Lab 2, Group B9

Anton Gefvert, Richard Friberg, Ruben Hillborg

12/10/2018

Contributions

- Assignment 1: Anton Gefvert
- Assignment 2:
- Assignment 3:

Assignment 1

1.1

Looking at Figure 1, we see that drawing a line to discriminate between male and female crabs on these data would be very feasable, thus classifying by linear discriminant analysis would be easy in this case.

1.2

If we compare Figure 2 to Figure 1, we see that they are very similar. There are some differences when carapace length is small, e.g the bottom left point is actually female. but is classified as male with the lda. There are also some differences when you look at the points which are kind of overlapping each other (some males classfied as females and vice versa). We notice though, that the bigger the carapace length is, the more accurate the lda model is!

When using this lda model we get a missclassification rate of 4.5%, this indicates that the model is very well fitted to this problem.

1.3

As seen in Figure 3,when using $p(Male) = 0.9, p(Female) = 0.1$ as a prior, we get (compared to Figure 3), as expected when weighting the male sex higher, a more male dominated graph. This can especially be seen in the lower bounds of carapace length and the values that are pretty close in between the two separation (e.g. the value around $CL = 38$ and $RW = 15$ is female in Figure 2 and male in Figure 3).

If we look at the missclassification rate when using the prior we get a missclassification rate of 8%, this is still very good, but almost twice as much when not using prior (or rather using a prior $p(Male) = p(Female) = 0.5$)

1.4

If we look at Figure 4 we see that it looks very similar to all the other figures. It looks very much like Figure 2, and is more keen to classify uncertain values as female than when using lda with prior.

Missclassification rate for this method is 3.5%, so slightly better than the other methods.

The equation for the decision boundary (green line in Figure 4) is $RW = 0.369CL + 1.08$

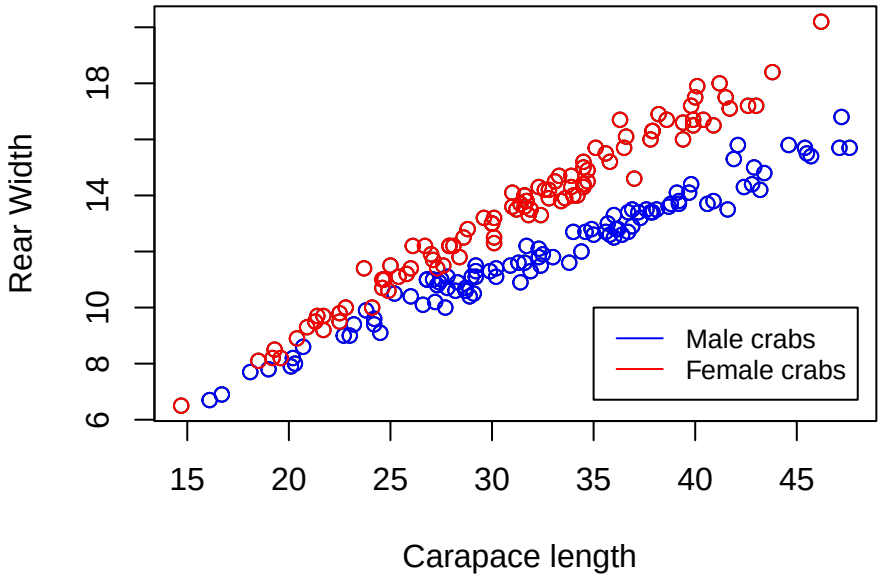

Figure 1: Carapace Length vs Rear Width classified by sex

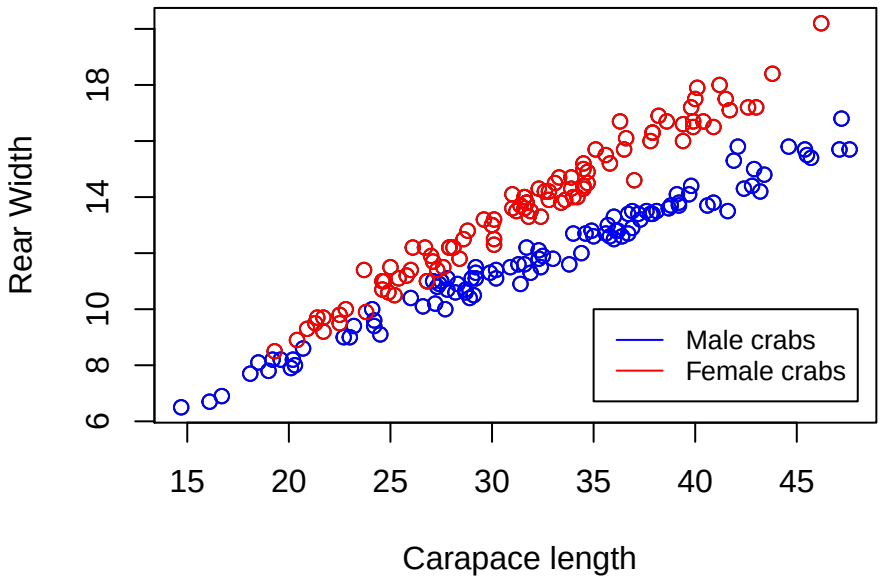

Figure 2: Carapace Length vs Rear Width classified by sex using lda

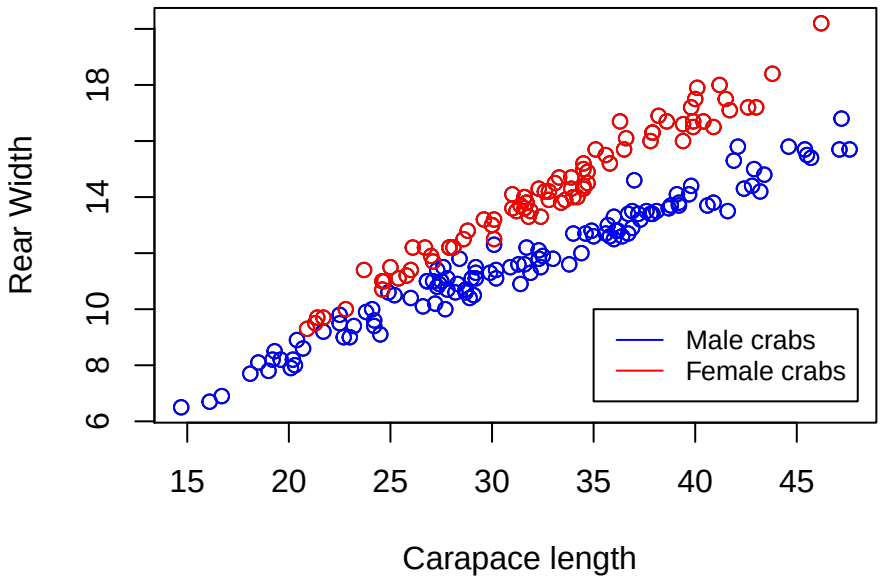

Figure 3: Carapace Length vs Rear Width classified by sex using lda (with prior)

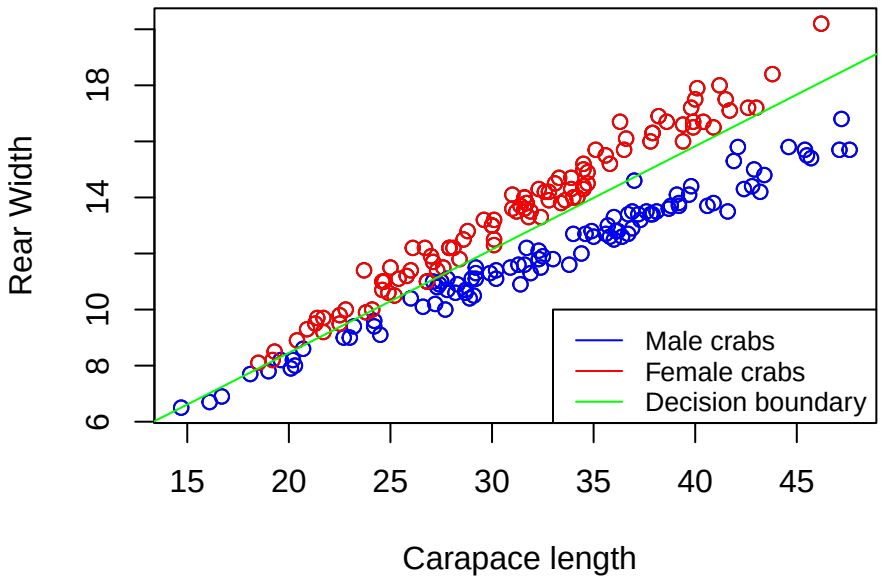

Figure 4: Carapace Length vs Rear Width classified by sex using glm

Assignment 2

Code appendix

```
#setup
knitr::opts_chunk$set(echo = FALSE, warning=F)
library(MASS)
library(readxl)
library(knitr)
library(tree)
library(e1071)
library(ROCR)
library(fastICA)

#1.1
data1 = read.csv("australian-crabs.csv")

male_data = data1[data1$sex == "Male",]
female_data = data1[data1$sex == "Female",]

plot(data1$CL, data1$RW, ylab="Rear Width", xlab=" Carapace length")
points(male_data$CL, male_data$RW, col="blue")
points(female_data$CL, female_data$RW, col="red")

legend(35, 10, legend=c("Male crabs", "Female crabs"),
       col=c("blue", "red"), lty=1, cex=0.8)

#1.2
lda_crabs = lda(sex ~ CL + RW, data=data1, CV=TRUE)
pred_female_data = data1[lda_crabs$posterior[,1] > lda_crabs$posterior[,2],]
pred_male_data = data1[lda_crabs$posterior[,2] > lda_crabs$posterior[,1],]

plot(data1$CL, data1$RW, ylab="Rear Width", xlab=" Carapace length")
points(pred_male_data$CL, pred_male_data$RW, col="blue")
points(pred_female_data$CL, pred_female_data$RW, col="red")

legend(35, 10, legend=c("Male crabs", "Female crabs"),
       col=c("blue", "red"), lty=1, cex=0.8)

pred_sex_list = lda_crabs$posterior[,2] > lda_crabs$posterior[,1] # True is male
sex_list = data1$sex == "Male" # True is male

wrongs = length(data1[sex_list != pred_sex_list, ][,1])

missclassification = wrongs / length(data1[,1])

missclassification

#1.3
lda_crabs = lda(sex ~ CL + RW, data=data1, CV=TRUE, prior=c(0.1, 0.9))
pred_female_data = data1[lda_crabs$posterior[,1] > lda_crabs$posterior[,2],]
pred_male_data = data1[lda_crabs$posterior[,2] > lda_crabs$posterior[,1],]

plot(data1$CL, data1$RW, ylab="Rear Width", xlab=" Carapace length")
points(pred_male_data$CL, pred_male_data$RW, col="blue")
```

```
points(pred_female_data$CL, pred_female_data$RW, col="red")

legend(35, 10, legend=c("Male crabs", "Female crabs"),
       col=c("blue", "red"), lty=1, cex=0.8)

pred_sex_list = lda_crabs$posterior[,2] > lda_crabs$posterior[,1] # True is male
sex_list = data1$sex == "Male" # True is male

wrongs = length(data1[sex_list != pred_sex_list, ][,1])

missclassification = wrongs / length(data1[,1])

missclassification

#1.4
crab_glm = glm(sex ~ CL + RW, data = data1, family=binomial())
glm_pred = predict(crab_glm, type="response")

male_glm = data1[glm_pred >= 0.5,]
female_glm = data1[glm_pred < 0.5,]

plot(data1$CL, data1$RW, ylab="Rear Width", xlab=" Carapace length")
points(male_glm$CL, male_glm$RW, col="blue")
points(female_glm$CL, female_glm$RW, col="red")

slope = coef(crab_glm)[2]/(-coef(crab_glm)[3])
intercept = coef(crab_glm)[1]/(-coef(crab_glm)[3])
abline(a=intercept, b=slope, col="green")

legend(33, 10, legend=c("Male crabs", "Female crabs", "Decision boundary"),
       col=c("blue", "red", "green"), lty=1, cex=0.8)

glm_sex_list = glm_pred > 0.5 # True is male
sex_list = data1$sex == "Male" # True is male

wrongs = length(data1[sex_list != glm_sex_list, ][,1])

missclassification = wrongs / length(data1[,1])

#missclassification
#slope
#intercept
```