Bayes Semester Project Report: Turkish Divorce Prediction

Problem Description

We applied a Bayesian Logistic Regression and built a Variational Autoencoder (VAE) to predict divorce among Turkish couples. This is a non-trivial problem as understanding what factors drive divorce can help develop effective couples therapy programs that can be implemented in Turkey and abroad. We implemented the aforementioned Bayesian methods to analyze the "Divorce Predictors" data set from the UCI Machine Learning Repository. Our goal was two-fold, 1. to know the probability of correct classification of couples as being either divorced or married, and 2. to find the likelihood of the importance of the various predictors as they relate to the outcome.

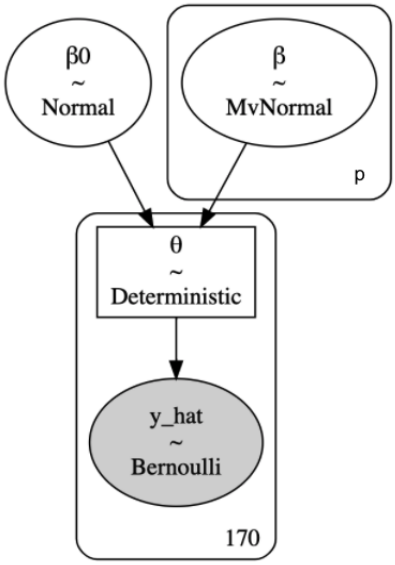
Yöntem et al., (2019) originally conducted a research study to find divorce prediction through correlation based feature selection and artificial neural networks among Turkish couples (we used their data for our project). Participants completed the Divorce Predictors Scale (DPS), which consists of 54 survey questions asked on a 0 to 4 Likert scale (where 0 = *strongly agree* and 4 = *strongly disagree*) to capture intensity of feelings for a given item. Furthermore, these survey questions are modeled after Gottman's Couples Therapy, in which he asserts that divorce prediction can be boiled down to four major tenets: criticism, contempt, stonewalling, and defensiveness (Gottman, 2014; Gottman & Gottman, 2012).

Data came cleaned and standardized with no missing values. Approximately half of the 170 couples reported being divorced, giving us a balanced data set. Importantly, of the married couples, roughly 43% married for love, while the remaining couples had arranged marriages. Yöntem et al., (2019) included only happily married couples (i.e. those who had no intention of getting divorced) in their study. Finally, the researchers claimed to have discovered the six most important DPS survey questions for divorce prediction (please refer to Appendix for list of survey questions). Our project focused on examining the uncertainty estimates around these six key divorce features.

Mathematical Linkage - Problem and Bayesian Methods

Classification of marital status (i.e. divorced or married) lends itself nicely to a Bayesian logistic regression model - the numeric categorical responses to each question in our data set can be used as predictors in such a model. To improve the interpretability of our model parameters, we shifted the predictors by -2, so that more negative values correspond to stronger feelings of agreement and more positive values represent stronger feelings of disagreement. Gaussian priors for model parameters were thought to be appropriate for this model. Our small sample size made MCMC sampling attractive, since training time did not take very long for reasonable sample and burn-in sizes. Moreover, a graphical representation of our model is displayed below, where $p < 55$ is the number of predictor variables included in the model.

Figure 1: GraphViz generated display of our Bayesian logistic regression model.

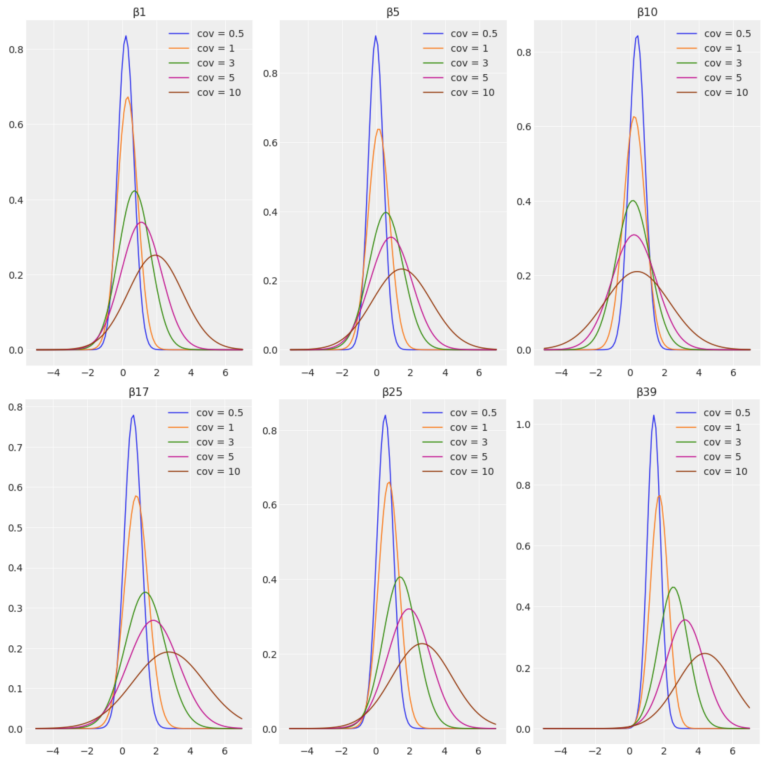


Also, the dimensionality of our model was of interest. Gottman proposes that there are four factors that contribute to the health and success of any marriage. Perhaps one could think of these four attributes as latent variables and to test the validity of the notion that survey responses could be generated by these latent features, we constructed a VAE with four bottleneck variables. Although the four bottleneck variables in our model would not explicitly represent criticism, contempt, stonewalling, and defensiveness, analyzing the loss in the decoding could give merit to the generative nature of Gottman's proposal. Lastly, the four bottleneck variables would be used in a logistic regression model to classify marital status, such that we could then analyze the posterior distribution of their associated model parameters.

Results

Due to the size of our data set and the number of predictors in our logistic regression model, we achieved high classification accuracy. However, we were more interested in the posterior distributions of the parameters associated with each survey question. A correlation matrix of survey responses showed that our predictors are highly correlated. Therefore, we concentrated on the six predictors that are deemed most important by Yöntem et al., (2019). Given the high correlation between the predictors, the scaling of our data, and the small size of our data set, we decided to investigate how distributions changed with different Gaussian priors centered at 0. Figure 2 presents the posterior distributions for these six predictor parameters with five different covariances used in the priors.

Figure 2: The posterior distributions for six model parameters with varying priors.



Through examination of these distributions, our suspicion that the posteriors are heavily dependent on the priors was confirmed – not only was the variance of the posteriors impacted by the priors, but also the position of the peaks. We believe this behavior can be attributed to the high correlation among the predictors and the bias-variance trade-off. As one posterior shifts, they all do. Also, as our priors become uninformed by increasing the covariances, the posterior variance dramatically increases and so too does the maximum a posteriori, which illustrates a case of bias-variance trade-off. Importantly, the intercept parameter is not displayed along with these posteriors, but it demonstrates a similar shift with the change in priors. Based on our examination of the priors' influence on the posteriors, we selected a Gaussian prior centered at 0 with a variance of 3 for each of the parameters of our different models.

Table 1: Model Selection and WAIC Scores.

Model	**# of Predictors**	**WAIC Score**	**Weight**
Reduced 1	46	-3.759229	0.471325
Full	54	-4.045046	0.352066
Recommended	6	-5.532581	0.118931
Reduced 2	32	-6.312563	0.057678

We generated a total of four models, each with a different number of predictors - to achieve this, we performed feature selection by extracting some of the highly correlated

predictors. Table 1 suggests that the Reduced 1 model with 46 predictors performed the best. Importantly, although the researcher-recommended model with 6 predictors did not produce the best classification accuracy, we are not too concerned with this performance. We wanted to explore the uncertainty estimations around the six most important predictors of divorce. So, we manipulated the priors to see how that would impact the subsequent WAIC scores and weights for Bayesian Model Averaging (BMA). We also produced a summary table as well as trace and forest plots for each respective model to further assess uncertainty. Because of the nature of the small data set all the models were overfit, so even the model with 6 predictors did well. Therefore, we care more about the distribution of the parameters and the importance of survey questions rather than the model accuracy.

We constructed a VAE to obtain dimensionality reduction. Upon constructing the VAE, it became clear that 170 survey responses were far too few for reasonable decoding from even a small reduction in dimensionality. This motivated us to explore a more classical approach to dimensionality reduction called Principal Component Analysis (PCA). However, we realized that the dimensionality reduction in PCA, like that of the VAE, would only be interesting with much more data. We felt that using four variables to train a classification model with the small data set was insufficient to serve as evidence for or against Gottman's theory. In our construction of the VAE, very rudimentary sampling methods were utilized to illustrate the pipeline that could be used to classify marital status using four bottleneck variables (note: this pipeline is included in our attached code for reference, but is omitted in this report).

Conclusion

Yöntem et al., (2019) placed a strong emphasis on the accuracy of models (mainly artificial neural networks) in classification of marital status. But, limited discussion about the possibility of overfitting or the analysis of uncertainty was mentioned. We investigated the uncertainty associated with the claims made by these researchers. Although the uncertainty in model accuracy for reduced models was low, it could be improved by using better train/test sets. We were hesitant to split the data into sufficient train/test sets because of its small size. Overall, the posterior distributions of parameters shed light on the correlation between even the most important predictors and the bias-variance trade-off associated with selecting priors with such a small data set. Our recommendation to the researchers would be to acquire more data using the DPS before putting too much stock in the accuracy of the trained models.

We must keep in mind that this research was conducted in Turkey with Turkish couples only, so it would be very interesting to reproduce the study in Turkey and other countries to see how views of divorce and marriage vary across cultures - for example, the notion of arranged marriage is not customary in the U.S. and many other western countries, but it is a common practice in eastern cultures. Differences in classification performance and uncertainty for marriage type (i.e. love or arranged) and country/culture would further provide context as to the generalizability of this analysis and resulting theories.

References

- https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set
- https://dergipark.org.tr/en/download/article-file/748448
- https://www.kaggle.com/alperenclk/for-beginner-divorce-prediction-whit-ann
- https://www.gottman.com/about/the-gottman-method/
- https://blog.keras.io/building-autoencoders-in-keras.html
- https://www.cdc.gov/nchs/data/dvs/national-marriage-divorce-rates-00-19.pdf
- Yöntem, M , Adem, K , İlhan, T , Kılıçarslan, S. (2019). DIVORCE PREDICTION USING CORRELATION BASED FEATURE SELECTION AND ARTIFICIAL NEURAL NETWORKS. Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi, 9 (1), 259-273. Retrieved from https://dergipark.org.tr/en/pub/nevsosbilen/issue/46568/549416.

Appendix

6 most important survey questions for divorce prediction according to Yöntem et al., (2019):

- “I know we can ignore our differences, even if things get hard sometimes.”
- “We don't have time at home as partners.”
- “I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.”
- “My spouse and I have similar ideas about how marriage should be.”
- “I know my spouse's basic anxieties.”
- “We're just starting an argument before I know what's going on.”