LDA and QDA analysis

Michael Kamp

September 15, 2025

Introduction

This assignment continues Problem 14 from Chapter 4.
I will use the same training/testing split from Week 1 to ensure consistency.
The goal is to use Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) to predict whether a car's `mpg` is above or below the median (`mpg01`) based on variables identified in Problem 14(b).

Data Preparation (from Week 1)

```r
# Load necessary libraries
library(ISLR)   # for the Auto dataset
library(MASS)   # for LDA and QDA

# Load the Auto dataset
data(Auto)

# Create the binary variable mpg01
mpg01 <- ifelse(Auto$mpg > median(Auto$mpg), 1, 0)
Auto$mpg01 <- mpg01

# Set up train/test split
set.seed(1)
train <- sample(1:nrow(Auto), nrow(Auto)/2)
test <- -train
```

Problem 14(d) – Linear Discriminant Analysis (LDA)

Question:
Perform LDA using the variables most associated with mpg01 from 14(b). Report the test error.

```r
# Fit LDA model
lda_model <- lda(mpg01 ~ cylinders + displacement + horsepower + weight +
acceleration + year,
                 data = Auto, subset = train)

# Predict on test data
```

```r
lda_pred <- predict(lda_model, Auto[test, ])

# Confusion matrix
lda_confusion <- table(lda_pred$class, Auto$mpg01[test])
lda_confusion
```

```
##
##      0  1
##   0 81  4
##   1 21 90
```

```r
# Test error rate
lda_error <- mean(lda_pred$class != Auto$mpg01[test])
lda_error
```

```
## [1] 0.127551
```

Explanation:
- The confusion matrix shows how many test observations were correctly or incorrectly classified.
- Test error rate: 0.127551.
- LDA assumes linear decision boundaries between the two classes (`mpg01 = 0` or `1`).
- Using the six key variables (`cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`) helps predict whether a car's mpg is above the median.
- This model provides a baseline for comparison with QDA.

Problem 14(e) – Quadratic Discriminant Analysis (QDA)

Question:
Perform QDA using the variables most associated with mpg01 from 14(b). Report the test error.

```r
# Fit QDA model
qda_model <- qda(mpg01 ~ cylinders + displacement + horsepower + weight +
acceleration + year,
                 data = Auto, subset = train)

# Predict on test data
qda_pred <- predict(qda_model, Auto[test, ])

# Confusion matrix
qda_confusion <- table(qda_pred$class, Auto$mpg01[test])
qda_confusion
```

```
##
##      0  1
```

```
##   0 89  6
##   1 13 88

# Test error rate
qda_error <- mean(qda_pred$class != Auto$mpg01[test])
qda_error

## [1] 0.09693878
```

Explanation:
- The confusion matrix shows how many test observations were correctly or incorrectly classified by the QDA model.
- Test error rate: 0.0969388.
- QDA allows for non-linear decision boundaries, which can better capture complex relationships in the data compared to LDA.
- Using the same six key variables from 14(b) allows a direct comparison of LDA and QDA performance.
- Comparing the test errors shows which method better predicts whether a car's mpg is above or below the median.
- This analysis highlights the relative importance of `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, and `year` in predicting mpg classification.