LDA QDA Classification

2025-05-04

```
library(MASS)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ggplot2)

wine_data <- read.csv("/Users/isabellachen/Downloads/wine-quality-white-and-red.csv")

wine_data$type <- factor(wine_data$type)

wine_data$quality_bin <- factor(
  ifelse(wine_data$quality >= 7, "1", "0"),
  levels = c("0","1")
)

# Split into 70/30 train/test
set.seed(1)
train_idx   <- createDataPartition(wine_data$quality_bin, p = 0.7, list = FALSE)
train_data  <- wine_data[train_idx, ]
test_data   <- wine_data[-train_idx, ]

fmla <- quality_bin ~ . - quality

# 4. Fit LDA and QDA
lda_model <- lda(fmla, data = train_data)
qda_model <- qda(fmla, data = train_data)

# 5. Predict on test set
lda_pred <- predict(lda_model, test_data)
qda_pred <- predict(qda_model, test_data)
```

```
# Evaluate LDA model
lda_conf <- confusionMatrix(lda_pred$class, test_data$quality_bin)
cat("LDA Model Performance:\n")
```

```
## LDA Model Performance:
```

```
print(lda_conf)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1456  257
##          1  110  126
##
##                Accuracy : 0.8117
##                  95% CI : (0.7936, 0.8288)
##     No Information Rate : 0.8035
##     P-Value [Acc > NIR] : 0.1888
##
##                   Kappa : 0.3026
##
##  Mcnemar's Test P-Value : 2.514e-14
##
##             Sensitivity : 0.9298
##             Specificity : 0.3290
##          Pos Pred Value : 0.8500
##          Neg Pred Value : 0.5339
##              Prevalence : 0.8035
##          Detection Rate : 0.7470
##    Detection Prevalence : 0.8789
##       Balanced Accuracy : 0.6294
##
##        'Positive' Class : 0
##
```

```
# Evaluate QDA model
qda_conf <- confusionMatrix(qda_pred$class, test_data$quality_bin)
cat("QDA Model Performance:\n")
```

```
## QDA Model Performance:
```

```
print(qda_conf)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1220  129
##          1  346  254
##
##                Accuracy : 0.7563
##                  95% CI : (0.7366, 0.7752)
##     No Information Rate : 0.8035
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3643
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7791
```

```
##             Specificity : 0.6632
##          Pos Pred Value : 0.9044
##          Neg Pred Value : 0.4233
##              Prevalence : 0.8035
##          Detection Rate : 0.6260
##    Detection Prevalence : 0.6921
##       Balanced Accuracy : 0.7211
##
##        'Positive' Class : 0
##
```

```
# Compare models
models <- c("LDA", "QDA")
accuracies <- c(
  lda_conf$overall["Accuracy"],
  qda_conf$overall["Accuracy"]
)
comparison <- data.frame(Model = models, Accuracy = accuracies)
print(comparison)
```

```
##   Model  Accuracy
## 1   LDA 0.8116983
## 2   QDA 0.7562853
```

```
ctrl <- trainControl(method = "cv", number = 10)

set.seed(1)
lda_cv <- train(
  quality_bin ~ . - quality,
  data      = train_data,
  method    = "lda",
  trControl = ctrl
)

set.seed(1)
qda_cv <- train(
  quality_bin ~ . - quality,
  data      = train_data,
  method    = "qda",
  trControl = ctrl
)
print(lda_cv)
```

```
## Linear Discriminant Analysis
##
## 4548 samples
##   13 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4093, 4094, 4094, 4093, 4093, 4094, ...
## Resampling results:
```

```
##
##   Accuracy   Kappa
##   0.8205743  0.3176433
```

```
print(qda_cv)
```

```
## Quadratic Discriminant Analysis
##
## 4548 samples
##   13 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4093, 4094, 4094, 4093, 4093, 4094, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7673704  0.3942065
```

```
library(caret)  # ensure caret is loaded
lda_preds <- predict(lda_cv, newdata = test_data)
lda_preds <- factor(lda_preds, levels = levels(test_data$quality_bin))
confusionMatrix(data=lda_preds, reference = test_data$quality_bin)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1456  257
##          1  110  126
##
##                Accuracy : 0.8117
##                  95% CI : (0.7936, 0.8288)
##     No Information Rate : 0.8035
##     P-Value [Acc > NIR] : 0.1888
##
##                   Kappa : 0.3026
##
##  Mcnemar's Test P-Value : 2.514e-14
##
##             Sensitivity : 0.9298
##             Specificity : 0.3290
##          Pos Pred Value : 0.8500
##          Neg Pred Value : 0.5339
##              Prevalence : 0.8035
##          Detection Rate : 0.7470
##    Detection Prevalence : 0.8789
##       Balanced Accuracy : 0.6294
##
##        'Positive' Class : 0
##
```

```r
qda_preds <- predict(qda_cv, newdata = test_data)
qda_preds <- factor(qda_preds, levels = levels(test_data$quality_bin))
confusionMatrix(data=qda_preds,reference = test_data$quality_bin)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1220  129
##          1  346  254
##
##                Accuracy : 0.7563
##                  95% CI : (0.7366, 0.7752)
##     No Information Rate : 0.8035
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3643
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7791
##             Specificity : 0.6632
##          Pos Pred Value : 0.9044
##          Neg Pred Value : 0.4233
##              Prevalence : 0.8035
##          Detection Rate : 0.6260
##    Detection Prevalence : 0.6921
##       Balanced Accuracy : 0.7211
##
##        'Positive' Class : 0
##
```

```r
lda_test_acc <- confusionMatrix(
  predict(lda_cv, newdata = test_data),
  test_data$quality_bin
)$overall["Accuracy"]

qda_test_acc <- confusionMatrix(
  predict(qda_cv, newdata = test_data),
  test_data$quality_bin
)$overall["Accuracy"]

data.frame(
  Model       = c("LDA","QDA"),
  CV_Accuracy = c(lda_cv$results$Accuracy,  qda_cv$results$Accuracy),
  Test_Accuracy = c(lda_test_acc, qda_test_acc)
)
```

```
##   Model CV_Accuracy Test_Accuracy
## 1   LDA   0.8205743     0.8116983
## 2   QDA   0.7673704     0.7562853
```