Final Project

Han Ambrose

May 8, 2021

Question 1

1a

Let 1 = Alaskan, 0 = Canadian.

All of the variables provided were used including gender, fresh, marine, weight and height

The dark shaded area corresponds to Result =1 (Alaskan) the light shaded area corresponds to Result =0 (Canadian). The probability that Result =1 (Alaskan) when diameter of rings for first year freshwater fish =140 is approximately 5%. The rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon. Similarly, The rings associated with marine growth are larger for the Alaskan-born than for the Canadian-born salmon.

The probability that Result =1 (Alaskan) when weight is more than 11.5 is approximately 10%. Alaskan salmon weighed less than Canadian salmon. Height is a little hard to tell. Variables that have significant power is fresh, marine and weight.

The algorithm did not seem to converge as we are dealing with perfect separation in the data set.

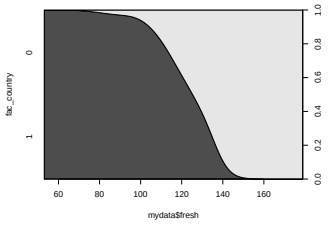


Figure 1: conditional density plot - fresh

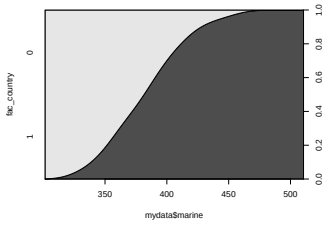


Figure 2: conditional density plot - marine

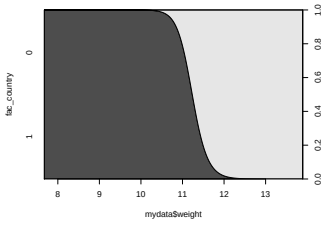


Figure 3: conditional density plot - weight

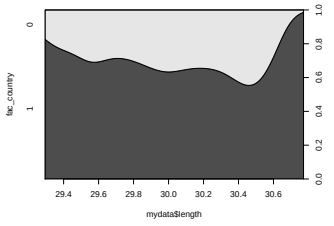


Figure 4: conditional density plot - length

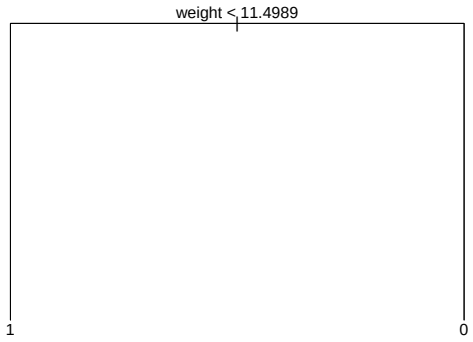
```
Call:
glm(formula = country ~ ., family = "binomial", data = mydata)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.446e-05  -2.100e-08   2.100e-08   2.100e-08   4.071e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.869e+00  2.619e+06   0.000    1.000
gender2      1.602e+01  7.402e+04   0.000    1.000
gender3     -2.200e+01  2.739e+05   0.000    1.000
fresh       -2.719e-01  8.797e+02   0.000    1.000
marine       7.866e-02  1.121e+03   0.000    1.000
weight      -1.991e+01  2.466e+04  -0.001    0.999
length       7.346e+00  8.658e+04   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.9095e+02  on 149  degrees of freedom
Residual deviance: 3.2113e-09  on 143  degrees of freedom
AIC: 14

Number of Fisher Scoring iterations: 25
```

1b

Notice that weight is the only variable used in tree construction. This means that by just weight alone we are able to separate the data into two. Salmons that weight less than 11.4989 is identified as Alaskan and greater than 11.4989 is identified as Canadian (1 = Alaskan, 0 = Canadian). As shown in the figure of importance variable, we could confirm that weight is the most importance variable.

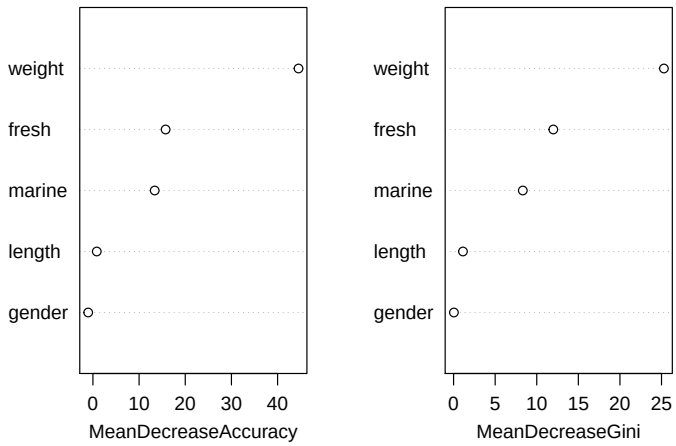
Training error rate for both tree and random forest is 0% and testing accuracy is 100%. As we have a perfect separation, this is somewhat expected.

```
Classification tree:
tree(formula = country ~ ., data = train.set)
Variables actually used in tree construction:
[1] "weight"
Number of terminal nodes:  2
Residual mean deviance:  0 = 0 / 103
Misclassification error rate: 0 = 0 / 105
```




```
Call:
 randomForest(formula = country ~ ., data = train.set, importance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0%
Confusion matrix:
   0  1 class.error
0 36  0           0
1  0 69           0
```

train_rf

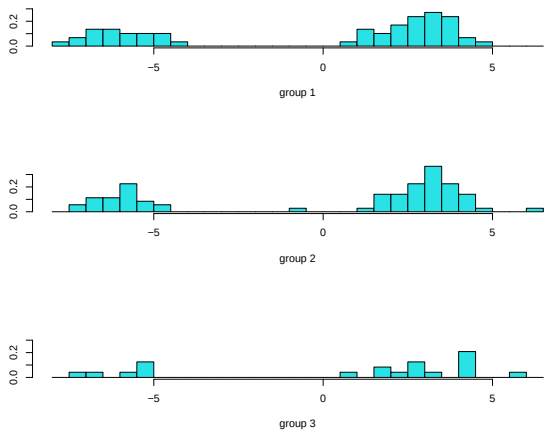


Weight is the most important, then fresh, marine. Length and gender does not seem to be important.

1c

As we can see from the summary of linear discriminant analysis below, weight again has the highest coefficient and the most important variable. The histogram below shows that the data were separated into 2 groups. 100% of data points are predicted correctly

```
Call:
lda(country ~ ., data = mydata)

Prior probabilities of groups:
        0         1
0.3333333 0.6666667

Group means:
  gender2 gender3  fresh marine   weight   length
0    0.48    0.00 137.46 366.62 13.10629 30.04107
1    0.46    0.02  98.38 429.66  9.01787 30.01658

Coefficients of linear discriminants:
                LD1
gender2 -0.064465127
gender3  0.001953653
fresh   -0.019761361
marine   0.006534095
weight  -1.887330692
length  -0.193223257
```



1d

It seems like tree/random forest is the best. Linear Discriminant is also a good choice as we can see data set were separated perfectly into two and testing accuracy is high.

Since we are dealing with well separate dataset, logistics regression did not converge. We could try to take out the weight variables as this variable is causing the perfect separation, then the algorithm will converge.

Question 2

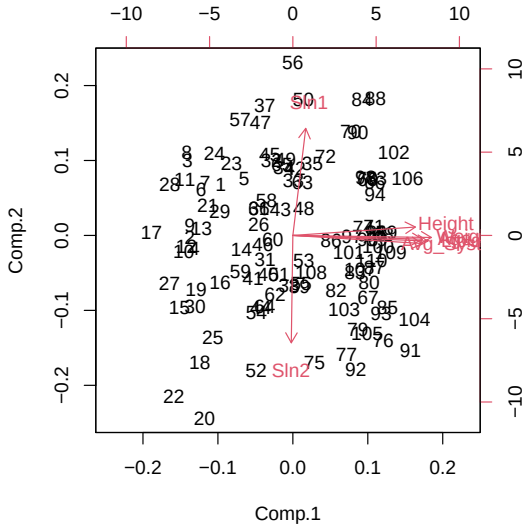
2a

Yes we can reduce dimension by using PCA and SPCA.

First ananlyzing with PCA

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5    Comp.6      Comp.7
Standard deviation     2.0989227 1.0891508 0.9039882 0.59145292 0.36762631 0.3112409 0.096139713
Proportion of Variance 0.6293538 0.1694642 0.1167421 0.04997379 0.01930701 0.0138387 0.001320406
Cumulative Proportion  0.6293538 0.7988180 0.9155601 0.96553388 0.98484089 0.9986796 1.000000000

Loadings:
            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Height       0.418                0.785         0.321  0.317
Weight       0.472                0.153        -0.137 -0.854
Age          0.442               -0.435 -0.654  0.402  0.162
Avg_Systole  0.440               -0.396  0.748  0.277
Lipid        0.460                              -0.800  0.375
Sln1                 0.703  0.701
Sln2                -0.705  0.708
```

Most of the variation is explained by the first 2 or three PCs. The first 2 PCs explain about 80% of variation.



We can also see from the biplot that the results of the height, weight, age, avg_systole, lipid are highly correlated. These are variables in the first component.

On the other hand, the second and third component indicating sln1 and sln2 uncorrelated with those from the first component as they are far apart from each other from the biplot.

These variables might explain 3 factors here. First is anatomic characteristics (height, weight, age, avg_systole, lipid) and second (sln1) or third factor (sln2) is genetic characteristics

In the case of SPCA, it does not seem to be more beneficial than just PCA as we already got 0 coefficients in some of the variables from the PCA loadings summary above.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Height	-0.2757806	0.0000000	0.0000000	1	0	0	0
Weight	-0.6185428	0.0000000	0.0000000	0	0	0	0
Age	-0.4273067	0.0000000	0.0000000	0	0	0	0
Avg_Systole	-0.4414750	0.0000000	0.0000000	0	0	0	0
Lipid	-0.4047946	0.0000000	0.0000000	0	0	0	0
Sln1	0.0000000	0.7124864	-0.7036787	0	0	0	0
Sln2	0.0000000	-0.7016859	-0.7105183	0	0	0	0

Figure 5: SPCA with penalty - Loadings

Rotation seems to help as we can see loading values moves to a after being rotated, especially for factor 2.

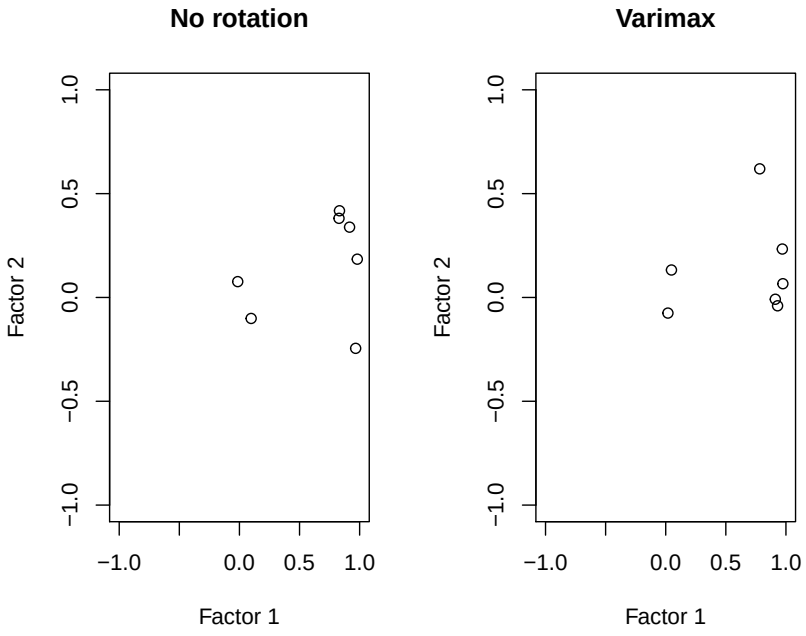


Figure 6: with and without rotation

2b

Below is the performance of hierarchical clustering (using euclidean distance), k-means and model based clustering.

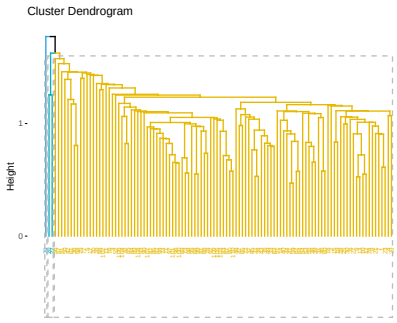


Figure 7: hierarchical clustering - single Link

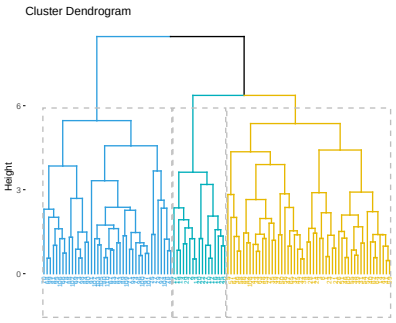


Figure 8: hierarchical clustering - complete Link

Single Linkage is space contracting that is why we see a big yellow cluster. Complete linkage is space dilating as clusters are more spread out.

Figure 9: k means

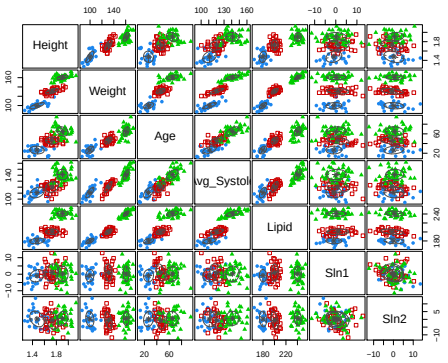


Figure 10: model based with 3 clusters

K-means and model-based seem to do better than hierarchical clustering as they are more evenly spread out or space conserving. In addition, k-means is good to pre-specify the number of clusters beforehand. In this case, we already know that we want 3 clusters therefore k-means seem to be an appropriate clustering method to use.