Homework 5

Andrew Boschee

No Collaborators. Outside Resources: Elements of Statistical Learning, Rdocumentation.com

Probability Calculations

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,

$$\beta_0 = -6$$

,

$$\beta_1 = 0.05$$

and

$$\beta_2 = 1$$

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

Lay out formula:

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

Plug in values:

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

Solve:

Table 1: Probability of Getting an A

0.3775

Given the student studies 40 hours and has a GPA of 3.5, they have roughly 38 percent probability of getting an A

Part 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?

Set the equation equal to 0.5:

$$0.5 = \frac{e^{-6+0.05\times 40+1\times 3.5}}{1 + e^{-6+0.05\times 40+1\times 3.5}}$$

Which becomes equal to:

$$log(\frac{0.5}{1-0.5}) = -6 + 0.05X_1 + 1 \times 3.5$$

Solve:

Table 2: Estimated Hours to Study

50

Given, the student has a 3.5 gpa, they should study 50 hours for a 50 percent chance of getting an A

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, which equals 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 dividend was $X = 10$, while the mean for those that didn't was $X = 0$. In addition, the variance of X for these two sets of companies was

$$\sigma^2 = 36$$

. 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.

Results:

$$p(4) = \frac{0.8e^{-(1/72)(4-10)^2}}{0.8e^{-(1/72)(4-10)^2} + 0.2e^{-(1/72)(4-0)^2}}$$

Solve

Table 3: Probability of Paying Dividend

0.7519

Plugging values into the equation, the probability of a dividend being paid comes out to roughly 75%.

Mclust Classification - Weekly Data

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.

i) Do a summary of your model.
-What is the best model selected by BIC? Report the Model Name and the BIC. (See https://www.rdocumentation.org/packages/mclust/versions/5.4/topics/mclustModelNames)
-What is the training error? What is the test error?
-Report the True Positive Rate and the True Negative Rate.

Table 4: Weekly Model Summary

Cluster Model	Model Type	BIC	Train Error	Test Error	Accuracy
V	univariate, unequal variance	-4327.804	0.442	0.452	0.548

Table 5: True Positives and True Negatives

TP	TN
0.852459	0.2093023

EDDA

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

-What is the best model selected by BIC?

-Find the training and test error rates.

-Report the True Positive and True Negative Rate.

Table 6: Weekly Model Summary

Model	Type	BIC	Train Error	Test Error	Accuracy
E	univariate, equal variance	-4429.15	0.446	0.375	0.625

Table 7: True Positives and True Negatives

TP	TN
0.9180328	0.1162791

iii) Compare the results with Homework #3 & 4. Which method performed the best? Justify your answer. *Here you need to list the previous methods and their corresponding rates.*

Table 8: Method Comparison - Weekly

Logistic Regression	GLM	LDA	QDA	KNN (k = 15)	EEV	VVV
0.56	0.625	0.625	0.59	0.55	0.55	0.625

For each method, the training and test sets were split 75/25 as before and used the mclust function to find the method. From the summaries, cofusion matrices are available and classification rates are shown. Finding true positive and true negatives was done in similar manner as prior assignments writing out the equations and pulling values from the confusion matrix. When comparing across all methods, there is not as much improvement as expected. I'm guessing that the limited size of the data is partially responsible for this situation with GLM, LDA, and VVV getting the exact same accuracy. I believe if we had a larger sample that it may give us some more insight on the effectiveness of using various modeling methods. While KNN and logistic regression are the simple and easy to explain methods, they definitely would not be the method of choice in the end here.

Mclust Classification - Auto Data

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.

i) Do a summary of your model. -What is the best model selected by BIC? Report the model name and BIC.

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

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

Table 9: Model Summary

Model	Model Type	BIC	Train Error	Test Error	Accuracy
EEV	ellipsoidal, equal volume and shape	-10418.83	0.037	0.112	0.888

Table 10: True Positives and True Negatives

TP	TN
0.8958333	0.8627451

EDDA

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

-What is the best model selected by BIC?

-Find the training and test error rates.

-Report the True Positive and True Negative Rate.

Table 11: Auto EDDA Model Summary

Model	Model Type	BIC	Train Error	Test Error	Accuracy
VVV	ellipsoidal, varying volume, shape, and orientation	-12130.58	0.092	0.112	0.888

Table 12: True Positives and True Negatives

TP	TN
0.8958333	0.8627451

iii) Compare the results with Homework #3 & 4. Which method performed the best? Justify your answer. *Here you need to list the previous methods and their corresponding rates.*

Table 13: Method Comparison - Auto

Logistic Regression	GLM	LDA	QDA	KNN (k = 10)	EEV	VVV
0.56	0.625	0.89	0.89	0.89	0.89	0.89

As expected from prior assignments, the models performed fairly well and comparable to the LDA, QDA, and KNN models. Similar to the Weekly dataset, it would be interesting to see a larger dataset that would make it much less likely to have the exact same accuracy and be able to see a more in depth comparison of the models.

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.*

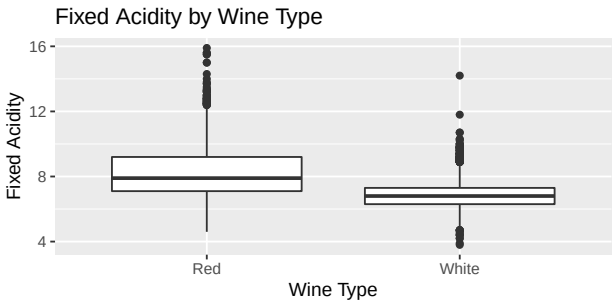
The study was composed of radio addresses by Ronald Reagan between the years of 1975 and 1979 to aid in making his presence known during his presidential campaign. Through this time there were roughly 1000 addresses with about 2/3 known to be written by Reagan himself, 39 by his staff, and the remaining uncertain. As expected in politics, these addresses were across various topics that contribute to events and opinions relevant to his campaign. With uncertainty of who authored the roughly 1/3 of his remaining radio addresses, an attempt was made to find characteristics of writing styles by authors of the given data to classify those remaining radio addresses. With a lack of data coming from non-Reagan addresses, I had immediate

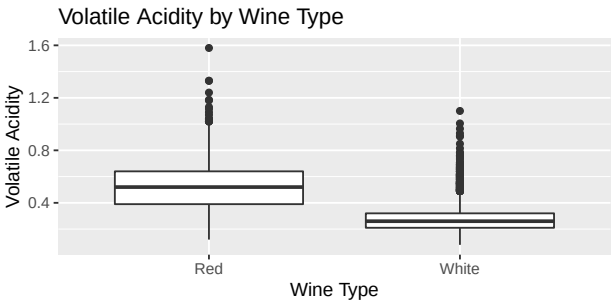
concern on how distinguishable it really could be when comparing between authors. This also varied by year as we could see from the table breaking down texts by author over the five-year period. Surprisingly, M. Anderson only had one authorship through that whole timespan and I am curious what may have caused the variance of authorship in relation to political and campaigning circumstances. Did Reagan need to change his tone or communication style over time in response to political bias and criticism? This also potentially factors in to how distinguishable the writing styles may be over those years. Multiple methods were used in EDA as well as classification methods through this study. To help analyze grouping of words, n-grams were used. This stuck out the most to me since it can be modified and help see phrases or common habits in an authors writing style. This also relates to the principal component analysis looking at the highest frequency words while eliminating words that don't provide any distinguishable benefit. This allows for clustering of word use to also get a broader picture of each authors habits. They seemed to have a strong liking to the Negative-Binomial model (while stating it's optimistically biases) with accuracies between 92 and 99 percent. Was a little surprised with their conclusion that Reagan drafted 77 speeches, and his collaborators 71, in 1975. Then, from 1976-79, Reagan drafts 90 and Hannafort 74. That seems much more balanced than I expected from my first impression that Reagan did the majority by himself. I was glad that in the end they were very thorough about the timing of authorships and separating the models for authorship in 1975 and 1976-1979. It's interesting to see the difference in how they analyzed radio addresses back then while we have so many social media outlets and biased news stations today to potentially impact these types of studies.

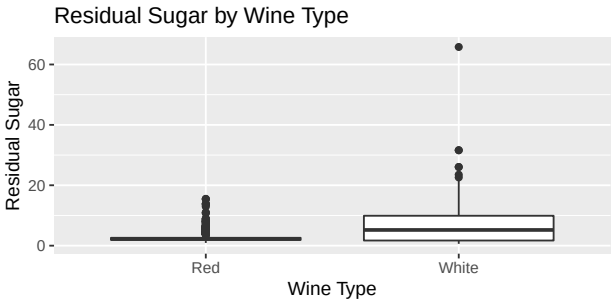
Wine Quality EDA

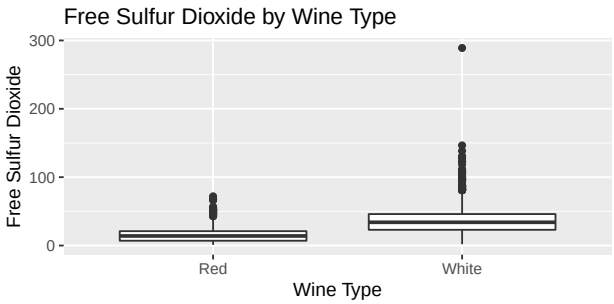
6) Last homework you chose a dataset from this website. Please do some initial exploration of the dataset. If you don't like the dataset you chose you may change it with another. It has to be a new dataset that we haven't used in class. Please report the analysis you did and dicuss the challenges with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

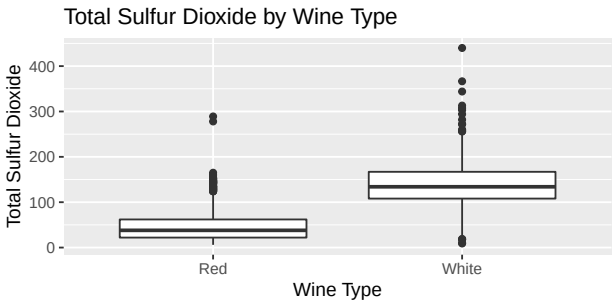
After combining the red and white wine datasets from the csv files, I added a type column to create the dependent variable for classification. With not much background/conceptual knowledge of wine, I am just looking to get a high level view of the independent variables for the model. To try and see what variables are likely to have significant role in relation to classification of red/white, boxplots were made for all independent variables in relation to classification. The ones that stood out the most and that I expect to have a strong impact on classification are shown below with others commented out.








Free Sulfur Dioxide by Wine Type
Free Sulfur Dioxide
300
200
100
0
Red
White
Wine Type


Total Sulfur Dioxide by Wine Type
Total Sulfur Dioxide
400
300
200
100
0
Red
White
Wine Type


Density by Wine Type
Density
1.04
1.03
1.02
1.01
1.00
0.99
Red
White
Wine Type


Sulphates by Wine Type
Sulphates
2.0
1.5
1.0
0.5
Red
White
Wine Type