The Use of Bayesian Techniques with Binary and Vector Data

Shaymaa Riyadh Thanoon

[1] Department Basic Sciences, College of Nursing, Mosul University, Nineveh, Iraq.

Email: shaymaa.riadh@uomosul.edu.iq

Abstract

This research provides a conceptual framework and examples for applying Bayesian techniques to binary and vector data. For the binary data, for observations take on one of two possible values, Bayesian logistic regression and Bayesian networks are techniques, applicable Bayesian logistic regression places priors on the coefficients and derives the posterior using the likelihoods under a logistic model. Bayesian networks represent dependencies between binary variables graphically and perform inference using conditional probability tables. For vector data, where observations are multi-dimensional, Bayesian linear regression places priors on the regression coefficients and finds posterior using the likelihoods under linear model. Gaussian process regression models the relationship between inputs and outputs as a draw from a Gaussian process prior and computes the posterior process given observed data. The research provides the conceptual framework underlying Bayesian analysis, including key concepts such as prior and posterior distributions. It highlights the advantages of Bayesian methods like the ability to incorporate domain knowledge and model uncertainty. Numerical examples demonstrate how Bayesian techniques can be applied to binary and vector data classification tasks. The abstract summarizes the core ideas and contributions of the research on this topic.

Keywords*:* Binary data; Gaussian process; Logistic regression; Vector data

1. Introduction

Bayes methods are a group of statistical methods that use Bayes' theory to construct a probability distribution for the parameter x based on our previous information about this parameter. In Bayesian methods, probabilities are interpreted as degrees of belief, and the goal is to infer the most likely hypothesis given the available data.

There is a wide range of problems to which Bayesian methods can be applied (Albert, 2009) including classification problems with binary and vector data. In the case of binary data, each observation is represented by a vector of binary values, where each value corresponds to the presence or absence of a particular feature. Bayesian methods can be used to model the probability of each class given the observed binary values and to make predictions for new observations.

In the case of vector data (Gelman et al., 2013; Gelman & Hill, 2006) each observation is represented by a vector of continuous or discrete values, where each value corresponds to a particular feature. Bayesian methods can be used to model the probability of each class given the observed vector values and to make predictions for new observations.

Bayesian methods have several advantages over other statistical techniques, including the ability to incorporate prior knowledge, to handle missing data, and quantify uncertainty. However, Bayesian methods can also be computationally intensive and may require the specification of prior distributions, which can be subjective and difficult to choose.

Overall, Bayesian methods are a powerful and flexible class of statistical techniques that can be used to solve a large number of problems, for example, classification problems with binary and vector data .

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

2. **Research Method**

In many cases (Barber, 2012) we have additional information from our previous experiences about parameter θ . e may notice that it takes different values and that there is evidence that θchanges and that this change and additional information can be represented by a probability distribution $\pi(\theta)$ for parameterθ. That is, θbecomes a random variable with a probability distribution $\pi(\theta)$.

Thus, the difference between the traditional statistical methods and the Bayesian method (McElreath, 2020) is that parameter θ is considered a random variable that has a probability distribution $\pi(\theta)$ that expresses our previous information about parameter θ and describes the degree of our belief in the possible values of this parameter or describes our previous experience about the parameter before obtaining the sample, and accordingly this distribution is called prior distribution: It describes the information and past experience that we have about parameter θ.

In Bayesian analysis, the prior distribution (McElreath, 2020) $\pi(\theta)$ for parameter θ represents the initial knowledge or beliefs about the parameters of interest before observing any data. It quantifies our uncertainty about the parameters before incorporating the information from the data.

The prior distribution is specified based on prior knowledge, previous studies, expert opinions, or subjective beliefs. It can take various forms, such as a normal distribution, uniform distribution, beta distribution, or any other probability distribution that is appropriate for the parameter being modeled.

The choice of the prior distribution (Gelman et al., 2013; Gelman & Hill, 2006) can have a significant impact on the posterior distribution and subsequent inferences. A prior can be informative, where it assigns a relatively higher probability to certain values of the parameter based on strong prior knowledge, or it can be non-informative, where it assigns a relatively equal probability to a wide range of values. Non-informative priors are often used when there is limited prior knowledge or when we want the data to dominate the inference.

One of the advantages of the Bayesian approach is that it allows for the iterative update of the prior distribution as more data becomes available. This is particularly useful when dealing with sequential or streaming data.

It's important to note that the choice of prior distribution can be subjective, and different individuals may have different prior beliefs. Sensitivity analysis or robustness checks can be performed to assess the impact of different prior specifications on the results.

2.1 Posterior distribution

the posterior distribution (McElreath, 2020) is the updated probability distribution of the parameters of interest after incorporating the information contained in the observed data.

Mathematically, the posterior distribution (Robert, 2007) is calculated as:

Posterior distribution ∝ Prior distribution × Likelihood function

$$\pi(\theta/x) \propto \pi(\theta)l(\theta/x). \quad (1)$$

The posterior distribution represents our updated beliefs about the parameters given the observed data. It provides a complete probability distribution that reflects the uncertainty in the parameter estimates.

2.2 Binary data

Binary data (Hosmer Jr et al., 2013) refers to a type of categorical data where each observation can take one of two possible outcomes or categories. These outcomes are typically represented as 0 and 1, or as "success" and "failure," "yes" and "no," or any other appropriate labels.

Binary data is commonly encountered in various fields, including biology, social sciences, finance, and machine learning.

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

When analyzing binary data, various statistical methods can be employed. Some commonly used techniques include:

Proportions and percentages: Calculating the proportion or percentage of observations falling into each category.

Chi-square test: Assessing the independence or association between two categorical variables.

Logistic regression: Modeling the relationship between binary response variables and predictor variables.

Odds ratio: Measuring the odds of an event occurring in one category compared to another.

Binomial distribution: Modeling the probability of observing a specific number of successes in a fixed number of trials.

Bayesian analysis can also be applied to binary data, where prior distributions, likelihood functions, and posterior distributions are used to estimate parameters and make inferences.

2.3 Vector data

Vector data (Bivand et al., 2008) is a type of spatial data representation that uses points, lines, and polygons to represent geographic features. It is commonly used in Geographic Information Systems (GIS) and spatial analysis. Vector data represents real-world features by defining their geometry and attributes.

Here are the main components of vector data:

Points: Points represent individual locations or specific features with a single set of coordinates. They are often used to represent landmarks, cities, or sampling locations. Each point can have associated attributes such as a name, population, or temperature.

Lines: Lines represent linear features, such as roads, rivers, or pipelines. They are composed of a series of connected points. Lines can have attributes associated with them, such as road type, length, or speed limits (Wan et al., 2013).

.

Polygons: Polygons represent areas or regions. They are enclosed by a series of connected lines, forming a closed shape. Examples of polygons include countries, parks, or administrative boundaries. Polygons can also have associated attributes, such as area, population density, or land use, (Bardenet et al., 2017).

Bayesian methods can be used with both binary and vector data. Here are some examples of how Bayesian methods can be applied to analyze binary and vector data, (Rana et al., 2016):

3. Results And Discussions

To present and discuss the results, both methods with binary and vector directions and coordinates must be presented and both methods must be discussed in detail as follows:

3.1. Bayesian Methods With Binary Data

Logistic regression is a popular method for modeling binary data. In Bayesian logistic regression, a prior distribution is placed on the coefficients of the logistic regression model, and the posterior distribution is obtained using Bayes' theorem. The posterior distribution can be used to make predictions and perform hypothesis testing.

The equation for logistic regression can be represented as (Maignan & Scott, 2016)

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta n x n \quad (2)$$

where:

logit (p): represents the logarithm of the odds of the binary outcome.

p : the probability of the binary outcome.

$\beta_0, \beta_1, \beta_2, \ldots, \beta n$: are the coefficients or parameters associated with each independent variable.

$x_1, x_2, \ldots, xn$:are the values of the independent variables.

And a Bayesian network is a graphical model that represents the probabilistic relationships among a set of variables. Bayesian networks can be used to model binary data by representing each binary variable as a node in the network and specifying conditional probabilities between the nodes (Koller & Friedman, 2009).

The equations of Bayesian networks are based on the principles of conditional probability and the chain rule of probability. A Bayesian network is a graphical model that represents the probabilistic relationships among a set of variables. The equations associated with Bayesian networks include the joint probability distribution, the conditional probability tables (CPTs), and the inference equations (Jensen & Nielsen, 2007).

(a) The joint probability distribution: The joint probability distribution of a Bayesian network is defined as the product of the conditional probabilities of each variable given its parents in the network. Mathematically, it can be expressed as:

$$P(X_1, X_2, \ldots, Xn) = P(X_1 \mid \text{Parents}(X_1)) * P(X_2 \mid \text{Parents}(X_2)) * \ldots * P(Xn \mid \text{Parents}(Xn)) \quad (3)$$

This equation represents the factorization property of Bayesian networks, where each variable's probability is conditioned on its parents.

(b) Conditional probability tables (CPTs): CPTs are used to represent the conditional probabilities of each variable given its parents. Each entry in the CPT specifies the probability distribution of a variable given the possible combinations of states of its parents. The CPTs provide the necessary information for updating probabilities and performing inference in the network (Levin et al., 2011).

(c) Inference equations: In Bayesian networks, inference involves calculating the posterior probabilities of variables given observed evidence. The most common inference task is computing the probability of a specific variable given evidence on other variables. This can be done using Bayesian inference, which employs the joint probability distribution and the evidence to calculate the desired posterior probability (Jeliazkov & Rahman, 2013).

$$P(X \mid E) = \alpha * P(X, E) = \alpha * \Sigma_u P(X, u, E) \quad (4)$$

Where:

X: represents the target variable,

E: represents the evidence,

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

α : the normalization constant and u represents the unobserved or hidden variables.

3.2 Bayesian Methods With Vector Data

Bayesian linear regression is similar to Bayesian logistic regression, but is used to model continuous outcomes instead of binary outcomes. In this method, a prior distribution is placed on the coefficients of the linear regression model, and the posterior distribution is obtained using Bayes' theorem (Gelman et al., 2013).

The equation for Bayesian linear regression can be written as (Bishop & Nasrabadi, 2006):

$y = X\beta + \varepsilon$ (5)

where:

y is the vector of observed dependent variable values

X : the matrix of independent variables, also known as the design matrix.

β : the vector of unknown coefficients or parameters that represent the relationship between the independent variables and the dependent variable

ε : the vector of random errors or noise term,

In Bayesian linear regression, the prior distribution is specified for the coefficients β. Typically, a normal distribution or a multivariate normal distribution is used as the prior distribution for β.

And Gaussian process regression (Murphy, 2012) is a non-parametric method that can be used to model vector data. In this method, a prior distribution is placed on the function that maps the input vectors to the output vectors, and the posterior distribution is obtained using Bayes' theorem. Gaussian process regression is often used in machine learning applications where the relationship between the input and output vectors is complex and unknown, (Mező & Baricz, 2017).

In Gaussian Process Regression (GPR), the relationship between the input variables (often denoted as X) and the output variable (often denoted as y) is modeled as a draw from a Gaussian process. The basic equation for Gaussian process regression can be written as follows (Bishop & Nasrabadi, 2006):

$y = f(X) + \varepsilon$ (6)

where:

y : the vector of observed output variable values.

X : the matrix of input variables.

f(X) : the unknown underlying function that represents the relationship between the input variables and the output variable. It is modeled as a draw from a Gaussian process.

ε : the vector of random errors or noise term.

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

To fully specify the Gaussian process, we need to define its mean function and covariance function (kernel). The mean function represents the expected value of the underlying function f(X) for a given input, while the covariance function determines the similarity or correlation between different input-output pairs.

Given a set of observed input-output pairs (X, y), the goal of Gaussian process regression is to estimate the distribution of the output variable y for new, unseen inputs. This is achieved by computing the posterior distribution of the Gaussian process conditioned on the observed data. The posterior distribution combines the prior distribution of the Gaussian process with the likelihood of the observed data, (Briceño-Arias et al., 2019).

To make predictions for new inputs, Gaussian process regression uses the posterior distribution to estimate the mean and variance of the output variable at each input point. The mean prediction represents the expected value of the output variable, while the variance provides a measure of uncertainty or confidence in the prediction (Bach et al., 2012).

In summary, the equation for Gaussian process regression is a combination of the underlying function f(X) and a noise term ε, where f(X) is modeled as a draw from a Gaussian process. The specific form of the Gaussian process and the choice of mean function and covariance function (kernel) depend on the problem at hand and the assumptions made about the underlying relationship between the input and output variables (Perrone & Favaro, 2015).

In summary, Bayesian methods can be used with binary and vector data. According to the type of problem and the nature of the data, the appropriate method is chosen (Lichman, 2013).

3.3 **Numerical Cases**

To illustrate the concept of Bayes with binary data suppose we have a dataset of 1000 emails, where each email is labeled as either spam (1) or not spam (0). We want to use Bayes' theorem to classify new emails as spam or not spam based on their content. We can model each email as a binary vector, where the i-th element of the vector is 1 if the i-th word appears in the email and 0 otherwise. We can then use the following steps:

Calculate the prior probabilities of spam and not-spam emails in the dataset. Let's say that 100 of the 1000 emails are labeled as spam, so the prior probability of spam is P(spam) = 0.1, and the prior probability of not spam is P(not spam) = 0.9.

Calculate the likelihood probabilities of each word given spam and not spam emails. For example, suppose that the word "free" appears in 50 of the 100 spam emails, and in 5 of the 900 not spam emails. Then the likelihood probability of "free" given spam is P("free "|spam) = 0.5, and the likelihood probability of "free" given not spam is P("free "|not spam) = 0.005.

Given a new email, calculate the posterior probabilities of spam and not spam using Bayes' theorem. For example, suppose that the new email contains the words "free" and "buy". We can calculate the posterior probability of spam as follows:

P(spam "free", "buy") = P("free", "buy "|spam) * P(spam) / P("free", "buy")

= P("free "|spam) * P("buy "|spam) * P(spam) / P("free", "buy")

= 0.5 * 0.2 * 0.1 / P("free", "buy")

Similarly, we can calculate the posterior probability of not spam as follows:

P(not spam "free", "buy") = P("free", "buy "|not spam) * P(not spam) / P("free", "buy")

= P("free "|not spam) * P("buy "|not spam) * P(not spam) / P("free", "buy")

= 0.005 * 0.01 * 0.9 / P("free", "buy")

The denominator P("free", "buy") is the probability of observing the words "free" and "buy" in any email, and can be calculated as follows:

P("free", "buy") = P("free", "buy "|spam) * P(spam) + P("free", "buy "|not spam) * P(not spam)

= P("free "|spam) * P("buy "|spam) * P(spam) + P("free "|not spam) * P("buy "|not spam) * P(not spam)

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

= 0.5 * 0.2 * 0.1 + 0.005 * 0.01 * 0.9

Once we have calculated the posterior probabilities of spam and not spam, we can classify the new email as spam if P(spam “free", "buy") > P(not spam “free", "buy"), and as not spam otherwise.

And the Bayesian concept with vector data suppose we have a dataset of 1000 images, where each image is represented as a 28x28 pixel grayscale matrix, and labeled as a digit from 0 to 9. We want to use Bayes' theorem to classify new images based on their pixel values. We can model each image as a 784-dimensional vector, where each element corresponds to a pixel value between 0 and 255. We can then use the following steps:

Calculate the prior probabilities of each digit in the dataset. Let's say there are 100 images of each digit, so the prior probability of each digit is P(digit) = 0.1.

Calculate the likelihood probabilities of each pixel value given each digit. For example, suppose that the pixel at location (i, j) has value 128 in 20 of the images of digit 3, and in 10 of the images of digit 8. Then the likelihood probability of pixel (I,j) having value 128 given digit 3 is P(pixel(i,j)=128|digit=3) = 0.2, and the likelihood probability of pixel (i, j) having value 128 given digit 8 is

P(pixel(i,j)=128|digit=8) = 0.1.

Given a new image, calculate the posterior probabilities of each digit using Bayes' theorem. For example, suppose that the new image has pixel values as follows:

pixel(1,1) = 100, pixel(1,2) = 200, ..., pixel(28,28) = 50

We can calculate the posterior probability of digit 3 as follows:

P(digit=3|image) = P(image|digit=3) * P(digit=3) / P(image)

where P(image digit=3) is the likelihood probability of the image given digit 3, and P(image) is the probability of the image appearing in any digit category. We can calculate the likelihood probability as follows:

P(image digit=3) = P(pixel(1,1)=100|digit=3) * P(pixel(1,2)=200|digit=3) * ... * P(pixel(28,28)=50|digit=3)

Similarly, we can calculate the posterior probability of each digit, and classify the new image as the digit with the highest posterior probability.

In summary, Bayesian methods can be used with binary and vector data. The choice of method depends on the specific problem and the nature of the data.

4. Conclusion

In conclusion, Bayesian methods are a powerful tool for classification problems with binary and vector data. These methods allow us to incorporate prior knowledge, handle missing data, and quantify uncertainty in our predictions. For binary data, Bayesian methods can be used to model the probability of each class given the observed binary values and make predictions for new observations. For vector data, Bayesian methods can be used to model the probability of each class given the observed vector values and make predictions for new observations.

One of the main advantages of Bayesian methods is their ability to incorporate prior knowledge. This can be particularly useful in cases where we have some prior information about the problem, such as the expected distribution of the features or the prevalence of the classes. By incorporating this prior.

References

[1] Albert, J. (2009). *Bayesian computation with R*. Springer.
[2] Bach, F., Jenatton, R., Mairal, J., & Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, *4*(1), 1–106.
[3] Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
[4] Bardenet, R., Doucet, A., & Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, *18*(47).

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023

[5] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, Issue 4). Springer.
[6] Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717). Springer.
[7] Briceño-Arias, L. M., Chierchia, G., Chouzenoux, E., & Pesquet, J.-C. (2019). A random block-coordinate Douglas–Rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, *72*, 707–726.
[8] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.
[9] Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
[10] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.
[11] Jeliazkov, I., & Rahman, M. A. (2013). Binary and ordinal data analysis in economics: Modeling and estimation. *Mathematical Modeling with Multidisciplinary Applications*, 123–150.
[12] Jensen, F. V, & Nielsen, T. D. (2007). *Bayesian networks and decision graphs* (Vol. 2). Springer.
[13] Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
[14] Levin, A., Weiss, Y., Durand, F., & Freeman, W. T. (2011). Efficient marginal likelihood optimization in blind deconvolution. *CVPR 2011*, 2657–2664.
[15] Lichman, M. (2013). *UCI machine learning repository*. Irvine, CA, USA.
[16] Maignan, A., & Scott, T. C. (2016). Fleshing out the generalized Lambert W function. *ACM Communications in Computer Algebra*, *50*(2), 45–60.
[17] McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.
[18] Mező, I., & Baricz, Á. (2017). On the generalization of the Lambert *W* function. *Transactions of the American Mathematical Society*, *369*(11), 7917–7934.
[19] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
[20] Perrone, D., & Favaro, P. (2015). A clearer picture of total variation blind deconvolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(6), 1041–1055.
[21] Rana, R., Jain, A., Shankar, A., Bednarek, D. R., & Rudin, S. (2016). Scatter estimation and removal of anti-scatter grid-line artifacts from anthropomorphic head phantom images taken with a high resolution image detector. *Medical Imaging 2016: Physics of Medical Imaging*, *9783*, 1619–1628.
[22] Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation* (Vol. 2). Springer.
[23] Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. (2013). Regularization of neural networks using dropconnect. *International Conference on Machine Learning*, 1058–1066.

Doi: https://doi.org/10.54216/GJMSA.0100104
Received: May 20, 2023 Revised: June 25, 2023 Accepted: July 30, 2023