HW4-LianZuo

Lian Zuo

2023-11-13

#install.packages("palmerpenguins")  
library(palmerpenguins)  
#head(penguins,3)  
names(penguins)

[1] "species" "island" "bill\_length\_mm"   
[4] "bill\_depth\_mm" "flipper\_length\_mm" "body\_mass\_g"   
[7] "sex" "year"

dim(penguins)

[1] 344 8

new.pengu=data.frame(na.omit(penguins[,c("species", "bill\_length\_mm", "bill\_depth\_mm",  
"flipper\_length\_mm", "body\_mass\_g")]))  
head(new.pengu,3)

species bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
1 Adelie 39.1 18.7 181 3750  
2 Adelie 39.5 17.4 186 3800  
3 Adelie 40.3 18.0 195 3250

names(new.pengu)

[1] "species" "bill\_length\_mm" "bill\_depth\_mm"   
[4] "flipper\_length\_mm" "body\_mass\_g"

dim(new.pengu)

[1] 342 5

#install.packages("mvnormtest")  
# library(mvnormtest)  
# mshapiro.test(t(new.pengu[,2:5]))

# 1.

Check the assumption of homogeneity of covariance matrix. Write down the conclusion of the test.

#install.packages("heplots")  
library(car)  
library(carData)  
library(heplots)  
boxM(new.pengu[,2:5],new.pengu[,1])

Box's M-test for Homogeneity of Covariance Matrices  
  
data: new.pengu[, 2:5]  
Chi-Sq (approx.) = 76.795, df = 20, p-value = 1.365e-08

Conclusion:The test statistic value is found to be 76.795 with a p-value = 1.365e-08. Therefore, we reject the null hypothesis and conclude that the population covariance matrices are not significantly different,so the data shows evidence of heterogeneity of covariance matrices at 5% level of significance (p-value = 1.365e-08).

# 2.

Based on the conclusion in part 1, what type of discriminant function would you consider to predict group membership of penguins?

Conclusion:Based on the conclusion in part 1, we can apply quadratic discriminant analysis instead of linear discriminant analysis for the data.

# 3.

1. Split the data into training set and testing set in the ratio 80:20. Use seed as 1234.

# Split the data into training set (80%) and test set (20%)  
set.seed(1234)  
index=sample(1:nrow(new.pengu), floor(0.8\*nrow(new.pengu)), replace=F)  
#head(index,3)  
train.data=new.pengu[index,]  
head(train.data,3)

species bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
284 Chinstrap 51.3 19.9 198 3700  
336 Chinstrap 46.8 16.5 189 3650  
101 Adelie 41.0 20.0 203 4725

test.data=new.pengu[-index,]  
head(test.data,3)

species bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
1 Adelie 39.1 18.7 181 3750  
3 Adelie 40.3 18.0 195 3250  
8 Adelie 34.1 18.1 193 3475

1. Scale the variables if necessary.

#install. packages("tidyverse")  
library(tidyverse) # to use select() function  
#library(dplyr)  
train.trs=data.frame(scale(select(train.data, where(is.numeric))))#scaling all numeric variables  
test.trs=data.frame(scale(select(test.data, where(is.numeric))))  
train.trs$Species=train.data$species  
test.trs$Species=test.data$species  
head(train.trs,3)

bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g Species  
284 1.3468026 1.3364828 -0.1554657 -0.5825673 Chinstrap  
336 0.5413959 -0.4058253 -0.7952967 -0.6457414 Chinstrap  
101 -0.4966838 1.3877271 0.1999960 0.7125024 Adelie

head(test.trs,3)

bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g Species  
1 -1.098850 1.0622785 -1.6557246 -0.7394392 Adelie  
3 -0.854792 0.7091604 -0.6389223 -1.3428846 Adelie  
8 -2.115758 0.7596059 -0.7841798 -1.0713342 Adelie

1. Perform the appropriate discriminant analysis.

#install.packages("MASS")  
#library(MASS)  
qda.pengu=MASS::qda(Species~., data=train.trs)  
qda.pengu

Call:  
qda(Species ~ ., data = train.trs)  
  
Prior probabilities of groups:  
 Adelie Chinstrap Gentoo   
0.4688645 0.1978022 0.3333333   
  
Group means:  
 bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
Adelie -0.8944932 0.5662161 -0.7547519 -0.5838012  
Chinstrap 0.9421106 0.5867969 -0.2805355 -0.5468856  
Gentoo 0.6991336 -1.1446449 1.2281006 1.1456964

1. Find the model accuracy, and comment on the model efficacy.

#prediction  
prediction.qda=predict(qda.pengu, test.trs, prior = qda.pengu$prior)  
#Model Accuracy  
accuracy=mean(prediction.qda$class==test.trs$Species)  
accuracy

[1] 0.9275362

Comment:The model efficacy is found to be about 92.75%, which is reasonably a high percentage.

1. Compute the confusion matrix, and comment on the findings.

# Finding Confusion matrix  
#install.packages("caret")  
library(caret)  
confusion.matrix.qda <- confusionMatrix(prediction.qda$class, test.data$species)  
print(confusion.matrix.qda)

Confusion Matrix and Statistics  
  
 Reference  
Prediction Adelie Chinstrap Gentoo  
 Adelie 23 5 0  
 Chinstrap 0 9 0  
 Gentoo 0 0 32  
  
Overall Statistics  
   
 Accuracy : 0.9275   
 95% CI : (0.8389, 0.9761)  
 No Information Rate : 0.4638   
 P-Value [Acc > NIR] : 2.345e-16   
   
 Kappa : 0.8837   
   
 Mcnemar's Test P-Value : NA   
  
Statistics by Class:  
  
 Class: Adelie Class: Chinstrap Class: Gentoo  
Sensitivity 1.0000 0.6429 1.0000  
Specificity 0.8913 1.0000 1.0000  
Pos Pred Value 0.8214 1.0000 1.0000  
Neg Pred Value 1.0000 0.9167 1.0000  
Prevalence 0.3333 0.2029 0.4638  
Detection Rate 0.3333 0.1304 0.4638  
Detection Prevalence 0.4058 0.1304 0.4638  
Balanced Accuracy 0.9457 0.8214 1.0000

comment:The model has an overall accuracy of 92.75%, indicating good predictive performance.Adelie and Gentoo penguins are well-predicted, with high counts on the diagonal. However, Chinstrap penguins are occasionally confused with Adelie, as indicated by 5 counts in the cell.

1. Use the discriminant function to predict the species of two penguins with the following measurements on the variables bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, and body\_mass\_g, respectively. The variables are measured on the scale as the original <data:Obs>. 1 = (43.6, 13.9, 217, 4900),Obs. 2 = (39.3, 20.6,190, 3650)

# Prediction of new observations  
qda.pengu2=MASS::qda(species~., data=train.data)  
bill\_length\_mm=c(43.6,39.3)   
bill\_depth\_mm=c(13.9,20.6)  
flipper\_length\_mm=c(217,190)  
body\_mass\_g=c(4900,3650)  
new.untr <- data.frame(bill\_length\_mm, bill\_depth\_mm,flipper\_length\_mm, body\_mass\_g)  
predict.new.untr.qda=predict(qda.pengu2, new.untr, prior = qda.pengu2$prior)  
predict.new.untr.qda

$class  
[1] Gentoo Adelie  
Levels: Adelie Chinstrap Gentoo  
  
$posterior  
 Adelie Chinstrap Gentoo  
1 7.865437e-15 5.735527e-18 1.000000e+00  
2 1.000000e+00 4.461547e-11 1.542826e-41

#prediction with scaled new observations  
 meancol=colMeans(train.data[,2:5])  
 sdcol=sapply(train.data[,2:5], sd)  
 new.tr=data.frame(scale(new.untr,meancol,sdcol))  
 qda.pengu=MASS::qda(Species~., data=train.trs)  
 predict(qda.pengu, new.tr, prior = qda.pengu$prior)

$class  
[1] Gentoo Adelie  
Levels: Adelie Chinstrap Gentoo  
  
$posterior  
 Adelie Chinstrap Gentoo  
1 7.865437e-15 5.735527e-18 1.000000e+00  
2 1.000000e+00 4.461547e-11 1.542826e-41

Comment:Obs.1 belongs to Gentoo group and Obs.2 belongs to Adelie group,respectively with posterior probability of 1 or very close 1.