**CS5900/STAT 46700 Topics in Data Science Spring 2025**

**Lab 10  
[Vaishak Balachandra - 037831852]**

1. The penguins dataset included in the palmerpenguins package provides the size measurements for adult foraging penguins near Palmer Station, Antarctica.
2. Access the data and determine its dimension.
3. How many species of penguins are provided in the dataset?
4. Perform the LDA and QDA to classify the species of penguins.

> ########################################################################################### Q1

>

> # a

>

> install.packages("palmerpenguins")

> library(palmerpenguins)

> data(penguins, package = "palmerpenguins")

> head(penguins)

# A tibble: 6 × 8

species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g sex year

*<fct>* *<fct>* *<dbl>* *<dbl>* *<int>* *<int>* *<fct>* *<int>*

1 Adelie Torgersen 39.1 18.7 181 3750 male 2007

2 Adelie Torgersen 39.5 17.4 186 3800 female 2007

3 Adelie Torgersen 40.3 18 195 3250 female 2007

4 Adelie Torgersen NA NA NA NA NA 2007

5 Adelie Torgersen 36.7 19.3 193 3450 female 2007

6 Adelie Torgersen 39.3 20.6 190 3650 male 2007

> names(penguins)

[1] "species" "island" "bill\_length\_mm" "bill\_depth\_mm"

[5] "flipper\_length\_mm" "body\_mass\_g" "sex" "year"

> dim(penguins)

[1] 344 8

> cat("There are 344 rows and 8 columns in the penguins dataset!")

There are 344 rows and 8 columns in the penguins dataset!

###################################################################################################

> # b

>

> # attach(penguins)

> table(penguins$species)

Adelie Chinstrap Gentoo

152 68 124

> cat("There are 3 different species in the penguins dataset namely:\n1. Adelie -- 152\n2. Chinstrap -- 68\n3. Gentoo -- 124")

There are 3 different species in the penguins dataset namely:

1. Adelie -- 152

2. Chinstrap -- 68

3. Gentoo -- 124

###################################################################################################

> # c

>

> # Before, performing the LDA and QDA. Let's remove the NA values in the dataset, if any!!

> sum(is.na(penguins))

[1] 19

> cat("There are NA values in the penguin dataset")

There are NA values in the penguin dataset

> clean\_data = na.omit(penguins)

> head(clean\_data)

# A tibble: 6 × 8

species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g sex year

*<fct>* *<fct>* *<dbl>* *<dbl>* *<int>* *<int>* *<fct>* *<int>*

1 Adelie Torgersen 39.1 18.7 181 3750 male 2007

2 Adelie Torgersen 39.5 17.4 186 3800 female 2007

3 Adelie Torgersen 40.3 18 195 3250 female 2007

4 Adelie Torgersen 36.7 19.3 193 3450 female 2007

5 Adelie Torgersen 39.3 20.6 190 3650 male 2007

6 Adelie Torgersen 38.9 17.8 181 3625 female 2007

> dim(clean\_data)

[1] 333 8

> cat("There are there are 11 such rows in the main dataset 'penguins', which has been removed!!")

There are there are 11 such rows in the main dataset 'penguins', which has been removed!!

>

>

> attach(clean\_data)

> names(clean\_data)

[1] "species" "island" "bill\_length\_mm" "bill\_depth\_mm"

[5] "flipper\_length\_mm" "body\_mass\_g" "sex" "year"

> install.packages("MASS")

> library(MASS)

> ################################################################################

>

> # 1. Performing LDA

>

> lda\_penguins = lda(species~., data = clean\_data)

> lda\_penguins

Call:

lda(species ~ ., data = clean\_data)

Prior probabilities of groups:

Adelie Chinstrap Gentoo

0.4384384 0.2042042 0.3573574

Group means:

islandDream islandTorgersen bill\_length\_mm bill\_depth\_mm flipper\_length\_mm

Adelie 0.3767123 0.3219178 38.82397 18.34726 190.1027

Chinstrap 1.0000000 0.0000000 48.83382 18.42059 195.8235

Gentoo 0.0000000 0.0000000 47.56807 14.99664 217.2353

body\_mass\_g sexmale year

Adelie 3706.164 0.500000 2008.055

Chinstrap 3733.088 0.500000 2007.971

Gentoo 5092.437 0.512605 2008.067

Coefficients of linear discriminants:

LD1 LD2

islandDream -1.20784332 -1.615747171

islandTorgersen -1.08912081 -0.128942981

bill\_length\_mm 0.09371102 -0.396269193

bill\_depth\_mm -0.92018212 -0.071302456

flipper\_length\_mm 0.10433497 -0.001273443

body\_mass\_g 0.00131498 0.001015129

sexmale -0.54812118 0.912511570

year -0.35906682 0.134141729

Proportion of trace:

LD1 LD2

0.8458 0.1542

> cat("Inference: LD1 accounts for 84.58% of the variance between the species groups. i.e., (LD1) is sufficient to capture most of the class separability in the data.")

Inference: LD1 accounts for 84.58% of the variance between the species groups. i.e., (LD1) is sufficient to capture most of the class separability in the data.

> plot(lda\_penguins, col = c("red","green","blue")[as.integer(clean\_data$species)])

A chart with different colored dots

AI-generated content may be incorrect.

> # let's consider only numerical variables in the model

>

> lda\_numerical = lda(species~bill\_length\_mm+bill\_depth\_mm+flipper\_length\_mm+body\_mass\_g, data = clean\_data)

> lda\_numerical

Call:

lda(species ~ bill\_length\_mm + bill\_depth\_mm + flipper\_length\_mm +

body\_mass\_g, data = clean\_data)

Prior probabilities of groups:

Adelie Chinstrap Gentoo

0.4384384 0.2042042 0.3573574

Group means:

bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g

Adelie 38.82397 18.34726 190.1027 3706.164

Chinstrap 48.83382 18.42059 195.8235 3733.088

Gentoo 47.56807 14.99664 217.2353 5092.437

Coefficients of linear discriminants:

LD1 LD2

bill\_length\_mm -0.085926709 -0.41660160

bill\_depth\_mm 1.041646762 -0.01042272

flipper\_length\_mm -0.084552842 0.01424552

body\_mass\_g -0.001347375 0.00168559

Proportion of trace:

LD1 LD2

0.8655 0.1345

> cat("Inference: LD1 accounts for 86.55% of the variance between the species groups. i.e., (LD1) is sufficient to capture most of the class separability in the data.")

Inference: LD1 accounts for 86.55% of the variance between the species groups. i.e., (LD1) is sufficient to capture most of the class separability in the data.

> plot(lda\_numerical, col = c("red","green","blue")[as.integer(clean\_data$species)])

A diagram of different colored dots

AI-generated content may be incorrect.

> # or

>

> # install.packages("ggplot2")

> library(ggplot2)

> lda\_pred <- predict(lda\_penguins)

> lda\_df <- data.frame(species = clean\_data[, "species"], LD1 = lda\_pred$x[, 1], LD2 = lda\_pred$x[, 2])

> ggplot(lda\_df) + geom\_point(aes(x = LD1, y = LD2, color = species), size = 4) + theme\_classic()

>

A group of dots with numbers

AI-generated content may be incorrect.

> ################################################################################

>

> # 2. Performing QDA

>

> library(MASS)

> library(ggplot2)

> set.seed(037831852)

> index <- sample(1:nrow(clean\_data), 0.7 \* nrow(clean\_data))

> train <- clean\_data[index, ]

> test <- clean\_data[-index, ]

>

> qda\_numerical <- qda(species~bill\_length\_mm+bill\_depth\_mm+flipper\_length\_mm+body\_mass\_g, data = train)

> qda\_numerical

Call:

qda(species ~ bill\_length\_mm + bill\_depth\_mm + flipper\_length\_mm +

body\_mass\_g, data = train)

Prior probabilities of groups:

Adelie Chinstrap Gentoo

0.4506438 0.2103004 0.3390558

Group means:

bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g

Adelie 38.71714 18.26190 190.2571 3695.714

Chinstrap 48.89592 18.52245 195.6531 3777.551

Gentoo 47.20380 14.95696 216.7089 5070.253

>

> predicted <- predict(qda\_numerical, test)

>

> head(predicted$class)

[1] Adelie Adelie Adelie Adelie Adelie Adelie

Levels: Adelie Chinstrap Gentoo

>

> head(predicted$posterior)

Adelie Chinstrap Gentoo

1 0.8285099 1.714901e-01 6.453494e-29

2 0.9995185 4.814871e-04 2.792383e-30

3 0.9999980 2.010203e-06 4.009137e-34

4 0.9999929 7.108147e-06 6.439711e-40

5 0.9999980 2.046592e-06 1.779171e-37

6 0.9999999 9.497511e-08 1.197489e-34

>

> mean(predicted$class == test$species)

[1] 0.99

> cat("Accuracy: 99%")

Accuracy: 99%

>

> install.packages("klaR")

> library(klaR)

> partimat(species ~ bill\_length\_mm + bill\_depth\_mm + flipper\_length\_mm + body\_mass\_g,

+ data = train, method = "qda")

A group of graphs showing different colors

AI-generated content may be incorrect.