Dean&Francis

ISSN 2959-6157

Predictive Analysis of Diseases Using Bayesian Inference and Markov Chain Monte Carlo Methods

Sirong Chen

Beijing National Day School, Beijing, China

Corresponding author: yangjunaidada@tzc.edu.cn

Abstract:

Predictive Analysis of Diseases Using Bayesian Inference and Markov Chain Monte Carlo Methods" focuses on applying Bayesian inference and Markov Chain Monte Carlo to make prediction and analyzation of certain disease, including a general case study and Alzheimer's Disease. The first application shows how, in the event that a patient receives a false diagnosis, the Bayesian inference can be used to test the likelihood that the patient has a condition. This example demonstrates how Bayesian inference can lead to a closer objective reality by iteratively modifying the prior probability to the posterior probability depending on new information. In the second application on Alzheimer's disease, the regression model parameters (intercept and slope) are estimated using the Gibbs sampling approach and two datasets including data on Alzheimer's patients. The final model, tested on a separate dataset, achieved an accuracy of 62%. The study demonstrates the potential of Bayesian and MCMC approaches in disease prediction, suggesting a pathway to more robust models in medical analytics.

Keywords: Bayesian inference, Gibbs Sampling, Markov Chain Monte Carlo, logistic regression.

1. Introduction

As some diseases have high mortality and high rate of misdiagnosis, fields like medical research and public health have long been finding robust predictive models for predicting disease accurately and effectively to offer patients the precise result, make the most wise decision, and efficient resource allocation. Traditional statistical methods for prediction have several limitations in dealing with complex situations and large dataset. When paired with Markov Chain Monte Carlo (MCMC) techniques, Bayesian inference provides a potent framework for addressing these issues. By integrating prior knowledge with new evidence, Bayesian inference is able to make more accurate results [1].

Recent studies have highlighted the effectiveness of combining Markov Chain Monte Carlo with Bayesian inference to make predictions in various fields, including using Bayesian approach to predict perfor-

mance of a student, deformation during tunnel construction, etc. Still, Bayesian inference and MCMC are valuable in predicting disease, and recent research uses gene expression data to predict cancer survival times, patient radio-sensitivity, etc. Furthermore, MCMC algorithms such as Gibbs sampling have proven effective in estimating complex model parameters where traditional methods fall short. These techniques are particularly valuable in the case of Alzheimer's disease since the various variables in such disease, including patients' age, genetic sensitivity. Given this complexity, the combination of Bayesian Inference and MCMC can provide with a more accurate predictions, ultimately leading to better decision-making and personalized treatment strategies [2].

The structure of this research study is as follows: The fundamentals of Markov Chain Monte Carlo and Bayesian inference are covered in Section 2, including dealing with random variable case, calculating the posterior distribution, selecting the prior distribution in Bayesian inference, and Importance sampling and Gibbs Sampling in MCMC. The third section are two applications of applying Bayesian inference and MCMC to make disease predictions, including predicting a general assumed disease and Alzheimer's disease. The fourth part of the paper demonstrates the conclusion, discussion, and future expectations.

2. Methods and Theory

2.1 Bayesian Inference

2.1.1 Introduction of Bayesian Inference

By using Bayes' theorem, the statistical reasoning method known as "Bayesian inference" updates a hypothesis's probability based on fresh information or evidence. At its core, Bayesian reasoning estimates a posterior probability in the form of a prior distribution by utilizing previous knowledge. Bayesian inference is an important technique in mathematics. Bayesian inference of updating information plays an important role in data series detection. In addition, Bayesian inference has applications in many fields due to its predictability, including scientific research, engineering, medical practice, sports, legal research, and philosophy. The formulation of Bayes' Theorem is [3]:

$$P(\theta \mid X) = \frac{P(X \mid \theta) P(\theta)}{P(X)} \tag{1}$$

Where $P(\theta \mid X)$ is the posterior probability, which shows how likely the parameters are based on the data that have been observed. $P(X \mid \theta)$ stands for probability or likelihood of the data given the parameters. $P(\theta)$ is the prior probability, the first impression of the parameters prior to looking at the data. $P(X)$ is the marginal likelihood, the total probability of the observed data.

2.1.2 Discrete Random Variable Case

When random variables in the hypothesis are discrete, the prior distribution is expressed as a discrete probability distribution. In this condition, the formula of calculating the posterior probability changes to summing the overall possible discrete states [4]:

$$\pi(\theta \mid x) = \frac{p(x \mid \theta_i)\pi(\theta_i)}{\sum_j p(x \mid \theta_j)\pi(\theta_j)} \tag{2}$$

For this equation, the parameter θ is discrete. Thus, by using the observed data and the previous distribution, one can still compute the posterior distribution.

2.1.3 Calculation of Posterior Distribution

The probability function and the prior distribution are combined to get the posterior distribution. The integral of the joint distribution over all possible values of θ is used to calculate the marginal density function, or m(x):

$$m(x) = \int p(x \mid \theta)\pi(\theta) d\theta \tag{3}$$

The posterior distribution is then:

$$\pi(\theta|x) = \frac{p(x \mid \theta)\pi(\theta)}{m(x)} \propto p(x \mid \theta)\pi(\theta) \tag{4}$$

This approach allows for the computation of the posterior distribution, even when θ is continuous, by integrating over all possible parameter values [5].

2.1.4 Selection of Prior Distributions

One important component of Bayesian inference is the selection of the prior distribution. In situations when previous data regarding the parameter θ is not accessible, one may choose a non-informative prior, like a uniform distribution:

$$\pi(\theta) = \begin{cases} c, \theta \in \Theta \\ 0, \theta \notin \Theta \end{cases} \tag{5}$$

where Θ is the parameter space, and c is a constant. In order to maximize the impact of the facts on shaping the posterior belief, non-informative priors are selected so as to have little effect on the posterior distribution.

2.2 Markov Chain Monte Carlo algorithm

2.2.1 Introduction to Markov Chain Monte Carlo algorithm

MCMC, Markov Chain Monte Carlo algorithm, is a dynamic method used to deal with the slow speed caused by the extensive computational complexity of static Monte Carlo algorithms based on Markov Chain and Bayesian theory. The Monte Carlo algorithms are widely used for

ISSN 2959-6157

numerical estimation in Bayesian inference. This technique is used to sample at random in the approach for estimating numerical outcomes; it is then considered important for obtaining solutions for complex mathematical integral calculations. So they allow for the approximation of integrals that are otherwise difficult to solve [6].

2.2.2 Introduction to Importance Sampling

While Monte Carlo algorithms typically rely on uniform distribution to obtain results such as integrals and areas, there are many cases where sampling from other distributions, such as normal or Poisson distributions, is more appropriate. This is where importance sampling comes into play. A weighted average of random drawings from a different distribution is used to approximate a mathematical expectation with regard to a target distribution in a group of Monte Carlo techniques known as importance sampling. Importance sampling, together with MCMC, serves as a basis for simulation-based methods of numerical integration."

2.2.3 Gibbs Sampling

Gibbs sampling is a special form of the Metropolis-Hastings algorithm. Due to its simplicity and ease of implementation, it is also a widely used MCMC algorithm. Gibbs sampling is primarily used for sampling and estimation of joint distributions of multivariate variables. It defines the full conditional distributions from the joint distribution, and for a given variable dimension, it fixes the other dimensions and uses the full conditional distribution for sampling, sequentially obtaining the iterative values for each dimension of the variables [7].

The process involves the following steps.

1. Initialize the state $X^0 = (x_1^0, x_2^0, ..., x_m^0)$

2. Set $t = 0$

3. For each variable x_i, sample from the conditional distribution given the other variables:

$p(x_i^{(t+1)} \mid x_i^{(t+1)}..., x_{i-1}^{(t+1)}, x_{i+1}^{t}, ..., x_m^{t})$

4. Increment t and repeat the process until convergence.

Gibbs Sampling offers several advantages, making it a powerful tool in MCMC methods. One of its benefits is its high convergence speed, as the acceptance probability is always 1, eliminating the need for rejections. Additionally, it is particularly effective in handling high-dimensional data, as it can efficiently explore complex probability spaces. Furthermore, by sequentially updating one variable at a time while keeping others constant, Gibbs Sampling simplifies the sampling process, reducing computational complexity and improving efficiency.

3. Applications

3.1 A basic example of Bayesian Inference

Given a disease with an incidence rate of 0.2%. If a person has the disease, the test accuracy is 90% (meaning there is a 10% chance of not detecting a positive case). If a person does not have the disease, the false positive rate is 5% (meaning there is a 5% chance of incorrectly reporting a positive case). What is the likelihood that a person reporting a positive case actually has the illness?

Let H represent having the disease, and E represent testing positive. Given that a test result is positive, one wishes to determine the probability, or P(H|E), that an individual has the illness. The computation is made using the Bayes theory and looks like this:

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)} \tag{6}$$

The likelihood of being ill is represented by P(H), where P(H) = 0.2%. P(E|H) = 90% is the likelihood of testing positive provided that an individual has the disease. P(E) represents the probability of testing positive.

To calculate P(E), one can scale up the number of people tested, for example, to 100,000. Based on the incidence rate. There are 200 patients and 99,800 healthy people. Among the 200 patients, 180 will test positive, while 20 will not be detected (negative). Among the 99,800 healthy people, 5% (4990 people) will test positive (false positives), while 95% (94,810 people) will test negative. Therefore,

$$P(E) = \frac{180 + 4990}{100000} = 0.0517 = 5.17\% \tag{7}$$

Enter the following into the Bayes theorem to get P(H|E)

$$P(H \mid E) = \frac{90\% \times 0.2\%}{5.17\%} = 3.482\% \tag{8}$$

The result is that the probability of disease is 3.543%, which is far less than 90%, and contrary to most people's intuition, the reason is that the number of false positives caused by a large number of healthy people is far more than that of patients, when there is evidence of "positive test", the probability of disease is increased from 0.2% to 3.543%, which is far from enough to confirm the diagnosis.

3.1.2 Further representation

In the above calculation, one finds that calculating P(E) is relatively difficult, and in many cases, it may even be impossible to know P(E). In this situation, one needs to use another representation of Bayes' theorem.

Let P(H) denote the probability of H occurring, and $\bar{H}$

represent the event of H not occurring, with P($\bar{H}$) denoting the probability of H not occurring. Obviously, $P\left(\bar{H}\right)=1-P(H)$. It is apparent that P(E) can be divided into two parts: the point where E and H converge is one part, and the junction of E and $\bar{H}$ is the other part. Thus, $P(E)=P(E\cap H)+P\left(E\cap\bar{H}\right)$. According to the previous formula $P(A\cap B)=P((A|B)\times P(B))$ substituting in, one can obtain

$$P(E)=P\left(E\cap\bar{H}\right)+P(E\cap H)=P(E|H)\times P(H)+P\left(E|\bar{H}\right)\times P\left(\bar{H}\right) \quad (9)$$

Another version of the Bayes theorem can be obtained by substituting P(E)

$$P(H|E)=\frac{P(E|H)\times P(H)}{P(E|H)\times P(H)+P\left(E|\bar{H}\right)\times P\left(\bar{H}\right)} \quad (10)$$

Using this formula, the author does not need to calculate P(E). Returning to the initial problem. The likelihood of having the illness is represented by P(H), which is P(H) = 0.2%. P(E|H) = 90% is the likelihood of testing positive provided that an individual has the disease. P(E|$\bar{H}$) indicates the five percent chance of testing positive in the absence of the illness. P($\bar{H}$) demonstrates the probability of not having the disease = 1 - P(H) = 99.8%. Substitute these values into the formula to calculate

$$P(H|E)=\frac{P(E|H)\times P(H)}{P(E|H)\times P(H)+P\left(E|\bar{H}\right)\times P\left(\bar{H}\right)} \quad (11)$$
$$=\frac{90\%\times 0.2\%}{90\%\times 0.2\%+5\%\times 99.8\%}=3.482\%$$

The likelihood that the patient has the illness rises from 0.1% to 4.721% with this positive test result. Assume this individual administers another test and receives a positive result once more. What is the likelihood that they currently have the illness?

The author will still use Bayes' theorem to calculate, but the prior probability $P(H)$ is no longer 0.2%; it is now 3.482%, while $P(E|H)$ and $P\left(E|\bar{H}\right)$ remain unchanged. Calculate the new $P(H|E)$

$$P(H|E)=\frac{90\%\times 3.482\%}{90\%\times 3.482\%+5\%\times(1-3.482\%)} \quad (12)$$
$$=39.37\%$$

The result is 39.37%. Two positive test results increase the prior probability from 0.1% to 3.482%, and then to 39.37%. As seen, the core idea of Bayes' theorem is to continually adjust the prior probability to the posterior probability based on new evidence, bringing it closer to the objective truth.

3.2 Alzheimer Disease Prediction

In this case, the author applies the algorithm to the investigate the Alzheimer patients, where age is the variable. First, the author selected two datasets of characteristics associated with Alzheimer's disease; the first was used to determine the regression model's prior distribution of the parameters, and the second served as an analytical data set. Second, from the first data set, the author randomly samples 60% of the data in a cyclic manner. Each time people apply traditional maximum likelihood method to fit the intercept and slope of the regression model and record it in an array. Then one plotted it and found that it's approximately conformed to normal distribution, so one used the built-in function in Python to fit it into the normal and obtain the mean and variance, which is the hyper-parameters one needs for the algorithm. Then, for the second data set, people divide it into to part to be used as the analytical data. 70% as the test set while 30% as the training set.

·The theoretical induction is the following. 1) Assume that two parameters in the logistic regression model are independent and Identically distributed as follows β N ($\bar{\beta}$, $|\Sigma|^2$)and α N($\bar{\alpha}$, σ^2). 2) Then one will have the probability density function $p(\beta)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}e^{-\frac{1}{2}\hat{\beta}^T\hat{\sum}^{-1}\hat{\beta}}$ and $p(\alpha)=\frac{1}{\sqrt{(2\pi)\sigma^2}}e^{\frac{(\alpha-\mu)^2}{2\sigma^2}}$. 3) If the author substitutes the previous function into this nonuniform one, one can get

$$p(\beta,\alpha|Y)\propto p(\beta)\,p(\alpha)\,p(Y|\beta,\alpha)= (2\pi)^{-\frac{d+1}{2}}e^{\frac{(\alpha-\mu)^2}{2\sigma^2}-\frac{1}{2}\hat{\beta}^T\hat{\sum}^{-1}\hat{\beta}}\prod_{i=1}^{n}\frac{e^{-z_i(1-y_i)}}{1+e^{-z^i}} \quad (13)$$

·The algorithm based on Gibbs sampling is the following. The author samples α_0 and β_0 from the prior distribution. According to previous induction, one can obtain the conditional probability

ISSN 2959-6157

$$P(\alpha|\beta,Y) \propto e^{-\frac{(\alpha-\mu)^2}{2\sigma_2^2}} \prod_{i=1}^{n} \frac{(e^{-(\beta x_i+\alpha)1-y_i}}{1+e^{-(\beta x_i+\alpha)}} \quad (14)$$

The conditional probability for β has the same form. Thus one can sample alpha and beta from their distribution and cyclically update them. Repeat the sampling process until convergence. From the first data set, one can draw a 3D diagram to show the distribution of the intercept and the slope. And people have the hyper-parameters as follows: $\mu_{coeff} = 0.0135$, $\mu_{intercept} = 0.7787$, $\sigma_{coeff} = 0.0102$, and $\sigma_{intercept} = 0.7899$. After applying the regression model and the algorithm to the second data set and the value above, one has the results as follows: Intercept = -0.3694 and Coefficient= -0.0037. After using the test set to test the result, the final accuracy rate is about 62%. The probable reason for the accuracy not being ideal is that people only used age to be the independent variable of machine learning. If people can obtain more other kinds of data in the future, the accuracy rate might be higher.

4. Conclusion

This study introduces the concepts of Bayesian inference and Markov Chain Monte Carlo (MCMC) methods, and explored the application of Bayesian inference and MCMC methods for disease prediction, focusing on a general prediction of having a disease and Alzheimer's disease. The two applications demonstrate the core idea of Bayes' theorem—the author can achieve a more accurate result by calculating the posterior probability infinitely as people gain more information. Additionally, by utilizing two datasets and employing the Gibbs sampling algorithm to estimate the regression model parameters, the accuracy of the model achieved 62%. These results demonstrate the potential of gaining a more accurate conclusion of complex medical cases by combining Bayesian inference and MCMC methods together. However, improvements can be made. Since the current model relies on a single variable (age), the predictive capacity is limited, and the final accuracy rate is about 62%, which is not an ideal value. In order to enhance the accuracy, future research should contain more patient data, such as genetic, environmental, and lifestyle factors. Additionally, future works can also include predictions of other disease based on this approach, or integrate machine learning techniques to create a more comprehensive models that might improve the strengths of both methodologies.

References

[1] Janda, T., Šejnoha, M., & Šejnoha, J. Applying Bayesian approach to predict deformations during tunnel construction. International Journal for Numerical and Analytical Methods in Geomechanics, 2018, 42(15): 1765-1784.

[2] Bekele, R., & Menzel, W. A bayesian approach to predict performance of a student (bapps): A case with ethiopian students. algorithms, 2005, 22(23): 1-6.

[3] Kaderali, L., Zander, T., Faigle, U., Wolf, J., Schultze, J. L., & Schrader, R. CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. Bioinformatics, 2006, 22(12): 1495-1502.

[4] Herschtal, A., Martin, R. F., Leong, T., Lobachevsky, P., & Martin, O. A. A bayesian approach for prediction of patient radiosensitivity. International Journal of Radiation Oncology Biology Physics, 2018, 102(3): 627-634.

[5] Goodarzi, M., Elkotb, M. A., Alanazi, A. K., Abo-Dief, H. M., Mansir, I. B., Tirth, V., & Gamaoun, F. Applying Bayesian Markov chain Monte Carlo (MCMC) modeling to predict the melting behavior of phase change materials. Journal of Energy Storage, 2022, 45: 103570.

[6] van de Schoot R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G. Bayesian statistics and modelling. Nature Reviews Methods Primers, 2021, 1(1): 1-13.

[7] Kruschke, J. K. Bayesian analysis reporting guidelines. Nature human behaviour, 2021, 5(10): 1282-1291.