Homework 5

Alex Soupir

February 14, 2020

Packages: ISLR, MASS, mclust, ggally

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

You do not need to include the above statements.

Please do the following problems from the text book ISLR or written otherwise.

1) Question 4.7.6 pg 170

Suppose we collect data for a group of students in a statistics class with variables $X1$ = hours studied, $X2$ = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1-e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$

$p(X) = \frac{e^{-6+0.05(40)+1(3.5)}}{1+e^{-6+0.05(40)+1(3.5)}}$

```
## [1] "Probability a student will get an A with studying 40 hours and a GPA of 3.5:"
```

```
## [1] 0.3775407
```

b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1-e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$

$0.50 = \frac{e^{-6+0.05X_1+1(3.5)}}{1+e^{-6+0.05X_1+1(3.5)}}$

$0.50(1 + e^{-6+0.05X_1+1(3.5)}) = e^{-6+0.05X_1+1(3.5)}$

$0.50 + 0.50e^{-6+0.05X_1+1(3.5)} = e^{-6+0.05X_1+1(3.5)}$

$0.50 = (1e^{-6+0.05X_1+1(3.5)}) - (0.50e^{-6+0.05X_1+1(3.5)}) = 0.50e^{-6+0.05X_1+1(3.5)}$

$0.50 = 0.50e^{-6+0.05X_1+1(3.5)}$

$1 = e^{-6+0.05X_1+1(3.5)} = e^{0.05X_1-2.5}$

$ln(1) = 0.05X_1 - 2.5$

```
## [1] "Hours needed for student from part (a) to get an A:"
```

```
## [1] 50
```

2) Question 4.7.7 pg 170

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a divident was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$. *You will need to use Bayes' theorem.*

π_k = fraction of training observations that belong to class k

$\pi; \pi_{yes} = 80\%$ 0.8

$\pi_{no} = 0.2$

$\hat{\sigma} = 6$

$\bar{x_{yes}} = 10$

$\bar{x_{no}} = 0$

$f(x)$ =normal distribution for x_k, σ_k, and the 4% quantile

$f_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}$; given $\pi = 0.8$, $\sigma^2 = 36$, $x = 4$, $\mu_{yes} = 10$

$f_2(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}$; given $\pi = 0.8$, $\sigma^2 = 36$, $x = 4$, $\mu_{no} = 0$

(4.11)

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}$$

```
## [1] "Probability that a company will issue a dividend given its percent profit was x = 4 last year:"
```

```
## [1] 0.7518525
```

3) Continue from Homework #3 & 4 using the **Weekly** dataset from 4.7.10), fit a model (using the predictors chosen for previous homework) for classification using the MclustDA function from the mclust-package.

```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

In the previous homework, Homework 4, I wasn't able to find any combinations of covariates that was able to increase the predictions accuracy of the models other than the `Lag2` value alone. Here, for the values for MclustDA I will use `Lag2` alone as well.

`i) Do a summary of your model.`

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
```

```
##
##  log-likelihood   n df       BIC
##       -2129.439 985 10 -4327.804
##
## Classes   n     % Model G
##    Down 441 44.77     V 2
##    Up   544 55.23     V 2
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   76 365
##   Up     70 474
## Classification error = 0.4416
## Brier score          = 0.2452
##
## Test confusion matrix:
##       Predicted
## Class  Down Up
##   Down    5 38
##   Up      9 52
## Classification error = 0.4519
## Brier score          = 0.2511
```

-What is the best model selected by BIC? Report the Model Name and the BIC. (See https://www.rdocume

The best model that was selected by mclust was simply V, which is variable/unequal variance (one-dimentional), guessign due to only having a single predictor was was found in the last assignment. This model type produced a BIC of -4327.804.

-What is the training error? What is the test error?

Training error is 44.2% and the testing error is 45.2%, which is similar.

-Report the True Positive Rate and the True Negative Rate.

True Positive rate and true negative rate calculated by the testing data are: True positive rate = 5/(5+38) = 0.116; True negative rate = 52/(9+52) = 0.852.

ii) Specify modelType="EDDA" and run MclustDA again. Do a summary of your model.

```
## ---------------------------------------------
## Gaussian finite mixture model for classification
## ---------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -2204.237 985  3 -4429.152
##
## Classes   n     % Model G
##    Down 441 44.77     E 1
##    Up   544 55.23     E 1
##
## Training confusion matrix:
```

```
##       Predicted
## Class  Down  Up
##   Down   22 419
##   Up     20 524
## Classification error = 0.4457
## Brier score          = 0.2462
##
## Test confusion matrix:
##       Predicted
## Class  Down Up
##   Down    9 34
##   Up      5 56
## Classification error = 0.375
## Brier score          = 0.24
```

-What is the best model selected by BIC?

The best model selected based on the BIC was the E model, or equal variance.

-Find the training and test error rates.

The training error rate is 44.57%, and the testing error rate is 37.5%.

-Report the True Positive and True Negative Rate.

The true positive rate of the testing data is 9/(9+34)=0.209, and the true negative rate of the testing data is 56/(5+56)=0.918.

iii) Compare the results with Homework \#3 \& 4. Which method performed the best? Justify your answer.

The first mclust model using just the Lag2 variable, as found to be the best in Homework 4 though trial and error with other variable combinations and transformations and interactions, produced a model accuracy of 54.8% and the model with EDDA produced an accuracy of 62.5%. Specifying model type to EDDA rather than MclustDA (default), increases the model's prediciton accuracy to the same as the logistic regression and LDA accuracy on the testing split.

4) Continue from Homework #3 & 4 using the **Auto** dataset from 4.7.11). Fit a classification model (using the predictors chosen for previous homework) using the MclustDA function from the mclust-package. Use the same training and test set from previous homework assignments.

```
## Columns that were determined to be most associated with mpg01 in (b):
## [1] "cylinders"    "displacement" "horsepower"   "weight"       "mpg01"
```

i) Do a summary of your model.

```
## ----------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -4706.255 312 98 -9975.324
##
## Classes   n     % Model G
```

```
##       0 152 48.72   EEE 7
##       1 160 51.28   EEE 9
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 138  14
##     1  11 149
## Classification error = 0.0801
## Brier score          = 0.0593
##
## Test confusion matrix:
##      Predicted
## Class  0  1
##     0 41  3
##     1  0 36
## Classification error = 0.0375
## Brier score          = 0.0355
```

-What is the best model selected by BIC? Report the model name and BIC.

The berest model selected is EEE, ellipsoidal, equal volume, shape, and orientation. The BIC produced by this model is -9975.324.

-What is the training error? What is the test error?

Training error rate is 0.0801, or 8%, and testing error for mclust was 0.0375, or 3.8%.

-Report the True Positive Rate and the True Negative Rate.

The true positive rate and the true negative rate using the testing data: true positive rate = 41/(41+3) = 0.932; true negative rate = 36/(36+0) = 1.000.

ii) Specify modelType="EDDA" and run MclustDA again. Do a summary of your model.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##        -5514.18 312 28 -11189.16
##
## Classes   n     % Model G
##       0 152 48.72   VVV 1
##       1 160 51.28   VVV 1
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 134  18
##     1  17 143
## Classification error = 0.1122
## Brier score          = 0.0975
##
```

```
## Test confusion matrix:
##      Predicted
## Class  0  1
##     0 38  6
##     1  1 35
## Classification error = 0.0875
## Brier score          = 0.0654
```

-What is the best model selected by BIC?

The best model that was selected using EDDA in `MclustDA` was VVV, or ellipsoidal, varying volume, shape, and orientation, with a BIC of -11189.16.

-Find the training and test error rates.

The training error of the model is 0.1122, or 8%, and the testing error of the model is 0.0875, or 8.9%.

-Report the True Positive and True Negative Rate.

True positive using the testing data with the model type specified to EDDA had a true positive rate of 38/(38+6) = 0.864 and a true negative rate of 35/(35+1) = 0.972.

iii) Compare the results with Homework \#3 \& 4. Which method performed the best? Justify your answer.

Previously for the `Auto` data, a logistic regression, linear discriminate analysis, quadratic discriminate analysis and KNN were performed. The logistic regression produced a error rate on the same testing split of 0.0625, while LDA and QDA had testing errors of 0.0875. THe KNN also produced a testing error of 0.0625. Now this week, adding to the list the Mclust with model types MclustDA and EDDA, we have 2 more testing errors which can be compared. First, the default for MclustDA is the model type of MclustDA, and this produced a testing error lower than any of the other models at 0.0375. This is almost half that of the logistic regression and KNN. However, the EDDA model type was the same as the LDA and QDA models. So the overall best model used thus far is created with MclustDA with the model type defaulting to MclustDA.

5) Read the paper "Who Wrote Ronald Reagan's Radio Addresses?" posted on D2L. Write a one page (no more, no less) summary. *You may use 1.5 or double spacing.*

Summary:

The data being used in this study is from the speeches on radio and news addresses from Ronald Regan. Many of the speeches are known in origin but there are a few that do not have a definitive author. It has been proposed before that using word length frequencies and length of documents, the author of the speech or document can be narrowed but this had been disproved and other methods of mathmatically providing evidence for authorship has been worked on. The goal of this paper was to give a mathmatical backing to who authored the unknown addresses given by Ronald Regan. They attempt to use a few different approaches to predict the authorship of the known make these predictions, including Bayesian models, cross validation, and other 'machinelearning' classification methods.

The authors used 3 different feature classes which included the word lengths, sequence of words of length n (n words in such sequence), and the featuers of semantics. There were 55 words that different between authors after applying a false discovery rate to the most frequent 3000 words. For the word sequences, 62 words were selected from a pool of 523 words. The semantics resulted in 6 discriminating features that weren't very strong and there was also a information gain variable that I do not fully understand, which cam from text classification from computer science, which ended up adding more discriminating words to the features that could be used. Additionally, more 4 word sequences were down selected. They show that discriminating factors can be visualized with principal components and that unsupervised clustering doesn't provide distinct clustering of authors even when using different metrics to calculate the 'close-ness' due to, most likely, not having enough data from the other authors.

The prediction accuracy was calculated using an 80/20 train/test split on the data and then looking at the predictions over 1000 iterations. The poisson predictions found that the Δ^2 statistic was good at correctly predicting the author of the address upwards of 90% of the time, but sometimes dropping lower. The negative-binomial using out of sample cross-validation produced prediction accuracies between 80% and 95%. They also showed that using cross-validation slightly decreased the prediction accuracy from 125/136 to 121.1/136 in the worst case over testing the model on the same data that was used to create it. A model voting was also done in a sort of ensemble method where if the majority of the models thought an address was from someone, that is where the ensemble vote went. They tested the models on 2 different addresses that had unknown authors of them. In both situations the authors of the publication think that Regan is the one who wrote them based on the tone and what is portrayed within them. In one of the addresses, all of the models predicted that Regan was the author but for the other address there was predictions that is was either him or Hannaford based on the model that was being used to create the prediction. However, for the second address, the Poisson model and the negative-binomial that isn't dependent on β all predict that it is written by Regan.

6) Last homework you chose a dataset from this website. Please do some initial exploration of the dataset. If you don't like the dataset you chose you may change it with another. It has to be a new dataset that we haven't used in class. Please report the analysis you did and dicuss the challenges with analyzing the data. The Mammographic Mass Data Set can be found here. *Any plots for this question need to be done using only GGplot2-based plots.*

Summary of the data after converting the *?* that they had instead of NAs to NAs and then resetting the factor levels. Also, R when importing thinks that Age is a factor and that the response is an integer, so I converted age to a number and Severity to a factor where 0 is benign and 1 is malignant.

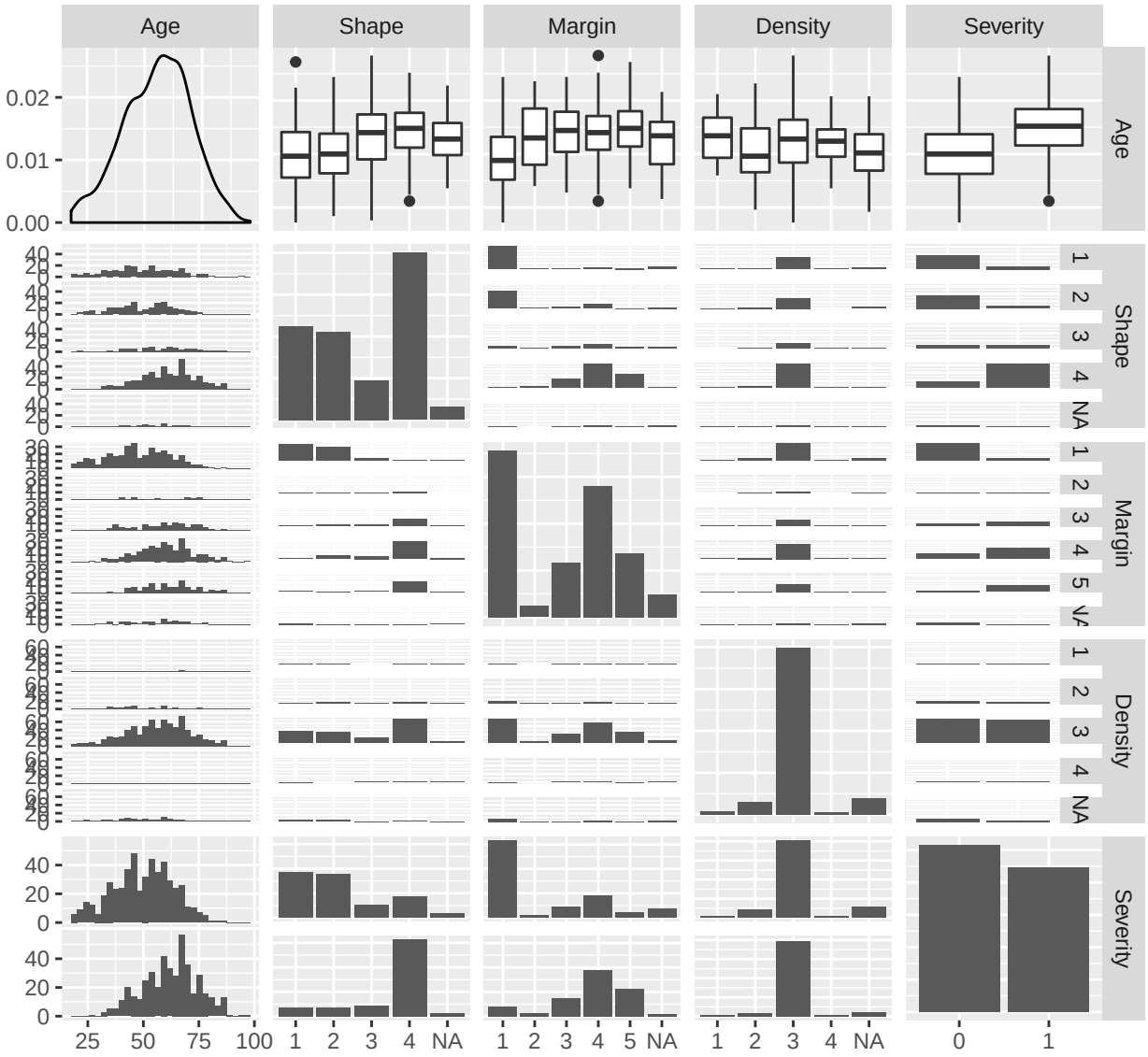
```
##      BIRADS         Age          Shape      Margin    Density    Severity
##  4      :547   Min.   :18.00   1   :224   1   :357   1   : 16   0:516
##  5      :345   1st Qu.:45.00   2   :211   2   : 24   2   : 59   1:445
##  3      : 36   Median :57.00   3   : 95   3   :116   3   :798
##  2      : 14   Mean   :55.49   4   :400   4   :280   4   : 12
##  6      : 11   3rd Qu.:66.00   NA's: 31   5   :136   NA's: 76
##  (Other):  6   Max.   :96.00              NA's: 48
##  NA's   :  2   NA's   :5
```

This shows wht was mentioned for Homework 4 with the missing values. Since this is just exploratory, I won't look to try and fix this other than looking at how many observations would be removed if I wanted to work with complete cases. There are 2 observatiosn that are missing BIRADS, which has for some reason a factor level of 6 so that is going to have to be explored more because BIRAD ranking is from 1 to 5. Age has 5 values missing with a range from 18 to 96 years old. Shape has 4 factor levels (good!) with 31 values missing. Margin has the 5 expected factor levels and 48 missing. And finally, Density have 4 factor levels with a high 76 values missing. The data has 516 benign breast masses and 445 malignant masses.

```
##  BIRADS        Age          Shape      Margin    Density    Severity
##  0 :  5   Min.   :18.00   1   :223   1   :355   1   : 15   0:512
##  2 : 14   1st Qu.:45.00   2   :210   2   : 24   2   : 58   1:435
##  3 : 36   Median :57.00   3   : 92   3   :111   3   :787
##  4 :547   Mean   :55.43   4   :393   4   :277   4   : 11
##  5 :345   3rd Qu.:66.00   NA's: 29   5   :134   NA's: 76
##  6 :  0   Max.   :96.00              NA's: 46
##  55:  0   NA's   :5
```

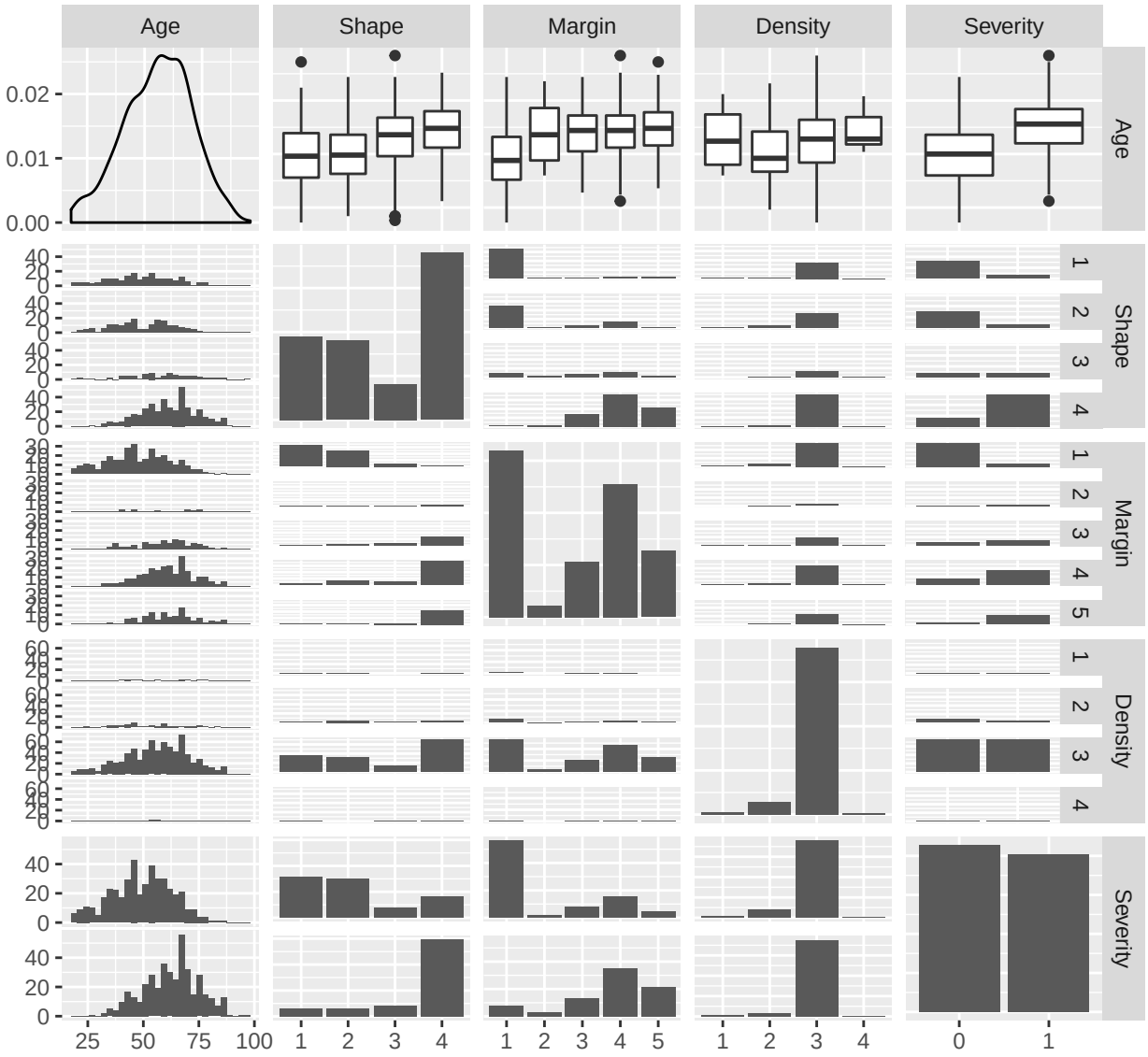
it appears that there is even 0's in the BIRAD which is not a given factor level in the documentation. BIRAD isn't a predictor, but it would be nice to be able to compare our model when it is made with the performance of the doctors. For not I will keep the original data frame but remove BIRADS since it isn't something that can be used to predict severity anyway.

On the following page is a plot matrix for the 4 predictors of the severity response. Looking at the Severity column, there are some differences in the distribution for Age, Shape, and Margin between benign and malignant patients. Density doesn't visually show a difference between benign and malignant with a large number being of Density 3 - low mass density. The distribution of age shows a large frequency of patients being around 60 years old, while the low and high end of age is tapered down which could cause issues (have read some places that unequal distribution can cause issues with not having enough data to be confident on the ends of the range; Ronald Regan paper also said something about issues possibly arrising from not having enough data from the other authors). This may also shows up in the Margin and Density data with a Margin of 1 and 4 having the largest frequencies, and Density of 3 having an large proportion of the samples belonging to that factor level. This may be something to keep in mind later on when building a model if training set ends up with almost all of Density 3 and then it tries to predict something later with a density of 1 or even 4; just a note.




```
##       Age         Shape    Margin   Density Severity
##  Min.   :18.00   1:190   1:320   1: 11   0:428
##  1st Qu.:46.00   2:180   2: 23   2: 56   1:403
##  Median :57.00   3: 81   3:106   3:756
##  Mean   :55.78   4:380   4:255   4:  8
##  3rd Qu.:66.00           5:127
##  Max.   :96.00
```

Removing all of the NAs in the data frame decreases the number of observations by 130. The distribution of Severity is now more balanced between benign and malignant. There still appears to be a difference in age distribution peaks and Shape frequency and Margin frequency.



Another challenge may eventually come form age being the only continuous variable for a predictor, and the rest being factors. I do like this data set though and the BIRAD will be interesting to compare the model created with the 2-doctor consensus.

References:

- StackOverflow
- rdrr.io - mclust documentation