

Lab 3 - Group report

Araya, Caroline, Gustav Stenelöv, Thomas

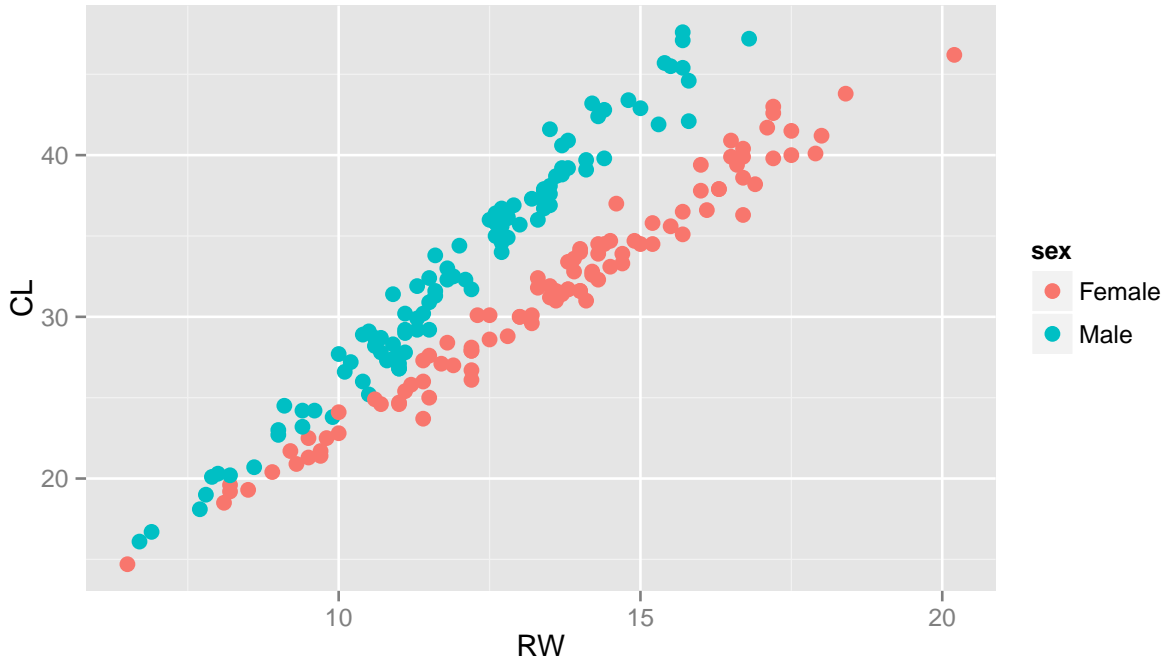
Monday, November 09, 2015

Assignment 1

In the first assignment a data material named australian crabs is analysed. It contains information about measurements of the frontal lobe, rear width etc. for 200 crabs.

1.1

A scatterplot visualizing the carapace length versus rear width where observations are colored by sex.



We can clearly distinguish two separate linear trends in the scatterplot. One trend for the male crabs and one trend for the female crabs. I think it will be easy to classify the sex for the larger crabs, since at larger values of CL and RW they are well separated from each other. At smaller values for the two variables it will be harder to classify with linear discriminant analysis, since the two trends are not well separated there.

1.2

The discriminant functions can be written as:

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

where

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i: y_i = c} x_i$$

$$\hat{\Sigma}_c = \frac{1}{N_c} \Sigma (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T$$

$$\hat{\Sigma} = \frac{1}{N} \Sigma_{c=1}^k N_c \hat{\Sigma}_c$$

$$\hat{\pi}_c = \frac{N_c}{N}$$

After fitting the model, the discriminant function for the male crabs can then be written:

$$\delta_{male}(x) = x^T 0.029410201 \begin{pmatrix} 11.990 \\ 32.851 \end{pmatrix} - \frac{1}{2} (11.990 \quad 32.851) 0.029410201 \begin{pmatrix} 11.990 \\ 32.851 \end{pmatrix} + \log(0.50)$$

and the female function:

$$\delta_{female}(x) = x^T 0.029410201 \begin{pmatrix} 13.487 \\ 31.360 \end{pmatrix} - \frac{1}{2} (13.487 \quad 31.360) 0.029410201 \begin{pmatrix} 13.487 \\ 31.360 \end{pmatrix} + \log(0.50)$$

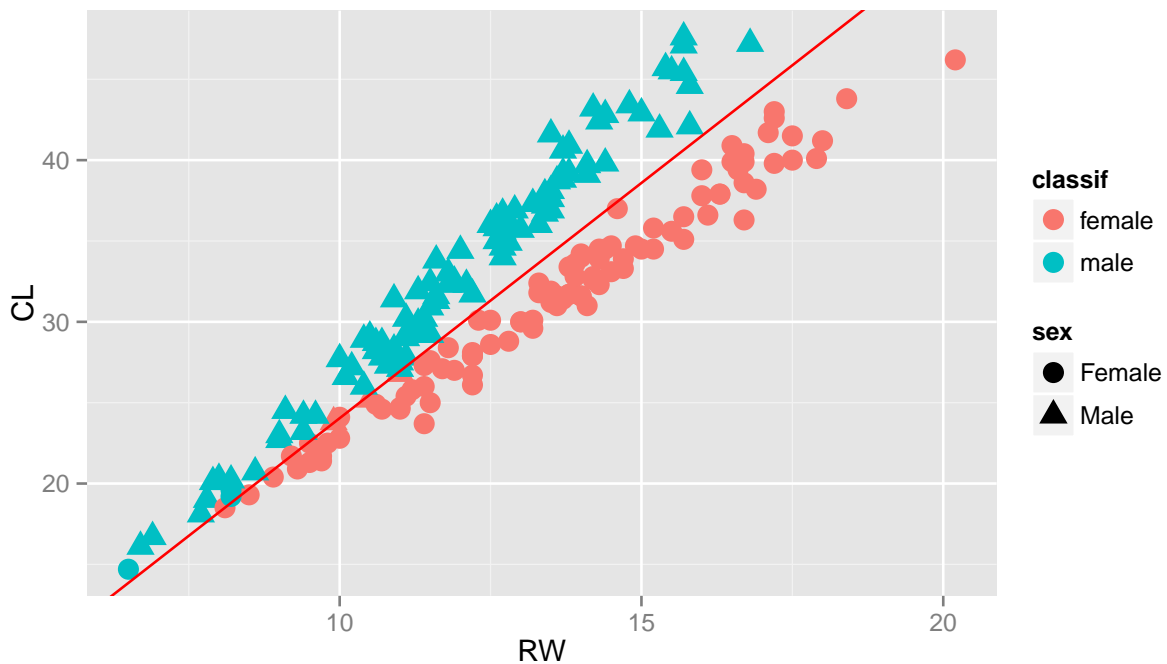
An observation is then classified to which ever discriminant function is greater.
Furthermore, the decision boundary is located where

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

And the estimated decision boundary is : CL = 2.91RW - 5.06

1.3

The original data of australian-crabs.csv is plotted and coloured by the classification labels which are obtained from the function.

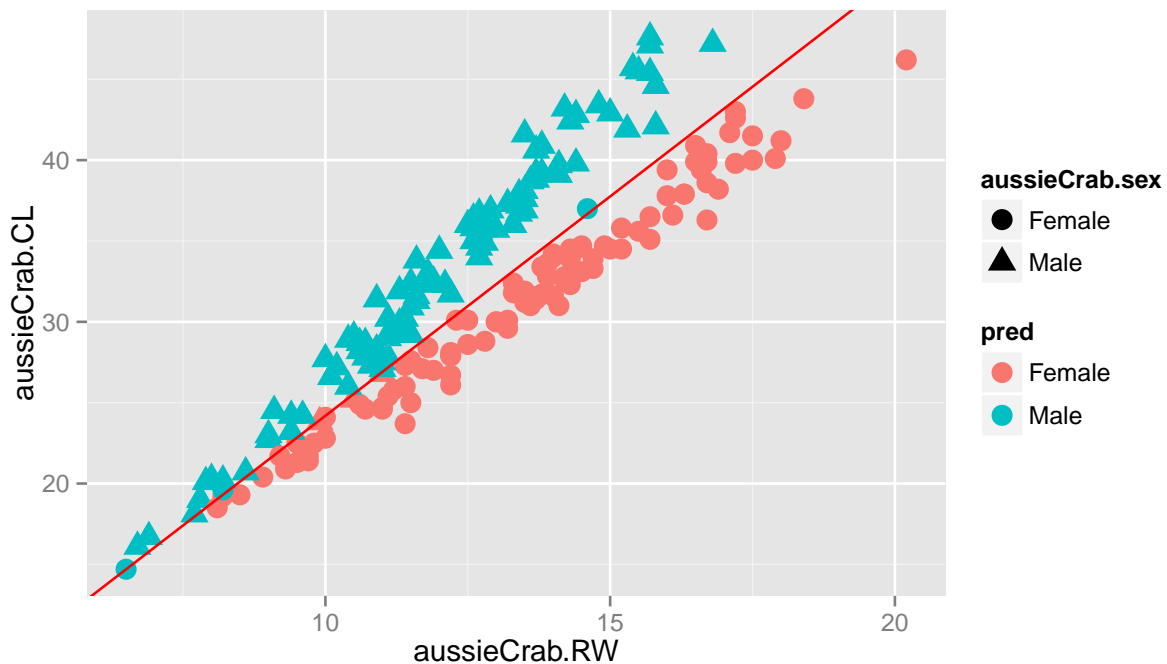


```
##      predict
## actual  female male
##   Female    97    3
##    Male     4   96
```

The red line indicates the decision boundary between female and male group. From looking at this plot, there are some misclassified data happened. The blue triangle indicates group as male while the orange circle represents female. Other symbols than these two (the blue triangle or the orange circle) suggest that it is misclassified. For example, the first point located at the left of the plot is in orange circle. The original is in the female group but when using LDA it is classified into male group. The quality of fit for using LDA is quite good. According to the confusion matrix which is given above, there are 7 observations which go into the wrong class.

1.4

Another method to classify the observations are by a logistic regression model. In the graph below the classifications are plotted in the same way as in 1.3.



The decision boundary for the logistic regression model: $CL = 2.713RW - 2.94$

The obtained result with logistic regression is very similar to the result given with the *lda* model. We see that the slope is a little bit less steep in the logistic regression decision boundary as compared to the LDA decision boundary.

Assignment 2

2.1 - 2.2

We use two measures of node impurity. For *fittree1* we use `split = "deviance"` which essentially minimizes tree information entropy to pick the tree splits, and for *fittree2* we use `split = "gini"`, which minimizes tree Gini impurity. We find misclassification rates for the test data for the two trees.

Deviance: Misclassification rate for the training and test data, respectively.

```
## [1] 0.212
```

```
## [1] 0.268
```

Gini index: Misclassification rate for the training and test data, respectively.

```
## [1] 0.236
```

```
## [1] 0.364
```

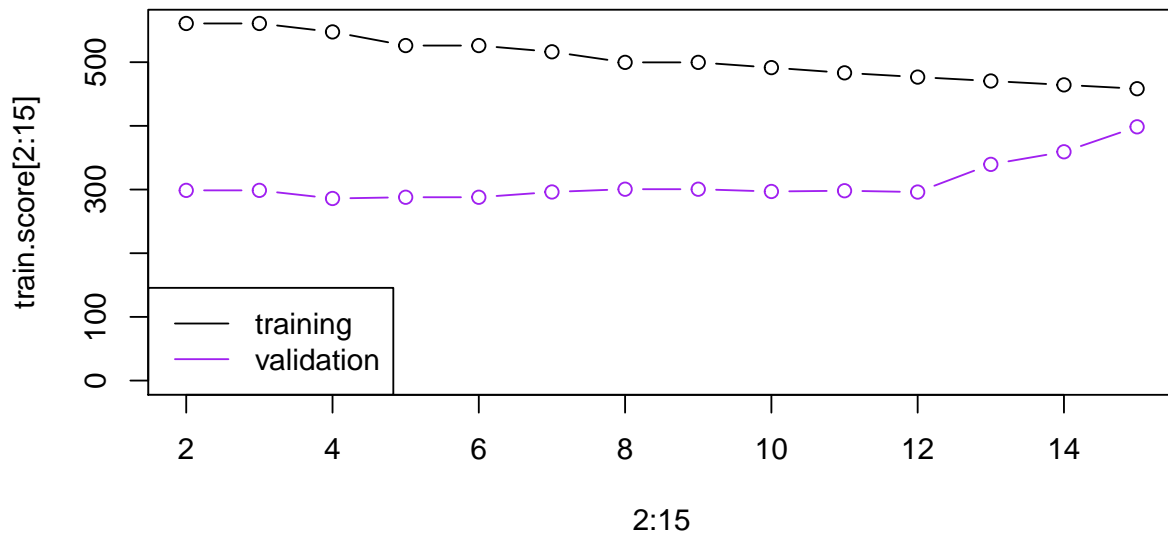
We see that the impurity measure of information entropy is better, so we will continue to evaluate the decision tree *fittree1*.

2.3

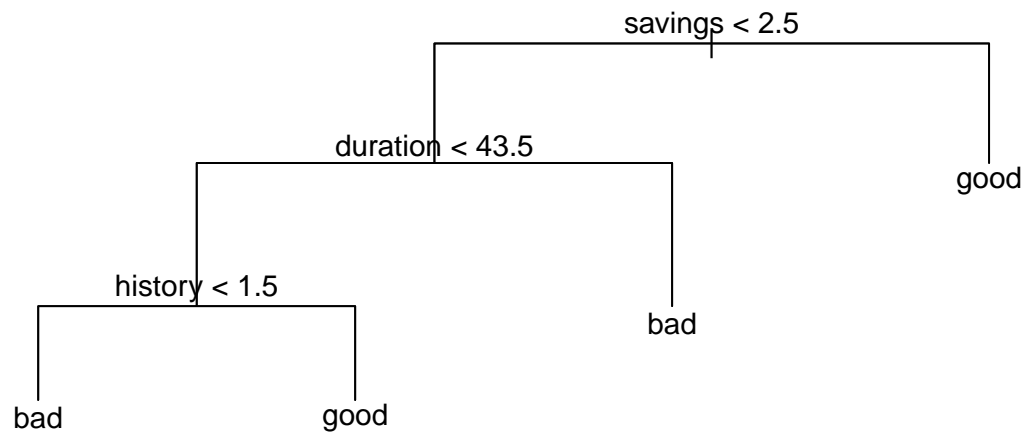
We are going to prune the tree `fittree1` by minimizing

$$\text{Deviance of tree} + \lambda|T|$$

where $|T|$ is the number of terminal nodes and λ is an algorithmically determined cost parameter. Below you will find the plot of deviance of the tree for training data and for validation data against number of terminal nodes in the greedily and recursively pruned subtree of `fittree1`.



The purple line for validation data have the lower score than the training data. When the number of leaves is equal to four the lowest score is obtained. Therefore this plot suggests that the optimal tree for this dataset should have four leaves.



The depth is 3. Uses the variables savings, duration and history. Much savings is good. Long duration is bad. Good history is good.

The misclassification for test data is ...

2.4

In 2.4 the classification of the customers is performed with the Naive Bayes classifier. The first of the tables below gives the confusion matrix for training data and the second for test data. For training data the misclassification rate is 0.4285714 and for test data it is 0.4619883.

2.5