Using R in Data mining for the masses, Chapter 6

Ulrik Hørlyk Hjort

April 6, 2014

1 Modeling

In the following text “the book” will refer to the the book: “Data Mining for the masses”

1.1 k-Means Clustering

Fiest we import the data set for chapter 4:

```
data = read.csv(‘‘Chapter06DataSet.csv’’, sep=’’,’’,header = TRUE)
```

In R we create a four group k-means cluster, like the one described in chapter 6 in the book, in the following way:

```
km<-kmeans(data, 4, iter.max = 10)
```

Printing the cluster shows the distribution of the observations across the four clusters:

```
> print(km)
K-means clustering with 4 clusters of sizes 140, 135, 154, 118

Cluster means:
    Weight Cholesterol    Gender
1 106.8500    119.5357 0.5428571
2 127.7259    154.3852 0.4592593
3 184.3182    218.9156 0.5909091
4 152.0932    185.9068 0.4406780

<<< Some of the output is left out in the example >>>
```

The numbering and order of the clusters generated by R differ a little compared to clusters shown in Figure 6-5 in the book. The relationship is (R -> "The book"): 1 -> 3, 2 -> 2, 3 -> 0 and 4 -> 1. So in this case cluster 3 has the highest average weight and cholesterol.
Filtering out the results of the data set contained in cluster 3 is done in the following way:
First we add an extra row to the data set which append the actual cluster realted to the observation

```
aggregate(data,by=list(km$cluster),FUN=mean)
clusters <- data.frame(data, km$cluster)
cluster3 = subset(clusters,clusters$km.cluster==3)
```

Inspecting the first rows of the data frame of cluster3 gives:

```
> head(cluster3)
   Weight Cholesterol Gender km.cluster
6     198         227      1          3
9     191         223      0          3
10    186         221      1          3
12    188         222      1          3
16    178         213      0          3
18    168         204      1          3
>
```

Filtered results for cluster 3:

```
> summary(cluster3)
     Weight       Cholesterol        Gender         km.cluster
 Min.   :167.0   Min.   :204.0   Min.   :0.0000   Min.   :3  
 1st Qu.:176.2   1st Qu.:212.2   1st Qu.:0.0000   1st Qu.:3  
 Median :183.5   Median :220.0   Median :1.0000   Median :3  
 Mean   :184.3   Mean   :218.9   Mean   :0.5909   Mean   :3  
 3rd Qu.:191.0   3rd Qu.:225.0   3rd Qu.:1.0000   3rd Qu.:3  
 Max.   :203.0   Max.   :235.0   Max.   :1.0000   Max.   :3  
```