STSCI 4740
Data Mining & Machine Learning
Final Project Report
Prediction on Divorce based on Questionnaire Answers

December 10th, 2019

Professor Dr. Yang Ning
STSCI 4740 FALL 2019

Project Group 1:
Gongpu Zhang gz257
Geyu (Kayanne) Chen gc549
Dongchen (Tony) Yuan dy334

Table of Contents:

1. Introduction

The goal of this project is to predict the status of respondents' marriage status based on people's questionnaire answers in the dataset. Our dataset is a questionnaire of 54 survey questions. Respondents in the study answer each question from 0 to 5 to indicate whether they agree or disagree with the questions, as 0 means strongly agree and 5 indicates strongly disagree. The dataset classifies the respondents into two classes. Class 0 denotes the respondents who never go through a divorce in their marriage, and class 1 denotes the divorced respondents. In total, we have 170 instances of survey results.

2. Parameter & Model Selection

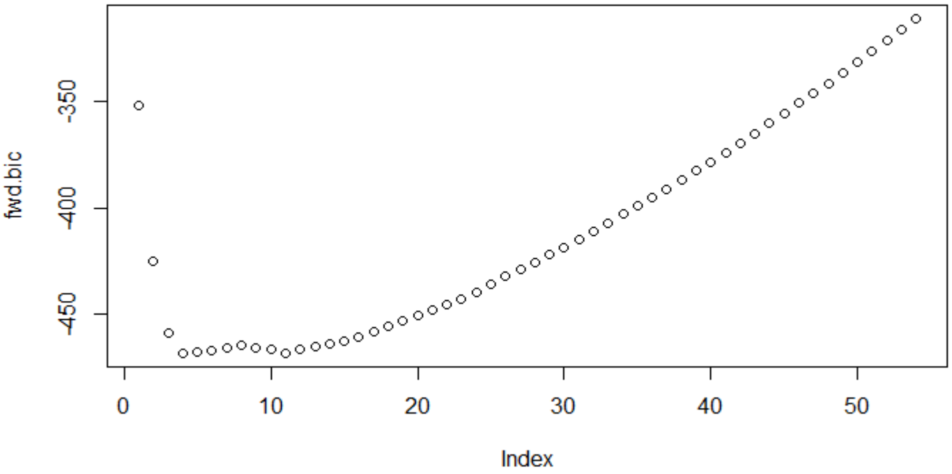
2.1. Parameter Selection (Dimension Reduction)

In the dataset, we have 170 observations with 54 potential features to fit into our prediction model. We began with selecting which features we should include in our prediction model. With a large number of features to choose from, we believed subset selection is the best approach for us to select the appropriate parameters in our model. Subset selection could not only improve the prediction accuracy but also help us better interpret the model. There are mainly three parameter-selection methods for us to choose from: the best subset selection, forward stepwise selection and backward stepwise selection.

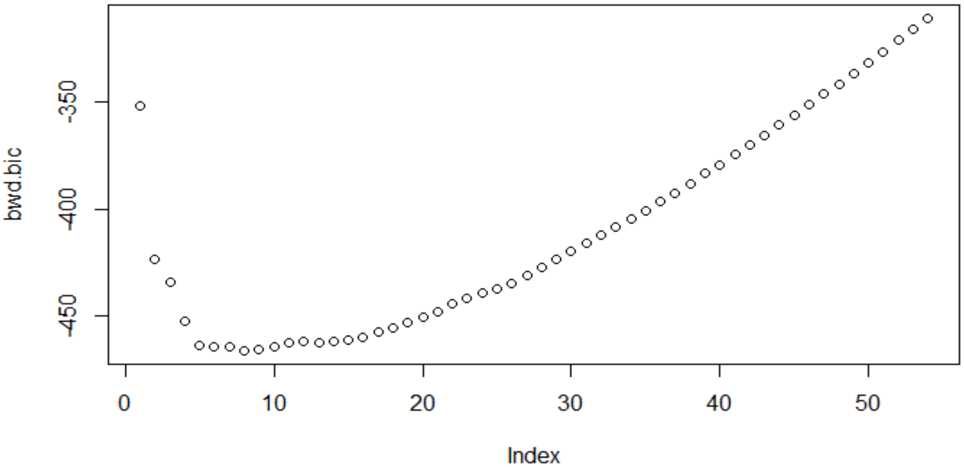
With a number of features (p) equals 54, the best subset selection creates 2^54 models, which would cost gigantic computation capacity. In fact, when we attempted to adopt the best subset selection, our program failed to provide us with a result in an hour. Consequently, we chose to use forward stepwise selection and backward stepwise selection to choose the features we would like to include in our prediction model. In this case, each method would generate 54*(54+1)/2 = 1485 models, which significantly reduced the computation capacity needed to drive for a result.

When we conducted the forward stepwise selection and backward stepwise selection, we decided to identify the best parameters-selection based on BIC (Bayesian Information Criterion) scores. The reason we chose to adopt the BIC scores for comparison was that the BIC scores placed a heavier penalty on models with more variables than the other scores such as Mallows' Cp and AIC. Since this is a classification problem, we chose not to use the adjusted R-squared score, since adjusted R-squared is not applicable in the setting. As mentioned before, there are 54 potential predictors and only 170 observations from our dataset. We would like our prediction model to have better interpretability with as few predictors as possible, so we examined BIC scores to find the most efficient features/predictors we would like to include in our model.

After conducting the forward stepwise selection and backward stepwise selection, we observed BIC scores of different models. The lower the BIC scores, the better the model performs. Here are the scatter plots of the BIC values from forward stepwise selection and backward stepwise selection. The horizontal axis is the number of predictors in the model, and the vertical axis is the BIC score.



Plot I. Forward stepwise selection BIC scores plot



Plot II. Backward stepwise selection BIC scores plot

We observed that there are two models achieving the lowest BIC scores for a forward stepwise selection. One model has 4 predictors: Atr6, Atr18, Atr29, Atr40, and the other model has 11 predictors: Atr6, Atr17, Atr18, Atr24, Atr25, Atr26, Atr29, Atr40, Atr46, Atr49, Atr52. For the backward stepwise selection, the model with 8 parameters has the lowest BIC score. The 8 predictors are: Atr6, Atr17, Atr25, Atr26, Atr40, Atr46, Atr49, Atr52. Here is a table presentation of different parameter selection results:

Method	Parameters Selection													
F-4	6		18				29	40						
F-11	6	17	18	24	25	26	29	40				46	49	52
B-8	6	17			25	26		40	46	49	52			

Table I: 3 Models Parameter Choosing

From the table above, we observed that there is a significant overlapping of parameters among the three models we select. Having three models with different number of predictors to choose from, we started to find the model that predicts the divorce status with the highest accuracy (lowest test error rate).

2.2 Model selection

This project is a classification problem in nature, so we were looking for a classification model to fit the dataset. Intuitively, logistic regression seems to be the most reasonable model for us to choose from. We randomly divided the dataset by half into training dataset and testing dataset. After fitting the model in the training set, we computed the test MSE and compared the results by choosing the model with the smallest test error rate.

Initially, we ran the logistic regression with the four-predictors model (Atr6+Atr18 +Atr29+Atr40) and got the warning message from R: "*glm.fit: algorithm did not converge; glm.fit: fitted probabilities numerically 0 or 1 occurred."* This error message translated into "our dataset was well-separated". When the classes are well-separated, logistic regression could be extremely unstable. In fact, when we attempted to fit the logistic regression models with 8 and 11 predictors, the same error occurred and all the test error rates were above 60%. Linear Discriminant Analysis (LDA) could effectively solve this problem.

Consequently, the LDA model became a strong candidate for us to choose from. Meanwhile, QDA is another model that we should consider to apply to our dataset. Comparing to LDA, QDA model has more flexibility with lower bias. The side-effect of QDA comes from the trade-off of bias and variance, as increasing the flexibility would inevitably increase the variances of the model. To see which model performs better, we need to evaluate each model based on their performances on the test error rate.

3. Model Evaluation

To conduct a model evaluation, we first tried to fit the LDA model onto the dataset. In our first attempt, we fitted the data to the full model, namely in a total of 54 predictors, to see whether LDA solve the issue of well-separated classes as we expected before. The problem was solved and we obtained a test MSE of 0.09411765. We then fitted the LDA model with 4 predictors, getting a test MSE of 0.03529412. The LDA model with the 11-predictors model had a test MSE of 0.03529412. The LDA model with 8 predictors derived a test MSE of 0.02352941, which is the smallest among all.

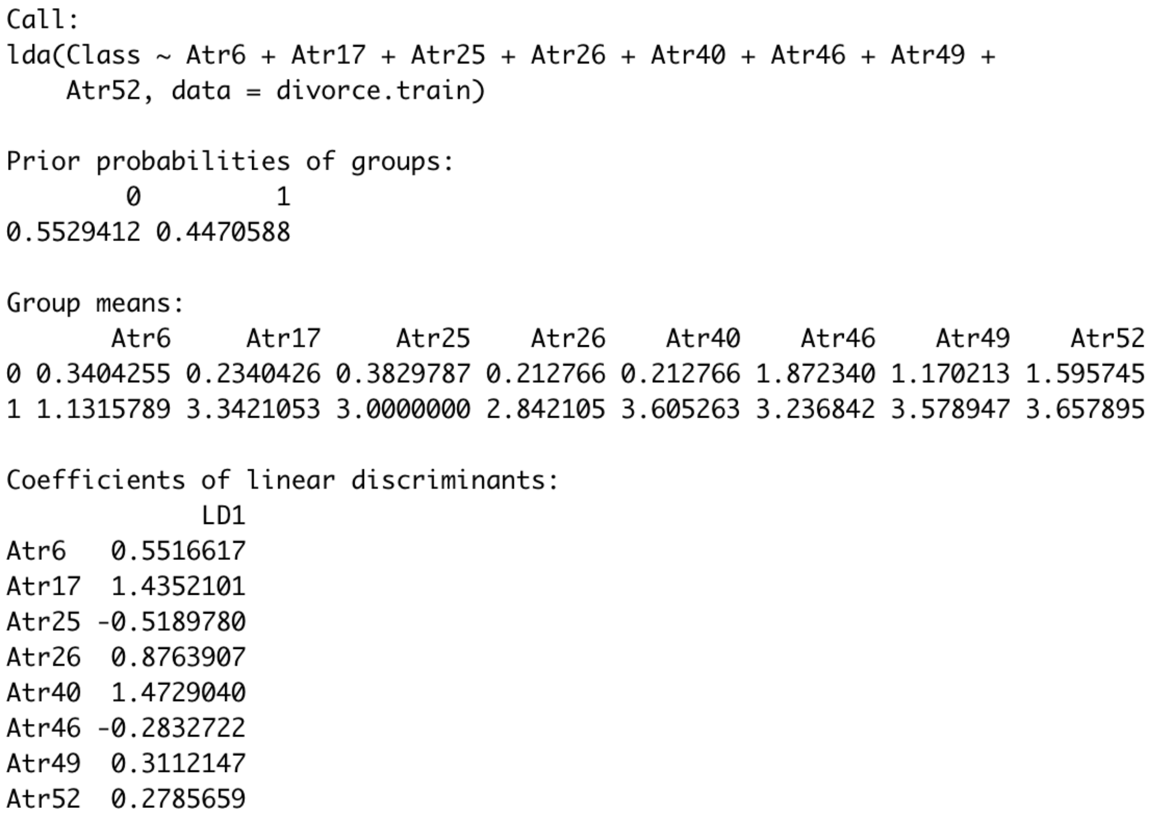
lda.pred \ divorce.test.Class	0	1
0	39	2
1	0	44

Table II: the output of the LDA model with 8 predictors

The table above summarizes how the LDA model with 8 predictors predicts the respondents in our test dataset. There are in total of 85 observations in the test data set. Among them, only 2 observations that should be class 1 were wrongly classified as class 0. So, the test error rate is approximately 2.85% in this case. Even though the model has already obtained terrific accuracy, we would also like to check how QDA performs with the dataset.

Next, our group tried the QDA approach, to see if the QDA model would out-perform the LDA model. Similar to what we did before, we first fitted the QDA model with all the predictors we have to have a generic look. However, the first fit of QDA did not work, as R reported "*Error in qda.default(x, grouping, ...) : some group is too small for 'qda'* ." Since QDA could suffer from the curse of dimensionality, where the matrix of our divorce dataset is not invertible. Therefore, QDA is not suitable for this problem.

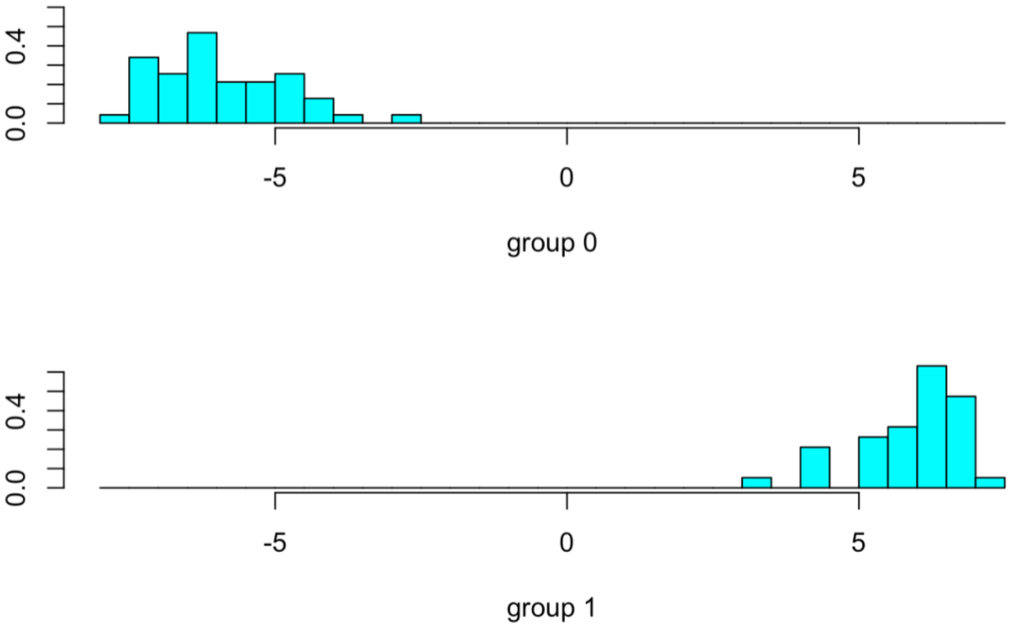
After getting the result of test MSE for different models, we decided to use the LDA model with 8 predictors. Since it has the lowest test MSE of 0.0235294 among all the models we tested in this practice. Below is the R code presenting the details of the LDA model with 8 predictors.



```
Call:
lda(Class ~ Atr6 + Atr17 + Atr25 + Atr26 + Atr40 + Atr46 + Atr49 +
    Atr52, data = divorce.train)

Prior probabilities of groups:
        0         1 
0.5529412 0.4470588 

Group means:
       Atr6     Atr17     Atr25    Atr26    Atr40    Atr46    Atr49    Atr52
0 0.3404255 0.2340426 0.3829787 0.212766 0.212766 1.872340 1.170213 1.595745
1 1.1315789 3.3421053 3.0000000 2.842105 3.605263 3.236842 3.578947 3.657895

Coefficients of linear discriminants:
             LD1
Atr6   0.5516617
Atr17  1.4352101
Atr25 -0.5189780
Atr26  0.8763907
Atr40  1.4729040
Atr46 -0.2832722
Atr49  0.3112147
Atr52  0.2785659
```

Table III: LDA output with 8 predictors

From the output of LDA, we see that πˆ0 = 0.553 and πˆ1 = 0.447. So, in our training dataset, 55.3% of the observations are those who have never been through a divorce. We could also find the group means of each predictor for the two classes. From the group means, we observed that class 1 tends to have higher values of each predictor. Such a pattern indicates that divorced individuals are more likely to disagree with the selected questions of this model in the survey. From the LDA output, the coefficients provide the linear combination of Lag1 to Lag 8 that is used to set up the LDA decision rule. The plot of the linear discriminants is attached below.



Plot III: linear discriminant coefficients calculated results of group 0 and 1

As we observed from the plot III, by calculating the different group results and plot those onto the histogram, we can see that group 0 is clustered on the left side of the axis around -5 while group 1 is on the right side around +5. This plot visualizes how discriminant coefficients are functioning as a linear combination of input variables which are used to calculate the posterior probability of class membership, during the second stage of LDA. This is because LDA has two stages: extraction and classification. Latent variables called discriminants are formed at the extraction stage, and data points are assigned to classes by discriminant coefficients, instead of original variables, at the classification stage.

Thus, if we multiply each answered survey value of LDA1 (the first linear discriminant) by the corresponding coefficient of each predictor and sum them up, for example: 0.5516617*1 + 1.4352101*5 - 0.5189780*0 + 0.8763907*2 + 1.4729040*0 - 0.2832722*4 - 0.3112147*2 + 0.2785659*3, we get a score for each respondent and we are able to assign the respondent to one of our classes. In this respondent, the sum output is 8.56 as the posterior probability, and according to the plot III, we probably would assign this respondent to group 1 where one has the highest probability. More specifically for each individual coefficient is: as for variable 6, with 0.5516617, it means holding other variables fixed, increasing one unit of the answer to survey question 6 stands for increasing 0.5516617 tendencies over the total tendency of being assigned to group 1, which means divorce ending. This coefficient could also be negative sometimes, for predictor 25 and 46, means a negative correlation to group 1. In these situations, increasing one unit of such an answer leads to higher probability of being assigned to group 0.

4. Conclusion

In conclusion, we recommend using the LDA model with 8 predictors for the divorce prediction of the survey answers. If a respondent presents his/her answer to these 8 questions below, our model will iterate through those answers and give a prediction on the marriage status of the respondent:

6. We don't have time at home as partners.
7. We are like two strangers who share the same environment at home rather than family.
25. I have knowledge of my spouse's inner world.
26. I know my spouse's basic anxieties.
40. We're just starting a discussion before I know what's going on.
46. Even if I'm right in the discussion, I stay silent to hurt my spouse.
49. I have nothing to do with what I've been accused of.
52. I wouldn't hesitate to tell my spouse about her/his inadequacy.

Based on the results from our LDA model, higher scores in questions 6, 7, 26, 40, 49 and 52 accompanying with lower scores in questions 25 and 46, would indicate higher possibility of assigning the individual into class 1, the divorce group.

5. References

1) Data source: *https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set#*
2) James, Gareth, *An Introduction to Statistical Learning: with Applications in R*,   Springer, 2017.