the prior

$$\pi(\alpha_1, \alpha_2, \ldots, \alpha_p) = \pi(\alpha_1|\alpha_2, \ldots, \alpha_p)\pi(\alpha_2|\alpha_3, \ldots, \alpha_p)\pi(\alpha_p).$$

Empirical Bayes

The empirical Bayes method uses the prior that is estimated from the data. This procedure is quite different from the standard Bayesian method, which the prior is chosen to be fixed before observing the data.

Consider the normal model with unknown variance again, if the parameter of interest is θ, for the hierarchical Bayes approach, one obtain $p(\theta|X) = \int p(\theta, \sigma^2|X)d\sigma^2$ through marginalization.

The empirical Bayes approach instead of putting a prior on σ^2, it plugging-in an estimated value of σ^2 (e.g., the mle) and use the prior $\pi(\theta) = N(0, \hat{\sigma}^2)$. Then the posterior of θ becomes $p(\theta|X, \hat{\sigma}^2) = p(\theta|X)$.

The empirical Bayes framework developed and named by Robbins (1956) and further developed by Bradley Efron.

Computation: the MCMC algorithm

The MCMC algorithm and the rise of high-speed computation power lead us possible to compute large and complex Bayesian hierarchical models.

The idea of the MCMC algorithm is really simple: it creates samples from the posterior distribution through Monte Carlo sampling.

Suppose you want to simulate the posterior distribution of two unknown parameters $p(\theta_1, \theta_2 \mid X)$, the MCMC algorithm is given as follows:

- Step 1. pick an initial value for $\theta_2^{(0)}$
- Step 2. Repeat $t = 1, \ldots, T$:
  - draw $\theta_1^{(t)}$ randomly from $p(\theta_1 \mid X, \theta_2^{(t-1)})$
  - draw $\theta_2^{(t)}$ randomly from $p(\theta_2 \mid X, \theta_1^{(t)})$
- Output $(\theta_1^{(1)}, \theta_2^{(1)}), \ldots, (\theta_1^{(T)}, \theta_2^{(T)})$

Discard the first few draws and use the remaining draws to calculate the posterior mean, credible intervals, etc..

Gibbs sampling

When the conditional posterior has a close form (e.g., using the conjugate prior), the algorithm in the last page is called the Gibbs sampling.

For the normal mean with unknown variance model: $x_i \sim N(\theta, \sigma^2)$, $\theta \mid \sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim \text{inverseGamma}(\alpha, \beta)$.

We can write the conditional posterior distributions

- $p(\theta \mid X, \sigma^2) = N\left(\frac{\sum_{i=1}^{n} x_i}{n+1}, \frac{\sigma^2}{n+1}\right)$
- $p(\sigma^2 \mid X, \theta) = \text{inverseGamma}\left(\frac{n+1}{2} + \alpha, \beta + \frac{\sum_{i=1}^{n}(x_i-\theta)^2+\theta^2}{2}\right)$

The Gibbs sampling repeatedly sample θ and σ^2 from the two conditional distributions.

Example: [here]

Metropolis-Hasting algorithm

When the prior is non-conjugate and the even the conditional posterior distribution does not has a closed form, we use the Metropolis-Hasting algorithm.

The metropolis-hasting algorithm samples a parameter, say θ, from a proposal density. In general we denote $\theta^{prop} \sim g(\theta|x, \vartheta)$. This proposal density usually has a closed form and easy to draw sample from.

After we obtain the proposal density, we go to the accept and reject step:

- Calculate $\alpha = p(\theta^{prop})/p(\theta^{(t-1)})$
- Sample a $u \sim$ Uniform$[0, 1]$
- Let $\theta^{(t)} = \theta^{prop}$ if $u < \alpha$

  Let $\theta^{(t)} = \theta^{(t-1)}$ if $u > \alpha$.

In fact, the Gibbs sampling is a special case of the metropolis-hasting algorithm when proposals are always accepted.

Issues with MCMC

- Draws are not independent
- No universal stopping rule, convergence of the algorithm can be slow.
- Difficult for parallel computing
- Need to run convergence diagnostics