STAT 536D Project

Violation of Conditional Independence Assumption

Zhipeng Zhu

Introduction

Conditional independence is defined that the occurrence of A and the occurrence of B are independent in their conditional probability distribution given some event C. That means, given the information of C, the observation of A will provide no information about the occurrence of B, and the observation of B will not reveal anything about A neither. [1]

Generally, the conditional independence assumption is the base of Bayesian models. However, real cases differ in that the input variables may not meet such an assumption, or the conditional independence is weak. We would like to check the potential damage to the model when the input variable can not hold the conditional independence assumption well.

Methodology

The generative model uses conditional independence to simplify the calculation and estimation. For a specific input data, it assumes the variables are conditionally independent given the observations of the response. When the variables indeed have conditional independence with each other, the model should have good outcomes. We can then modify the data and make the input variables a little correlated with the observations and train the model under default assumptions. That is, we will use a dataset violating the conditional independence assumption and check the outcome of such a mistaken model.

The data will include three/four variables listing in Table 1.

Table 1 Summary of Data

Variables	Generating method
Y	Sample from 2500 Y=1 and 2500 Y=0
X	Binomial (1, p) $p = (1-y) * \gamma_{0,tr} + y * \gamma_{1,tr}$ $\gamma_{0,tr} = 0.1$ and $\gamma_{1,tr} = expit(logit(\gamma_{0,tr}) + log(2)$
Xstr1	Binomial (1, p) $p = \big(1-(\theta * y + \theta * x)\big) * (1-0.85) + (\theta * y + \theta * x) * 0.75$ θ determines the contribution of Y on Xstr1
Xstr2	Binomial (1, p) $p = \big(1-(\theta * y + \theta * x)\big) * (1-0.95) + (\theta * y + \theta * x) * 0.60$ θ determines the contribution of Y on Xstr2

The generative model is constructed as

$$f(y, x, x_1^*, x_2^*)$$

$$\begin{aligned} &= f(y)f(x|y)f(x_1^*, x_2^*|x, y) \\ &= f(y)f(x|y)f(x_1^*, x_2^*|x) \\ &= f(x|y)f(x_1^*|x)f(x_2^*|x) \end{aligned}$$

The prior distributions of the model are assumed to follow uniform distribution, list as the following:

$$\gamma_0 \sim unif(0,1)$$

$$\gamma_1 \sim unif(0,1)$$

$$Sn_1 \sim unif(0.5,1)$$

$$Sp_1 \sim unif(0.5,1)$$

$$Sn_2 \sim unif(0.5,1)$$

$$Sp_2 \sim unif(0.5,1)$$

$$trgt = logit(\gamma_1) - logit(\gamma_0)$$

Three models are conducted in the report with different values of θ, which changes the correlation between Xstr and Y. We use the $\theta = 0$ as the control and assume it has met the conditional independence assumption. Then two more sets of data are generated with $\theta = 0.1$ and $\theta = 0.6$.

With the different values of θ, the estimations of the three models will absolutely be different. Hence, it would be better to evaluate the results by the estimator, standard error, Rhat and number of effective points to figure out the potential damage resulted from the violation of assumption.

Results and Comparison

The results of “trgt” in three models are summarized in Table 2.

Table 2 Summary of "trgt" in models

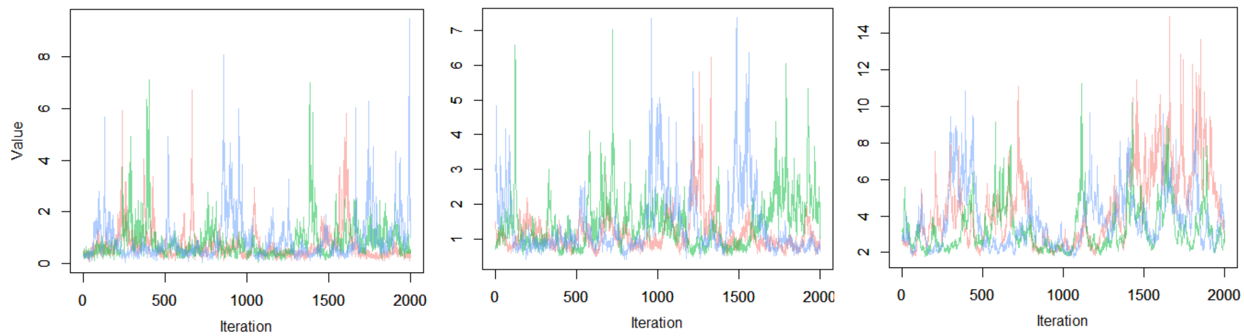
Model	Mean	Standard Error	Rhat	n.eff
$\theta = 0$	0.8396979	0.73888510	1.06	185
$\theta = 0.1$	1.3387087	0.76723081	1.09	104
$\theta = 0.6$	3.7973914	1.72982031	1.24	71

The estimations of mean value deviate a lot after the contribution of Y to Xstr increases, and the standard errors become more considerable. The difference in estimation values is resulted from the changes of observations, and such changes will obviously provide different posterior.

Meanwhile, the more important indexes, the Rhat and the n.eff, show that the highly correlated data results in the worst performance in convergence.

The performance of the models can also be accessed through the trace plots in Figure 1.

Figure 1 Trace of "trgt" in models

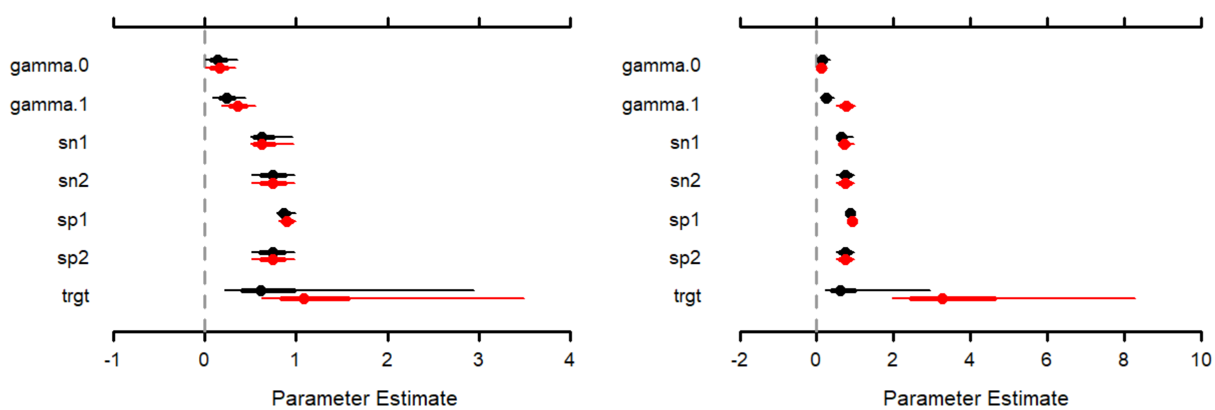


(a) M1: $\theta = 0$ (b) M2: $\theta = 0.1$ (c) M3: $\theta = 0.6$

According to the trace plots, all three models have some outstanding errors in literation. However, as the contribution of Y increases, the deviation of estimated values becomes more notable. This is corresponding to the Rhat reported in the model summary. While the first two models get Rhats close to 1, the third one is rated as 1.24, which indicates a terrible convergence.

Besides the estimation of “trgt”, the estimations of the other parameters can also reveal some differences between models. The confidence interval of each estimated parameters is summarized in Figure 2, with the first model as the reference.

Figure 2 Comparison of estimations betweem models



(a) M1 vs. M2 (b) M1 vs. M3

As shown in the plots of estimations, the sensitivity and specificity of both M2 and M3 are similar to the M1. However, the key parameters of M2 and M3, the γ_0 and γ_1, have a notable deviation from M1. This means the change of correlation has significant damage to the estimation of the conditional probability between X and Y. Compared to the true values of the γ_0 and γ_1, even the M1 has some deviations and the $\gamma_{1,\mathrm{tr}}$ is out of the confidence interval in M1.

Conclusion and Discussion

Based on the comparisons between models, a vogue implication about the damage of violating the conditional independence assumption has drawn out. Besides the worse performance of convergence, a deviation of the estimation is observed.

However, this conclusion may not indicate any causation between the violation and the observed results. The changes in the contribution of Y to the Xstr are crude and straightforward, which may further influence the model itself. The changes of contribution indicate the information loss of the data, and hence, the models use a wrong prior on the distribution of Xstr. The results reflect the effects of the assumption's violation and the information loss of the data.

Some methods may be practical to deal with both the problem of assumption's violation and information loss. One of them is to construct a weighted naïve Bayes model. The weighted naïve Bayes model assumes weighting as an approach to reduce the damage of violation of the independence assumption. [2] It weighs the attributes with potential correlations and keeps the structure of Bayes models as the same. This method can also be applied to the naïve Bayes classifier and improve the performance when dealing with dependent attributes.

On the other hand, it is recommended to check or test the conditional independence of the variables in a dataset and prevent the violation of assumption before conducting the model. One way to test the conditional independence is to utilize the X^2 or G^2 statistics across the individual test between occurrences of events given the other. The independence hypothesis can be rejected once such a rejection is reported in any single test. The other way is to use the Cochran-Mantel-Haenszel test. This test will produce the Mantel-Haenszel statistic and reflect the average partial association between variables. [3]

Reference

1. Dawid, A. P. Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B* 41 (1979), no. 1, 1–31.
2. Zaidi, Nayyar & Cerquides, Jesús & Carman, Mark & Webb, Geoffrey. (2013). Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. The Journal of Machine Learning Research. 14. 1947-1988.
3. 5.4.4 - Conditional Independence | STAT 504. (n.d.). PennState: Statistics Online Courses. Retrieved April 29, 2021, from https://online.stat.psu.edu/stat504/lesson/5/5.4/5.4.4

Appendix

Code: https://github.com/zhuzp98/Violation-of-Conditional-Independence