Computer Lab 3

Introduction to Machine Learning

Emil K Svensson

Sys.Date()

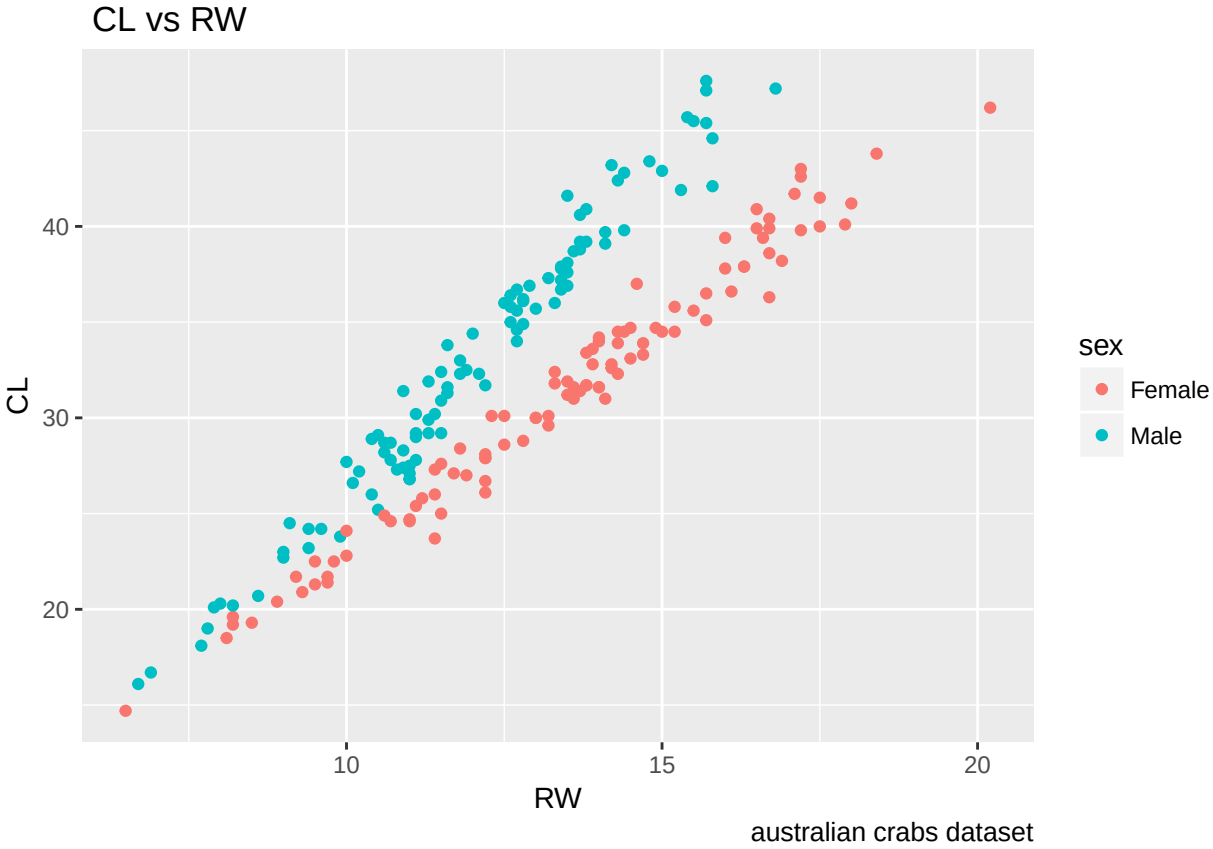
Assignment 1

1.1

```
crabs <- read.csv("australian-crabs.csv")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
p <- ggplot(data = crabs) + geom_point(aes(x = RW, y = CL, col = sex)) + labs(title = " CL vs RW",
    caption = "australian crabs dataset")
plot(p)
```



A line would be able to separate the genders pretty well. One could expect a couple of missclassifications when the variables has low values as the genders seem to have less separation there.

1.2

```
LDA <- function(X) {

    RW <- X[, 1]
    CL <- X[, 2]
    sex <- X[, 3]
    myMu <- aggregate(cbind(RW, CL), by = list(sex), FUN = mean, simplify = TRUE)
    myCov <- by(cbind(RW, CL), list(sex), cov, method = "pearson")
    myPi <- aggregate(cbind(RW, CL), by = list(sex), FUN = function(x) length(x)/nrow(cbind(RW,
        CL)), simplify = TRUE)


    mySig <- ((myCov[[1]] * myPi[2, 2] * length(RW)) + (myCov[[2]] * myPi[2,
        3] * length(RW)))/nrow(X)


    woMale <- -0.5 * as.matrix(myMu[2, 2:3], ncol = 2) %*% solve(mySig) %*%
        t(myMu[2, 2:3]) + log(myPi[2, 3])
    woFem <- -0.5 * (as.matrix(myMu[1, 2:3], ncol = 2)) %*% solve(mySig) %*%
        t(myMu[1, 2:3]) + log(myPi[1, 3])


    wM <- solve(mySig) %*% t(myMu[2, 2:3])
    wF <- solve(mySig) %*% t(myMu[1, 2:3])


    a <- (woMale - woFem)
    b <- wM - wF
    x <- cbind(X[, 1:2])

    # w0s is a w1s is b[1] w2s is b[2]
    myInter <- as.numeric(-a/b[2])
    mySlope <- as.numeric(-b[1]/b[2])

    X$myClass <- t(ifelse((a[1] + t(b) %*% t(x)) > 0, levels(X[, 3])[2], levels(X[,
        3])[1]))
    colnames(X)[4] <- "Predicted"
    retObj <- list(w0 = c(woMale, woFem), w1 = cbind(wM = wM, wF = wF), myClass = X,
        myModel = c(myInter = myInter, mySlope = mySlope))

    return(retObj)
}
```

The return object of the LDA-function returns a list with all answers and the decision boundary.

To get the decision boundary we have the set the two discriminantfunctions equal to each other and solve for one of the parameters depending on X so that we get where they intercect.

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

$w_i = \Sigma^{-1}\mu_i \; w_{oi} = (-1/2)\Sigma^{-1}\mu_k + log(\pi_k)$

$\delta_{male}(x) = \delta_{female}(x)$

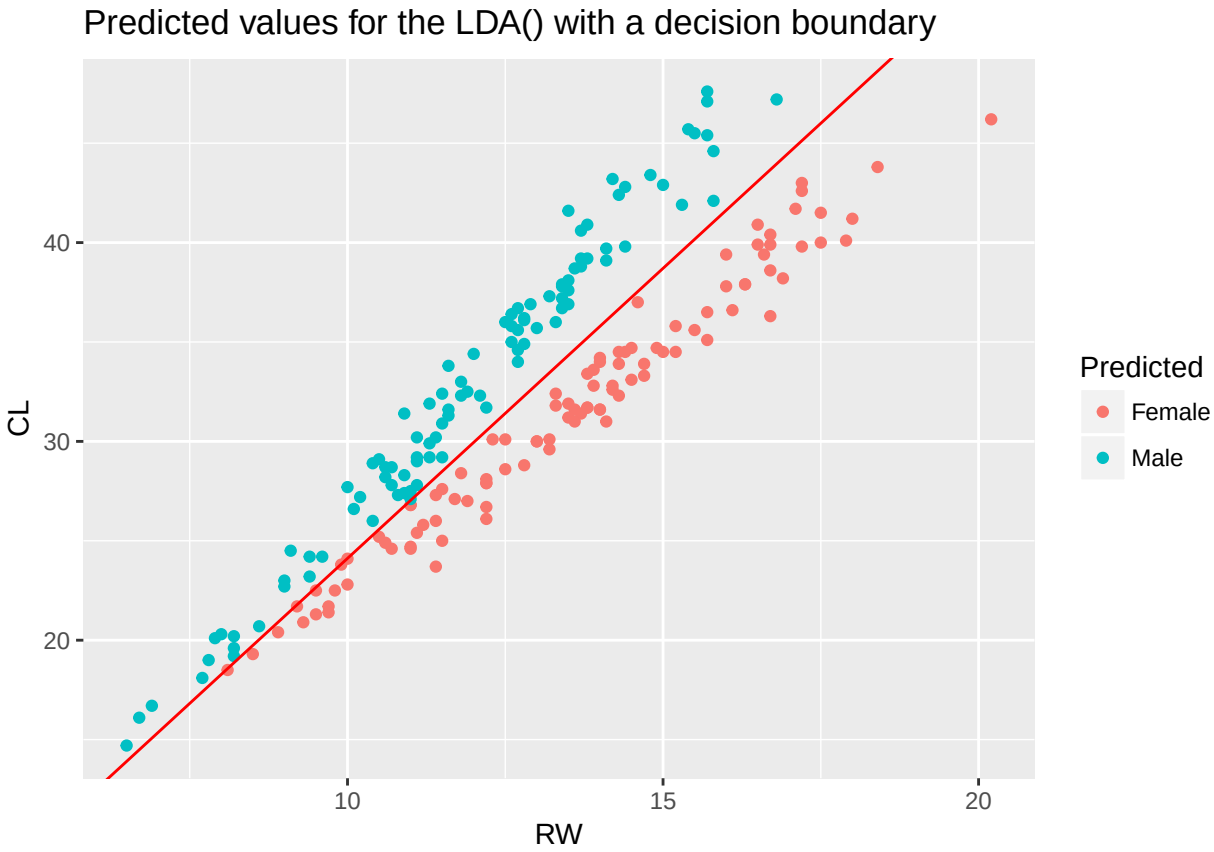
$\delta_{male}(x) - \delta_{female}(x) = 0$

$$x^T(w_{Male} - w_{Female}) + (w_{0Male} - w_{0Female}) = 0$$
$$x_{CL}^T(w_{Male} - w_{Female}) + x_{RW}^T(w_{Male} - w_{Female}) + (w_{0Male} - w_{0Female}) = 0$$
$$x_{RW}^T(w_{Male} - w_{Female}) + (w_{0Male} - w_{0Female}) = -x_{CL}^T(w_{Male} - w_{Female})$$
$$(x_{RW}^T(w_{Male} - w_{Female}) + (w_{0Male} - w_{0Female}))/-(w_{Male} - w_{Female}) = x_{CL}^T$$
$$(x_{RW}^T(w_{Male} - w_{Female}) + (w_{0Male} - w_{0Female}))/-(w_{Male} - w_{Female}) = x_{CL}^T$$

1.3

```
results <- LDA(crabs[, c(5, 6, 2)])
## 2.3 actualdata + desicion boundaries p + geom_abline(intercept =
## results$myModel[1], slope = results$myModel[2], col = 'Red')

# predicted classes + desicion boundaries
ggplot(data = results$myClass) + geom_point(aes(x = RW, y = CL, col = Predicted)) +
    geom_abline(intercept = results$myModel[1], slope = results$myModel[2],
        col = "Red") + labs(title = "Predicted values for the LDA() with a decision boundary")
```



The decision line divides the data nicely but i has some issiues when RW is below 12 where the two groups are closer in distance.

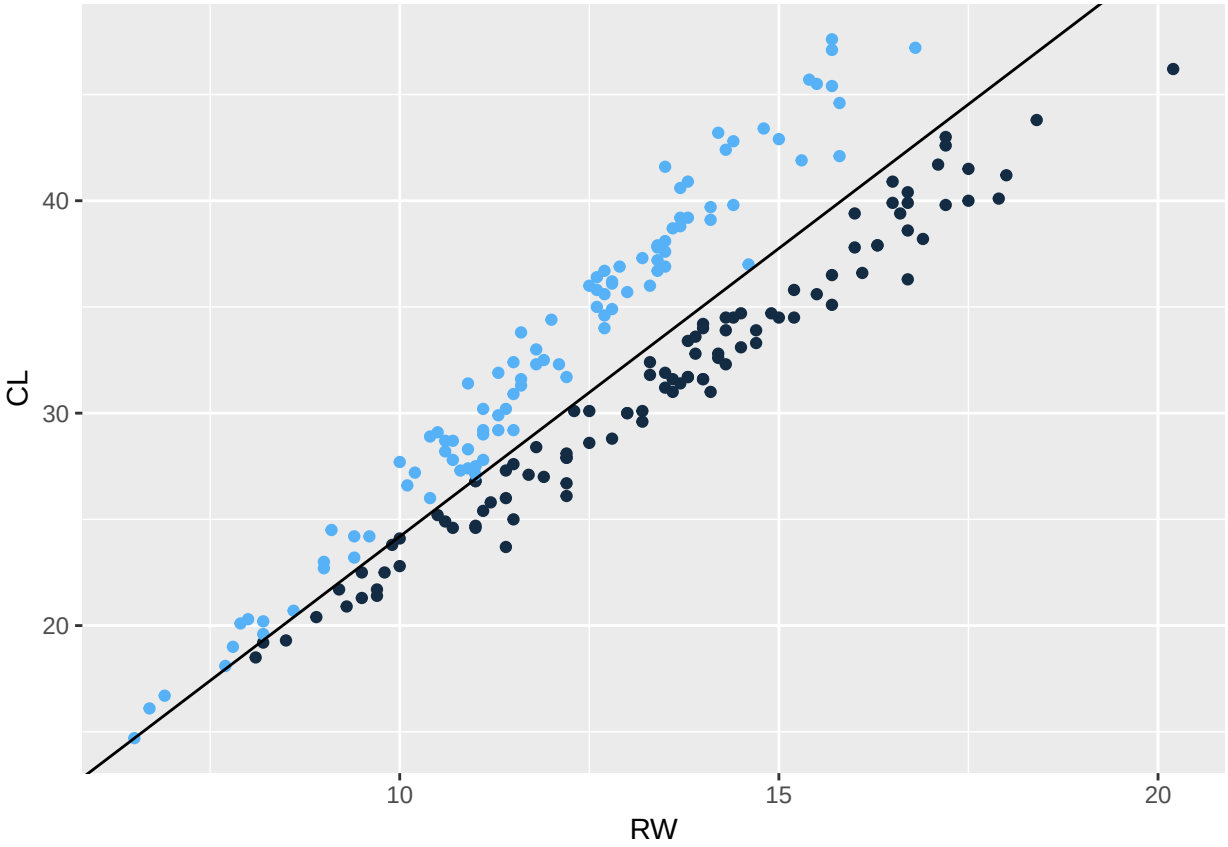
1.4

```
myLogit <- glm(sex ~ RW + CL, family = binomial(link = "logit"), data = crabs)
```

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

```
myDecLog <- coef(myLogit)[1:2]/-coef(myLogit)[3]

ggplot(data = results$myClass) + geom_point(aes(x = RW, y = CL, col = ifelse(myLogit$fitted.values >
    0.5, 1, 0)), show.legend = FALSE) + geom_abline(intercept = myDecLog[1],
    slope = myDecLog[2])
```



One visible difference in the plots are that an observation located close to the line at CL = 37 and RW = 14 is now classified as a male wheras it was classified as a female. Other than this it is hard to disguinsh any visible diffrences.

```
cat("For the Logistic regression \n")
```

```
## For the Logistic regression
```

```
t(table(Predicted = ifelse(myLogit$fitted.values > 0.5, "Male", "Female"), Observed = crabs$sex))
```

```
##          Predicted
## Observed Female Male
##   Female     97    3
##   Male        4   96
```

```
cat("\n")
```

```r
cat("For the LDA:\n")
```

```
## For the LDA:
```

```r
t(table(Predicted = results$myClass[, 4], Observed = crabs$sex))
```

```
##         Predicted
## Observed Female Male
##   Female     97    3
##   Male        4   96
```

Both classifiers has the same missclassificationrate 7/200 and has the same amount (but not necessery the same) of missclassifications in the anti-diagonals for the different categories.

Assignment 2

2.1

```r
CS <- read.csv2("creditscoring.csv")

# Suffle the rows
set.seed(12345)
CS <- CS[sample(nrow(CS)), ]

# Divide them up in different sets
csTrain <- CS[1:(nrow(CS) * 0.5), ]
csValid <- CS[((nrow(CS) * 0.5) + 1):floor(nrow(CS) * 0.75), ]
csTest <- CS[((nrow(CS) * 0.75) + 1):nrow(CS), ]
```

2.2

	Train	Test
Gini	0.212	0.248
Deviance	0.240	0.304

The Gini splitting criterion has a higher training and missclassification rate than the deviance splitting criterion. So for the following step the deviance criterion is used.

2.3