lab2

Priya Kurian pullolickal

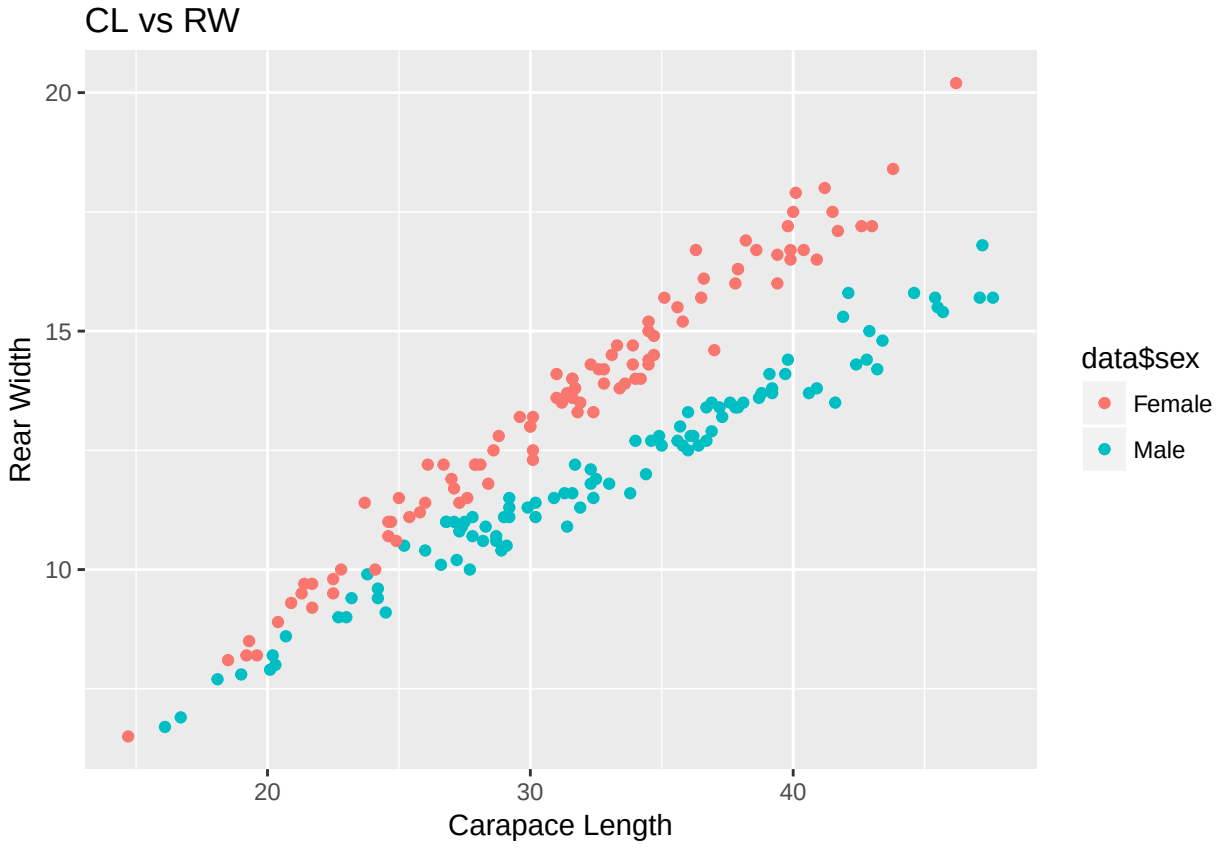
November 30, 2017

Assignment 1. LDA and logistic regression

1.1

linear discriminant analysis estimates the probability of a sample which belongs to a specific class. And Logistic regression is used to describe data and relationship between one dependent binary variable and one or more independent variables.

Given Australian-crabs.csv contains the measurements of different features of the crab.Also it has column for *sex*,where the crabs are divided as male or female. Two indepedent variables carapace length (CL) versus rear width (RW) are considered to classify data.



In the above plot we have shown the relation between CL and RW where the points are coloured on the basis of sex.From the scatterplot we can see two distinct groups of points,Red-Female and Blue-Male.Since we got two different groups of point for the variable sex,we can conclude that the data can be classified using Linear discriminant analysis.

1.2

LDA can be calculated using the discriminant function equation

$$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2} \mu_k^T \sum^{-1} \mu_k + \log \pi_k$$

The proportional priors are

$$\hat{\pi_k} = \frac{N_k}{N}$$

sigma,

$$\widehat{\sum}_c = \frac{1}{N} \sum_{c=1}^{k} N_c \widehat{\sum}_c$$

where the coefficients are,

$$w_{0i} = -\frac{1}{2} \mu_i^T \sum^{-1} \mu_i + \log \pi_i$$

$$w_i = \sum^{-1} \mu_i$$

which are implemented as below:

```
## [1]  5.682847 -1.947504 -9.865352
```

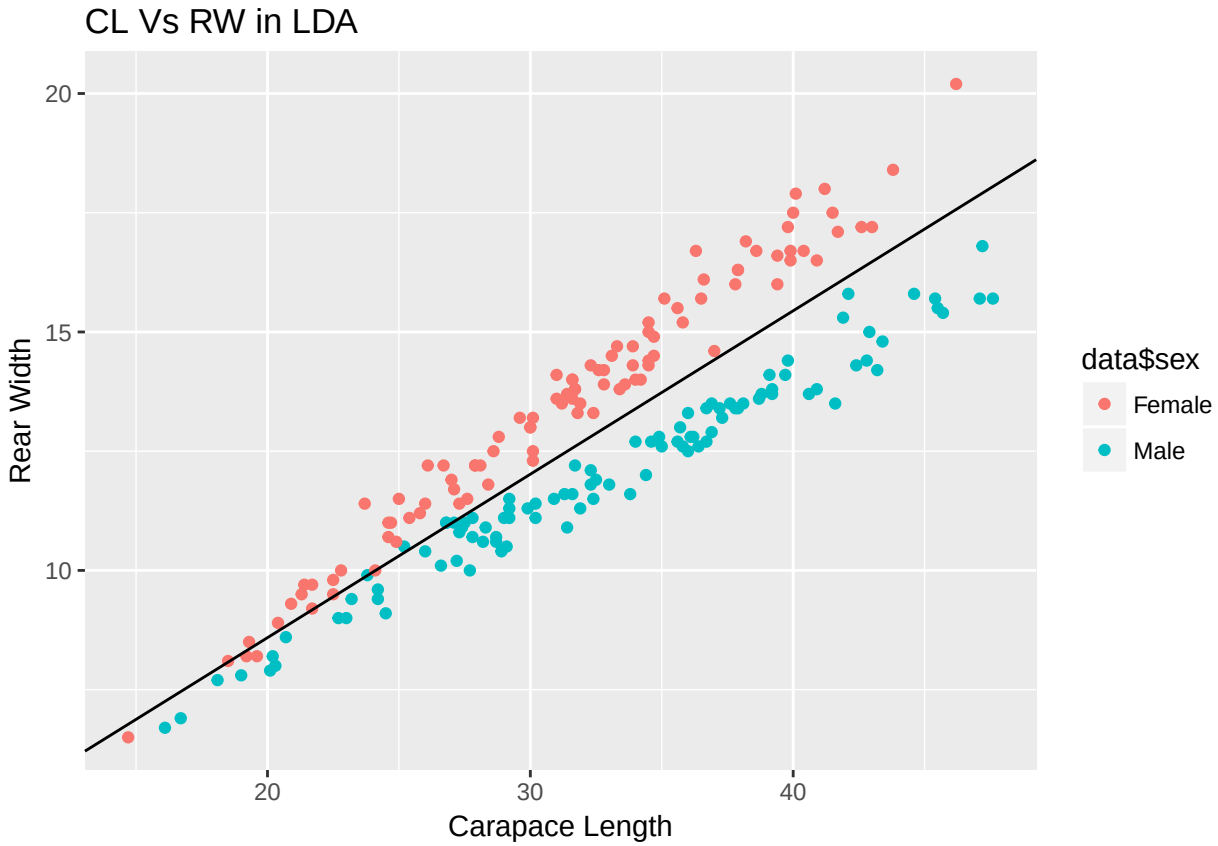
Finding the equation of the decision boundary :

```
## [1] 0.3426987
```

```
## [1] 1.735988
```

```
## Equation of the decision boundary: 1.735988 + 0.3426987 * k
```

1.3



The plot with decision boundary,the points above the black line are Female and the points below the line in are Male.From the plot it can be concluded that the prediction is good as the classified data looks almost similar to the data of first plot which is the original one.

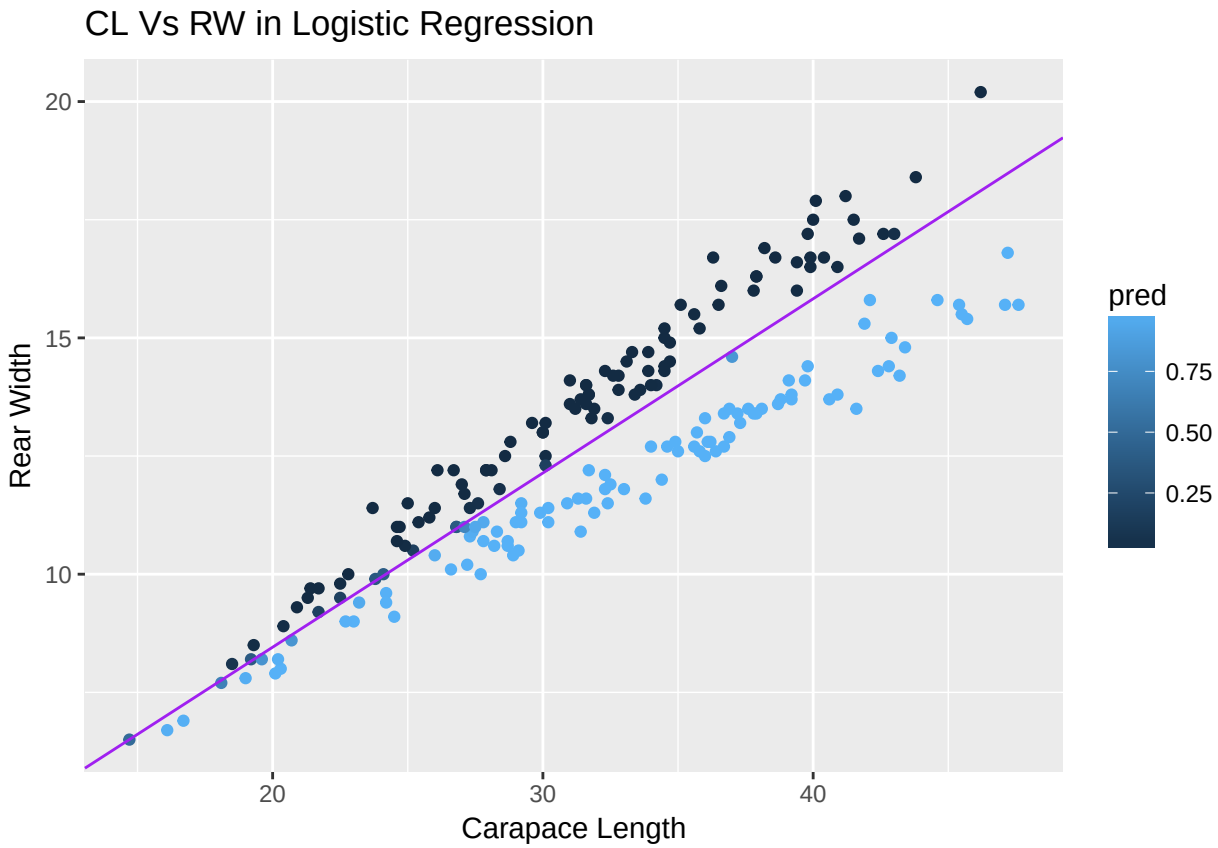
1.4 Classification using logistic regression from glm() function :

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##        CL
## 0.3685758
```

```
## (Intercept)
##     1.08379
```

Classified data plotted as below:



Equation of the decision boundary

```
## Decision boundary with Logistic Regression: 0.3685758 + 1.08379 * k
```

Based on both model , it can be concluded that both have almost same model but they fit differently.

APPENDIX

```
#Assignment 1

#ASSIGNMENT 1.2
setwd("D:/MastersLiu/Statistics & Data Mining/732A95Introduction to Machine Learning/Lab2")
crab_data=read.csv("australian-crabs.csv", sep = ",", dec = ".")

#1
library(ggplot2)

qplot(
  x =crab_data$CL,
  y = crab_data$RW,
  data = crab_data,
  color = crab_data$sex ,
  main="CL vs RW",
  xlab="Carapace Length", ylab = "Rear Width")


#2
Y=crab_data$sex
```

```r
X=data.frame(RW=crab_data$RW, CL=crab_data$CL)

disc_fun=function(label, S){
  A=X[Y==label,]
  A_mean<-sapply(A,mean)
  #MISSING: compute LDA parameters w1 (vector with 2 values) and w0 (denoted here as b1)

  b1<- (-(1/2)%*%t(A_mean)%*%solve(S)%*%A_mean )+ log(pi[1])

  w1<-(solve(S)%*%A_mean)
  return(c(w1[1], w1[2], b1[1,1]))

 }

X1=X[Y=="Male",]
X2=X[Y=="Female",]

#finding the covarainces
S=cov(X1)*dim(X1)[1]+cov(X2)*dim(X2)[1]
S=S/dim(X)[1]
pi<-c(dim(X1)[1]/dim(X)[1], dim(X2)[1]/dim(X)[1])

#discriminant function coefficients
res1=disc_fun("Male",S)
res2=disc_fun("Female",S)

# MISSING: use these to derive  decision boundary coefficients 'res'
res <- res2-res1

# classification

d=res[1]*X[,1]+res[2]*X[,2]+res[3]
Yfit=(d>0)
#plot(X[,1], X[,2], col=Yfit+1, xlab="CL", ylab="RW")

slope <- -(res[2] / res[1] )
intercept <- -(res[3] /res[1] )
cat("Equation of the decision boundary:" , intercept, "+", slope, "* k\n")
#1.3
#MISSING: use 'res' to plot decision boundary.

print(qplot(
  x =crab_data$CL,
  y = crab_data$RW,
  data = crab_data,
  color = Yfit ,
  main="CL vs RW",
  xlab="Carapace Length", ylab = "Rear Width")
+geom_abline(slope = slope, intercept = intercept)+ggtitle("CL Vs RW in LDA"))


#1.4
```

```
glm1 <- glm(sex ~ CL + RW,family=binomial(link="logit"), data=crab_data)
slope1 <- -(glm1$coefficients[2] / glm1$coefficients[3] )
intercept1 <- -(glm1$coefficients[1] /glm1$coefficients[3] )
pred <- predict(glm1, newdata=crab_data, type="response")
#plot(pred)

print(qplot(
  x =crab_data$CL,
  y = crab_data$RW,
  data = crab_data,
  color = pred,
  main="CL vs RW",
  xlab="Carapace Length", ylab = "Rear Width")
  +geom_abline(slope = slope1, intercept = intercept1,colour='purple')+ggtitle("CL Vs RW in Logistic Re

cat("Decision boundary with Logistic Regression:",
    slope1, "+",intercept1, "* k\n")


################################
```
‘‘‘