STA 135 Final Project

Analysis of Pima Indians Diabetes

Trevor Adam, Ameya Gaitonde, Jeff Lee, Marco Oviedo

June 13, 2023

Contributions

Trevor Adam - Box's M Test, QDA, conclusions

Ameya Gaitonde - QDA full dataset, confusion matrices, conclusions

Jeff Lee - QDA train/test split, confusion matrices, conclusions

Marco Oviedo - Box's M Test, Visualization, conclusions

I. Introduction

In the late 1930s there was a documented count of twenty-one Pima Indians diagnosed with type-2 diabetes[1]. The trend of increased diagnoses continued by almost ten times by the early 1950s. By 1970, almost 40% of Pima Indians that were aged thirty-five or older were diagnosed with type 2 diabetes[1]. In addition, they displayed higher obesity levels among similar aged groups of Caucasians[1]. The significant increase in diagnoses of type-2 diabetes among the Pima Indian population has become a public health concern.

To gain a better insight into this health concern, this report will focus on the Diabetes test results collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The data set focuses specifically on a population of women that are of Pima Indian heritage living near Phoenix, Arizona. We aim to create an early detection model using Quadratic Discriminant Analysis to predict the diagnosis of type-2 Diabetes in the Pima Indian Population using the health factors in the dataset.

II. Data Analysis

The data set we will be using consists of 768 observations on 9 variables regarding the female Pima Indian population. The 9 variables are the health parameters: number of times the individual was pregnant, plasma glucose concentration, blood pressure, triceps skin fold thickness, insulin levels, body mass index, diabetes pedigree function, age, and if they have diabetes.

Out of the 768 female Pima Indians in the dataset, only 268 were positive for diabetes while 500 were negative for diabetes. Both of these classes tended to have either normal distribution or skewed to the right. As shown in the figure above, we can see that pressure and mass had a normal distribution while pregnancy, triceps, insulin, pedigree and age were skewed to the right. Looking at the scatter plots we can see that for some variables such as glucose and triceps the separation of the two classes is more clear. Many of the other combinations of variables have no distinct separation between the two classes for diabetes. The variable combinations with a clear distinction between the two groups would be more important features to the model.

III. Box's M-Test

To determine whether we have to use Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA) we will perform Box's M Test to test if the covariance matrices for the Positive and Negative class for diabetes are equal to each other. This test is to help us know if the assumptions for LDA are met. The Null and Alternative Hypothesis for Box's M Test are as follows below:

$$H_0: Covariance\ Matrices\ are\ Equal\ \ \Sigma_1 = \Sigma_2$$
$$H_A: Covariance\ Matrices\ are\ Not\ Equal\ \ \Sigma_1 \neq \Sigma_2$$

After running the test in R, we find that the Chi-Square Value is 226.71, with 36 degrees of freedom. The resulting p-value from the test was 2.2e-16. Since we have a significantly small p-value we reject the null hypothesis at any reasonable level of significance (1%, 5%, 10%) and conclude that the covariance matrices are not equal.

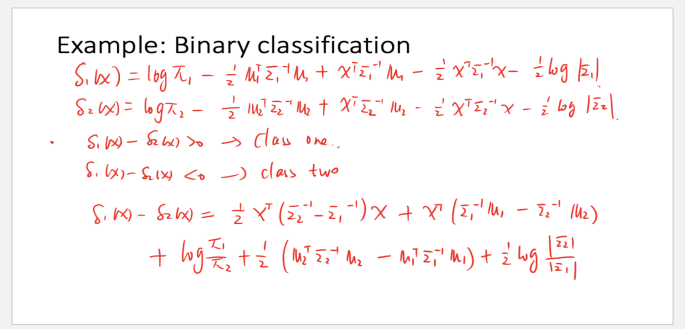
```
```{r}
res = boxM(PimaIndiansDiabetes[, 1:8], PimaIndiansDiabetes[, "diabetes"])
summary(res)
```

Summary for Box's M-test of Equality of Covariance Matrices

Chi-Sq: 226.7065
df: 36
p-value: < 2.2e-16
```

IV. QDA with full dataset

- Since the null hypothesis of equal covariance matrices is rejected, we need to use Quadratic Discriminant Analysis, instead of Linear Discriminant Analysis, for the purpose of classification of each individual as diabetic (positive) or non-diabetic (negative). Since covariance matrices are not equal, this suggests that the decision boundary might be of a quadratic nature. To find the decision boundary with QDA, we used the following equation from lecture 7-2.



- The decision boundary is given by the equation on the bottom, where S1(x) - S2(x) = 0, or in other words, P(k = 1 | X) - P(k = 2 | X) = 0. If this difference is greater than 0, the observation most likely belongs to class 1 (diabetic). If the difference is negative, the observation most likely belongs to class 2 (non-diabetic). To calculate this function in R, we extracted two subsets of the dataframe, one for each class, and converted each to a matrix. We used these matrices to calculate the value of the sample means and covariance matrices. Furthermore, we calculated the proportion of individuals in each class to obtain the values of π_1 and π_2, for the quantity $\log(\pi_1/\pi_2)$. Putting all these together, we built a function to manually calculate the decision boundary.
- **DECISION BOUNDARY**: The snippet of code below contains the function we wrote to calculate the decision boundary. We wrote coeff_one and coeff_two to simplify the expression a little bit, but the "value" variable is equal to P(k = 1 | X) - P(k = 2 | X), and the "constant" variable refers to all parts of the decision boundary expression that do not involve the observation X. If value > 0, the model predicts that the person has diabetes. If value < 0, that indicates the individual most likely does not have diabetes. This function simply returns either "pos" or "neg".

-

```
# Classification using the decision boundary
coeff_one = 0.5*(sigma2_inverse - sigma1_inverse)
coeff_two = (sigma1_inverse%*%mu1 - sigma2_inverse%*%mu2)

# x is an 8 by 1 vector containing health data on an individual
classifyDiabetes <- function(x) {
  value = ((t(x) %*%coeff_one%*%x) + (t(x)%*%coeff_two) + constant)

  answer = ""
  if (value > 0) {
    answer = "pos"
  }
  if (value < 0) {
    answer = "neg"
  }
  return(answer)
}
```

Confusion matrix: QDA with full dataset

Accuracy rate: 587/768 = 76% correctly predicted

	Actual positive	Actual negative
Predicted positive	155	68
Predicted negative	113	432

Confusion matrix: QDA with train/test split

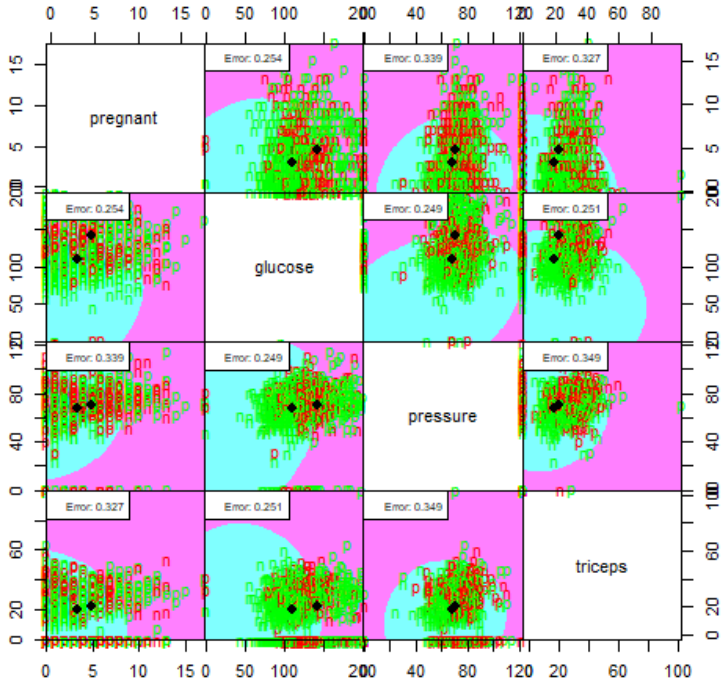
- We also evaluated the performance of QDA after splitting our data into train and test sets. Using the decision boundary formula from above, modeled using the train set, we create the following confusion matrix.

Accuracy rate: 112/153 = 73% correctly predicted

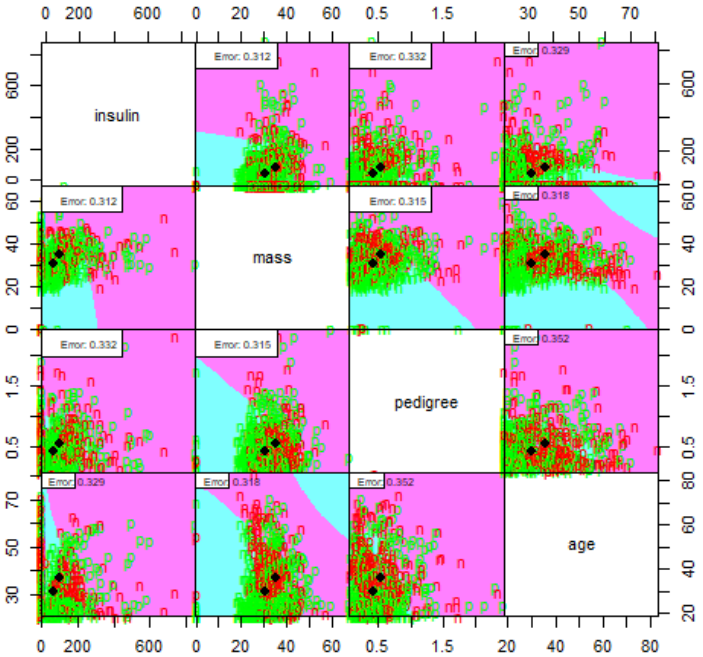
	Actual positive	Actual negative
Predicted positive	30	18
Predicted negative	23	82

QDA Partition Plots

To visualize the performance and classification of the QDA model on the data, we used QDA partition plots. Each plot below will show the classification outcomes for the variables in the Pima Indian data set. This will help us identify which combination of variables perform well with QDA.



Looking at the first QDA partition plot we can see that most of the variable combinations have an error rate of 0.3 or more. This means that 30% of the labels are misclassified by the QDA model. The combination of variables with the lowest error rate of 24.9% is Glucose and Pressure. The worst performing variables with an error rate of 34.9% was the combination of Pressure and Triceps.



In the second partition plot we can see that all the variable combinations have an error rate that is greater than 30%. The variable combination that performed the best with the QDA model was Insulin and Mass with an error rate of 31.2%. The worst performing combination of variables, not only in the second plot but overall, was Pedigree and Age with an error rate of 35.2%.

V. Conclusions

The QDA model on the full dataset has a 76% correct prediction rate, while the QDA model on the test dataset after the train/test split (80% of data in train set, 20% in test set), has a 73% correct prediction rate. This slight drop in accuracy is due to the possibility that we were likely overfitting our model by training on the full dataset. The data is also slightly imbalanced with only 35% of observations belonging to the positive class. This could be one reason why the model does not do well when predicting positive cases. In real world scenarios, we should use a similar train/test split to avoid overfitting on our data. This is because a model that has been trained on the entire given dataset might not be very accurate on a set of new, previously unseen data points.

References

[1] Leslie O. Schulz and Lisa S. Chaudhari. “High-Risk Populations: The Pimas of Arizona and Mexico”. In: web-article,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/R14 (2015).