705604096_stats101c_hw3

Jade Gregory 705604096

2023-10-23

Question 1

```
winetrain <- read.csv("WineTrain.csv")
winetest <- read.csv("WineTest.csv")
winetrain$Class <- as.factor(winetrain$Class)
winetrain$Wine.Color <- as.factor(winetrain$Wine.Color)
winetest$Class <- as.factor(winetest$Class)
winetest$Wine.Color <- as.factor(winetest$Wine.Color)
head(winetest)
```

```
##   X Wine.Color fixed.acidity volatile.acidity citric.acid residual.sugar
## 1 1          R           8.6             0.45        0.31            2.6
## 2 2          W           6.9             0.25        0.24            3.6
## 3 3          W           8.1             0.20        0.28            0.9
## 4 4          R           7.8             0.56        0.19            2.1
## 5 5          W           6.3             0.35        0.30            5.7
## 6 6          W           6.0             0.27        0.15            1.5
##   chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 1     0.086                 21                  50 0.99820 3.37      0.91
## 2     0.057                 13                  85 0.99420 2.99      0.48
## 3     0.023                 49                  87 0.99062 2.92      0.36
## 4     0.081                 15                 105 0.99620 3.33      0.54
## 5     0.035                  8                  97 0.99270 3.27      0.41
## 6     0.056                 35                 128 0.99360 3.12      0.45
##   alcohol Class
## 1     9.9  Good
## 2     9.5   Bad
## 3    11.1  Good
## 4     9.5   Bad
## 5    11.0  Good
## 6     8.8  Good
```

```
head(winetrain)
```

```
##   X Wine.Color fixed.acidity volatile.acidity citric.acid residual.sugar
## 1 1          W           7.3             0.23        0.41           14.6
## 2 2          R          10.0             0.32        0.59            2.2
## 3 3          W           6.2             0.27        0.43            7.8
## 4 4          W           6.6             0.25        0.32            5.6
```

```
## 5 5          W           6.9             0.24        0.39            1.3
## 6 6          W           7.1             0.23        0.39            1.6
##   chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 1     0.048                  73                  223 0.99863 3.16      0.71
## 2     0.077                   3                   15 0.99940 3.20      0.78
## 3     0.056                  48                  244 0.99560 3.10      0.51
## 4     0.039                  15                   68 0.99163 2.96      0.52
## 5     0.063                  18                  136 0.99280 3.31      0.48
## 6     0.032                  12                   65 0.98980 3.25      0.40
##   alcohol Class
## 1     9.4   Bad
## 2     9.6   Bad
## 3     9.0   Bad
## 4    11.1  Good
## 5    10.4  Good
## 6    12.7  Good
```

a)

```
# wine training data glm model
winetrainglm <- glm(Class ~ . - X - Class, data = winetrain, family = binomial())
summary(winetrainglm)
```

```
##
## Call:
## glm(formula = Class ~ . - X - Class, family = binomial(), data = winetrain)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.022e+02  7.014e+01   2.883 0.003938 **
## Wine.ColorW          -2.540e-01  2.691e-01  -0.944 0.345231
## fixed.acidity         2.212e-01  7.336e-02   3.015 0.002569 **
## volatile.acidity     -1.280e+00  3.656e-01  -3.500 0.000465 ***
## citric.acid           3.216e-01  3.445e-01   0.933 0.350576
## residual.sugar        9.878e-02  2.803e-02   3.524 0.000425 ***
## chlorides            -3.847e+00  1.601e+00  -2.403 0.016246 *
## free.sulfur.dioxide   8.239e-03  3.407e-03   2.418 0.015596 *
## total.sulfur.dioxide -3.128e-03  1.446e-03  -2.164 0.030486 *
## density              -2.127e+02  7.106e+01  -2.993 0.002759 **
## pH                    1.674e+00  4.141e-01   4.042 5.30e-05 ***
## sulphates             1.433e+00  3.492e-01   4.105 4.05e-05 ***
## alcohol               1.810e-01  8.768e-02   2.064 0.038999 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3881.2  on 2799  degrees of freedom
## Residual deviance: 3603.1  on 2787  degrees of freedom
## AIC: 3629.1
##
## Number of Fisher Scoring iterations: 4
```

```r
# wine training data confusion matrix
winetrainprob <- predict(winetrainglm, data = winetrain, type = "response")
wtrainpredl <- rep("Good", length(winetrainprob))
wtrainpredl[winetrainprob <= 0.5] <- "Bad"
table(wtrainpredl, winetrain$Class)
```

```
##
## wtrainpredl Bad Good
##        Bad  937  520
##        Good 481  862
```

```r
# misclassification rate
mean(wtrainpredl != winetrain$Class)
```

```
## [1] 0.3575
```

The misclassification rate of the wine training data for the glm model is 35.75%.

```r
# wine testing data confusion matrix
winetestprob <- predict(winetrainglm, data = winetrain, newdata = winetest, type = "response")
wtestpredl <- rep("Good", length(winetestprob))
wtestpredl[winetestprob <= 0.5] <- "Bad"
table(wtestpredl, winetest$Class)
```

```
##
## wtestpredl Bad Good
##       Bad  427  212
##       Good 205  356
```

```r
# misclassification rate
mean(wtestpredl != winetest$Class)
```

```
## [1] 0.3475
```

The misclassification rate of the wine testing data for the glm model is 34.75%.

b)

```r
# lda model
winelda <- lda(Class ~ . - X - Class, data = winetrain)
winelda
```

```
## Call:
## lda(Class ~ . - X - Class, data = winetrain)
##
## Prior probabilities of groups:
##       Bad      Good
## 0.5064286 0.4935714
##
## Group means:
```

```
##      Wine.ColorW fixed.acidity volatile.acidity citric.acid residual.sugar
## Bad    0.7320169      7.263047        0.3584908   0.3123836       5.640197
## Good   0.7858177      7.161505        0.3189146   0.3273444       5.176773
##       chlorides free.sulfur.dioxide total.sulfur.dioxide   density       pH
## Bad  0.06012623            30.29055             117.7475 0.9951660 3.210063
## Good 0.05074096            30.90630             113.7218 0.9940833 3.227366
##      sulphates  alcohol
## Bad  0.5253597 10.20609
## Good 0.5359624 10.82847
##
## Coefficients of linear discriminants:
##                                LD1
## Wine.ColorW          -3.849040e-01
## fixed.acidity         3.535272e-01
## volatile.acidity     -1.977917e+00
## citric.acid           5.221028e-01
## residual.sugar        1.572735e-01
## chlorides            -5.317770e+00
## free.sulfur.dioxide   1.265154e-02
## total.sulfur.dioxide -4.921709e-03
## density              -3.427890e+02
## pH                    2.645963e+00
## sulphates             2.225004e+00
## alcohol               2.805697e-01
```

```
library(caret)
```

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

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

```
# training data confusion matrix
t1 <- table(predict(winelda)$class, winetrain$Class)
print(confusionMatrix(t1))
```

```
## Confusion Matrix and Statistics
##
##
##        Bad Good
##   Bad  940  526
##   Good 478  856
##
##                Accuracy : 0.6414
##                  95% CI : (0.6233, 0.6592)
##     No Information Rate : 0.5064
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.2824
##
##  Mcnemar's Test P-Value : 0.138
##
##             Sensitivity : 0.6629
```

```
##             Specificity : 0.6194
##          Pos Pred Value : 0.6412
##          Neg Pred Value : 0.6417
##              Prevalence : 0.5064
##          Detection Rate : 0.3357
##    Detection Prevalence : 0.5236
##       Balanced Accuracy : 0.6411
##
##        'Positive' Class : Bad
##
```

```
# training data misclassification rate
winepred1 <- predict(winelda, winetrain)
wine.classify <- winepred1$class
wine.classperc <- mean(wine.classify != winetrain$Class)
wine.classperc
```

```
## [1] 0.3585714
```

The misclassification rate for the wine training data for the lda model is 35.857%.

```
# testing data confusion matrix
t2 <- table(predict(winelda, winetest)$class, winetest$Class)
print(confusionMatrix(t2))
```

```
## Confusion Matrix and Statistics
##
##
##        Bad Good
##   Bad  428  215
##   Good 204  353
##
##                Accuracy : 0.6508
##                  95% CI : (0.6231, 0.6778)
##     No Information Rate : 0.5267
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.299
##
##  Mcnemar's Test P-Value : 0.6252
##
##             Sensitivity : 0.6772
##             Specificity : 0.6215
##          Pos Pred Value : 0.6656
##          Neg Pred Value : 0.6338
##              Prevalence : 0.5267
##          Detection Rate : 0.3567
##    Detection Prevalence : 0.5358
##       Balanced Accuracy : 0.6493
##
##        'Positive' Class : Bad
##
```

```r
# misclassification rate for testing data
winepred2 <- predict(winelda, winetest)
wine.classify2 <- winepred2$class
wine.classperc2 <- mean(wine.classify2 != winetest$Class)
wine.classperc2
```

```
## [1] 0.3491667
```

The misclassification rate for wine testing data for the lda model is 34.916%.

c)

```r
# qda model
wineqda <- qda(Class ~ . - X - Class, data = winetrain, method = "mle")
wineqda
```

```
## Call:
## qda(Class ~ . - X - Class, data = winetrain, method = "mle")
##
## Prior probabilities of groups:
##       Bad      Good
## 0.5064286 0.4935714
##
## Group means:
##      Wine.ColorW fixed.acidity volatile.acidity citric.acid residual.sugar
## Bad    0.7320169      7.263047        0.3584908   0.3123836       5.640197
## Good   0.7858177      7.161505        0.3189146   0.3273444       5.176773
##       chlorides free.sulfur.dioxide total.sulfur.dioxide   density       pH
## Bad  0.06012623            30.29055             117.7475 0.9951660 3.210063
## Good 0.05074096            30.90630             113.7218 0.9940833 3.227366
##      sulphates  alcohol
## Bad  0.5253597 10.20609
## Good 0.5359624 10.82847
```

```r
# training confusion matrix
wqdapred <- predict(wineqda)$class
table(wqdapred, winetrain$Class)
```

```
##
## wqdapred  Bad Good
##     Bad   718  307
##     Good  700 1075
```

```r
# misclassification rate
mean(winetrain$Class != wqdapred)
```

```
## [1] 0.3596429
```

The misclassification rate for the qda model of the training data is 35.96%.

```r
# testing confusion matrix
wqdapred2 <- predict(wineqda, winetest)$class
table(wqdapred2, winetest$Class)
```

```
##
## wqdapred2 Bad Good
##      Bad  318  144
##      Good 314  424
```

```r
# misclassification rate
mean(winetest$Class != wqdapred2)
```

```
## [1] 0.3816667
```

The misclassification rate for the qda model for the testing data is 38.16%.

d)

```r
library(class)
```

```r
set.seed(113355)
winetestX <- winetest[-c(1,14)]
winetrainX <- winetrain[-c(1,14)]
winetestY <- winetest$Class
winetrainY <- winetrain$Class
winetrainX[-c(1)] <- scale(winetrainX[-c(1)])
winetestX[, -1] <- scale(winetestX[, -1])
winetrainX$Wine.Color <- as.numeric(winetrainX$Wine.Color)
winetestX$Wine.Color <- as.numeric(winetestX$Wine.Color)
wineknn <- knn(winetrainX, winetestX, winetrainY, k = 25)
table(wineknn, winetestY)
```

```
##        winetestY
## wineknn Bad Good
##    Bad  405  194
##    Good 227  374
```

```r
mean(wineknn != winetestY)
```

```
## [1] 0.3508333
```

The misclassification rate for the knn model is 35.08%.

e) Since the misclassification rate of the wine testing data for the glm model is 34.75%, I would argue that this is our best model. This is because the misclassification rate is the lowest among all of the models we produced. All of our models had misclassification rates that were near 30%, and they appeared to produce similar misclassification rates close in value.

Question 2

```r
olives <- read.csv("Olives.csv")
olives$region <- as.factor(olives$region)
olives$area <- as.factor(olives$area)
set.seed(1234567)
i = 1:dim(olives)[1]
i.train <- sample(i, 400, replace = FALSE)
O.train <- olives[i.train,]
O.test <- olives[-i.train,]
O.testY <- O.test$region
```

a)

```r
library(e1071)
```

```r
olive.nb <- naiveBayes(region ~ . - X - region, data = O.train)
olive.nb
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##      1      2      3
## 0.5725 0.1800 0.2475
##
## Conditional probabilities:
##    area
## Y    Calabria  Coast-Sardinia  East-Liguria  Inland-Sardinia  North-Apulia
##   1 0.18340611      0.00000000    0.00000000       0.00000000    0.05240175
##   2 0.00000000      0.31944444    0.00000000       0.68055556    0.00000000
##   3 0.00000000      0.00000000    0.33333333       0.00000000    0.00000000
##    area
## Y      Sicily  South-Apulia     Umbria  West-Liguria
##   1 0.12227074    0.64192140 0.00000000    0.00000000
##   2 0.00000000    0.00000000 0.00000000    0.00000000
##   3 0.00000000    0.00000000 0.31313131    0.35353535
##
##    palmitic
## Y       [,1]      [,2]
##   1 1341.371 144.25424
##   2 1109.458  39.70558
##   3 1100.303  78.35730
##
##    palmitoleic
## Y        [,1]     [,2]
##   1 156.32751 49.72011
##   2  96.72222 14.11843
##   3  83.30303 23.24161
##
```

```
##    stearic
## Y       [,1]     [,2]
##   1 227.1179 38.96838
##   2 226.2778 16.77626
##   3 232.8788 38.98515
##
##    oleic
## Y       [,1]     [,2]
##   1 7085.026 323.2466
##   2 7272.375 142.4920
##   3 7780.768 155.7132
##
##    linoleic
## Y        [,1]     [,2]
##   1 1040.4847 205.4955
##   2 1193.1111 107.1097
##   3  735.1414 141.9307
##
##    linolenic
## Y       [,1]      [,2]
##   1 38.21834  8.265093
##   2 26.90278  5.259893
##   3 21.25253 17.182302
##
##    arachidic
## Y       [,1]     [,2]
##   1 62.91703 11.34070
##   2 73.05556 11.74361
##   3 36.68687 30.69955
##
##    eicosenoic
## Y        [,1]      [,2]
##   1 26.820961 8.6042776
##   2  1.888889 0.7229742
##   3  1.909091 0.7297037
```

```
olive.fit <- predict(olive.nb, O.test)
table(olive.fit, O.test$region)
```

```
##
## olive.fit  1  2  3
##         1 89  0  0
##         2  0 26  0
##         3  5  0 52
```

```
mean(olive.fit != O.testY)
```

```
## [1] 0.02906977
```

The misclassification rate of our testing data is 2.9%.

b)

```
oliveslda <- lda(region ~ . - X - region, data = O.train)
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear
```

```
t_olive <- table(predict(oliveslda, O.test)$class, O.test$region)
print(confusionMatrix(t_olive))
```

```
## Confusion Matrix and Statistics
##
##
##      1  2  3
##   1 94  0  0
##   2  0 26  0
##   3  0  0 52
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9788, 1)
##     No Information Rate : 0.5465
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            1.0000   1.0000   1.0000
## Specificity            1.0000   1.0000   1.0000
## Pos Pred Value         1.0000   1.0000   1.0000
## Neg Pred Value         1.0000   1.0000   1.0000
## Prevalence             0.5465   0.1512   0.3023
## Detection Rate         0.5465   0.1512   0.3023
## Detection Prevalence   0.5465   0.1512   0.3023
## Balanced Accuracy      1.0000   1.0000   1.0000
```

```
olivespred <- predict(oliveslda, O.test)
olives.classify <- olivespred$class
mean(olives.classify != O.test$region)
```

```
## [1] 0
```

The misclassification rate is 0%.

c)

```
oliveslda
```

```
## Call:
## lda(region ~ . - X - region, data = O.train)
##
## Prior probabilities of groups:
##      1      2      3
## 0.5725 0.1800 0.2475
##
## Group means:
##   areaCoast-Sardinia  areaEast-Liguria  areaInland-Sardinia  areaNorth-Apulia
## 1          0.0000000         0.0000000            0.0000000        0.05240175
## 2          0.3194444         0.0000000            0.6805556        0.00000000
## 3          0.0000000         0.3333333            0.0000000        0.00000000
##   areaSicily  areaSouth-Apulia  areaUmbria  areaWest-Liguria  palmitic
## 1   0.1222707         0.6419214   0.0000000         0.0000000 1341.371
## 2   0.0000000         0.0000000   0.0000000         0.0000000 1109.458
## 3   0.0000000         0.0000000   0.3131313         0.3535354 1100.303
##   palmitoleic  stearic    oleic  linoleic linolenic arachidic eicosenoic
## 1   156.32751 227.1179 7085.026 1040.4847  38.21834  62.91703  26.820961
## 2    96.72222 226.2778 7272.375 1193.1111  26.90278  73.05556   1.888889
## 3    83.30303 232.8788 7780.768  735.1414  21.25253  36.68687   1.909091
##
## Coefficients of linear discriminants:
##                                LD1          LD2
## areaCoast-Sardinia   -0.651828788  1.278422491
## areaEast-Liguria     -0.134797838  0.502428409
## areaInland-Sardinia   0.651828788 -1.278422491
## areaNorth-Apulia     -2.396916072 -0.949239757
## areaSicily           -1.436300054  0.592584527
## areaSouth-Apulia     -3.507927484  2.089896331
## areaUmbria            1.229762200 -1.167751246
## areaWest-Liguria     -1.026227244  0.610415424
## palmitic              0.007108048  0.005477706
## palmitoleic           0.005271761  0.005050268
## stearic               0.010372433  0.002062706
## oleic                 0.010913925  0.004842669
## linoleic              0.015054351 -0.004535588
## linolenic            -0.051294858  0.007803169
## arachidic             0.038397315 -0.027860988
## eicosenoic           -0.094789705 -0.046556254
##
## Proportion of trace:
##    LD1    LD2
## 0.7628 0.2372
```

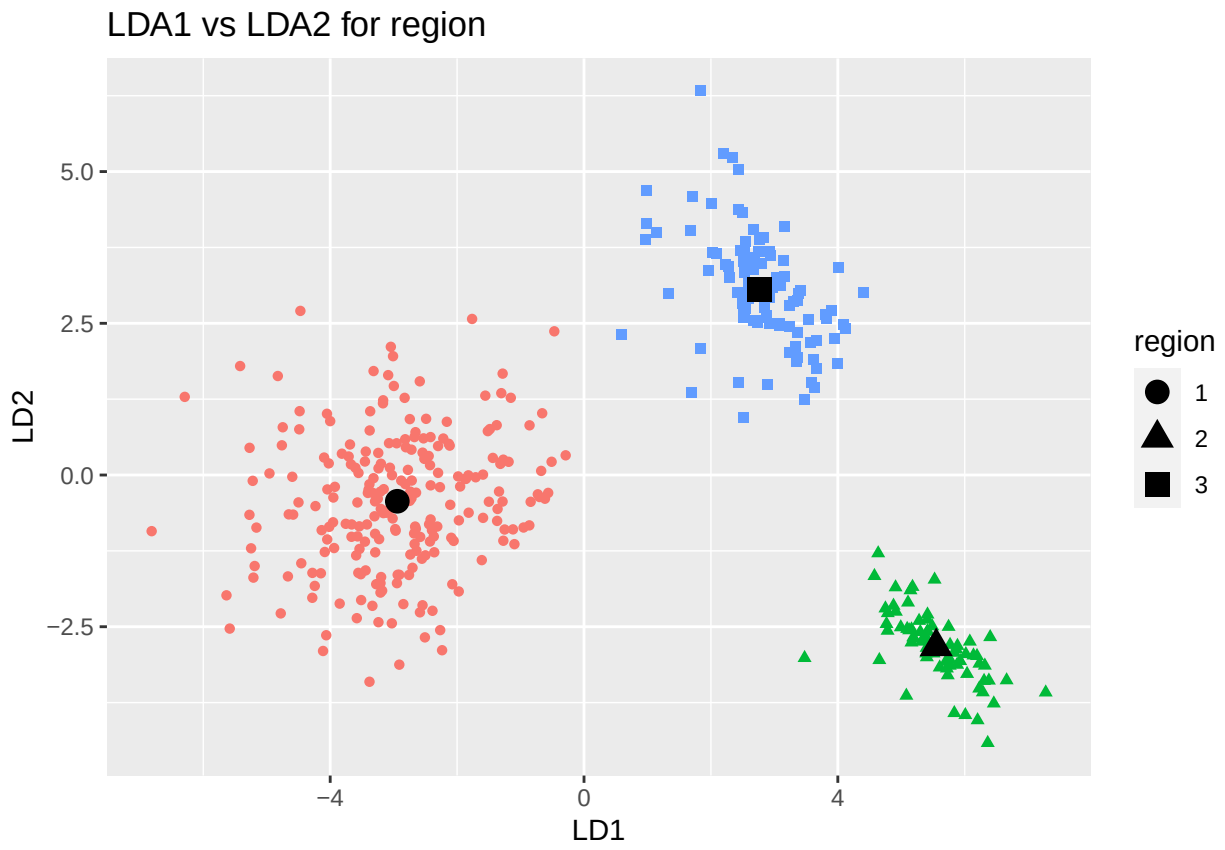
The proportion of tracee for LD1 is 0.7628, and the proportion of trace for LD2 is 0.2372.

d)

```
library(ggplot2)
```

```
LD1 <- predict(oliveslda)$x[,1]
LD2 <- predict(oliveslda)$x[,2]
centroids <- aggregate(data = O.train, cbind(LD1, LD2) ~ region, mean)
```

```
cenplot <- ggplot(data = O.train, aes(LD1, LD2, colour = region, shape = region)) + geom_point()
cenplot + ggtitle("LDA1 vs LDA2 for region") + geom_point(size = 1) + geom_point(data = centroids, size
```



e)

```
O.train$area <- as.numeric(O.train$area)
O.test$area <- as.numeric(O.test$area)
# qda model
olivesqda <- qda(region ~ . - X - region, data = O.train, method = "mle")
# testing confusion matrix
oqdapred <- predict(olivesqda, O.test)$class
table(oqdapred, O.test$region)
```

```
##
## oqdapred  1  2  3
##        1 94  0  1
##        2  0 26  0
##        3  0  0 51
```

```
# misclassification rate
mean(O.test$region != oqdapred)
```

```
## [1] 0.005813953
```

The misclassification rate is 0.58%.

f)

```
summary(olivesqda)
```

```
##         Length Class  Mode
## prior     3    -none- numeric
## counts    3    -none- numeric
## means    27    -none- numeric
## scaling 243    -none- numeric
## ldet      3    -none- numeric
## lev       3    -none- character
## N         1    -none- numeric
## call      4    -none- call
## terms     3    terms  call
## xlevels   0    -none- list
```

This is the summary of our QDA model.