Math 6020- Final Project

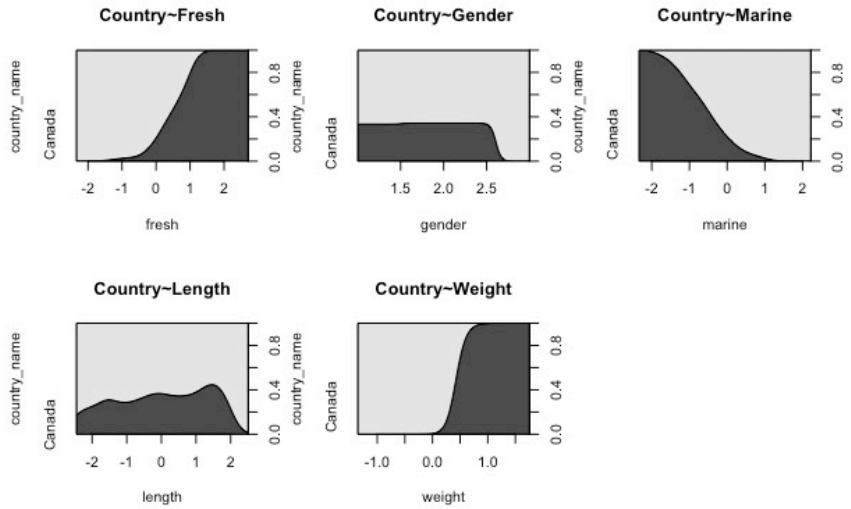
Hilary Hickingbotham

Introduction

For the final project of Math 6020, we are asked to look at two datasets, one about 150 Canadian and Alaskan salmon, and one with phenotype information for 110 patients. The methods applied to these datasets are: logistic regression, deviance analysis through a partial likelihood, classification trees, random forests, Fisher's Linear Discriminant Analysis, PCA and SPCA and clustering.

Salmon Dataset

The first question asked to analyze the data using logistic regression, deviance analysis through partial likelihood and to identify variables that have significant explanatory power. Creating density plots for all the variables:

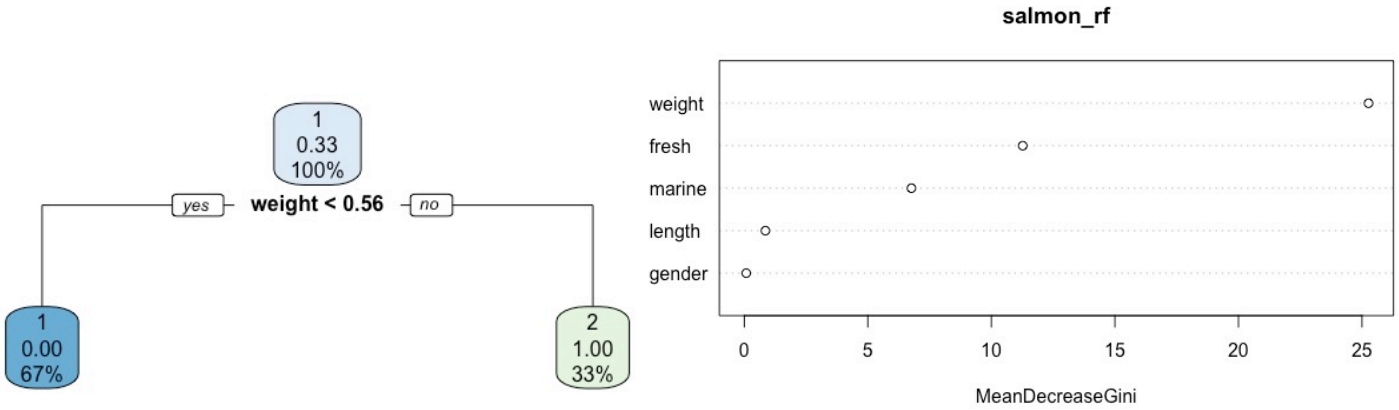


From the density plot, it looks as if, fresh, marine and weight may be significant, whereas gender and length may not be. Completing the logistic analysis, using the glm function, with the different combinations of variables, along with the deviance, and chi-squared tests (partial likelihood), it can be determined that the conditional graph was correct in identifying that length and gender are not significant, and the variables that are significant are fresh, marine, and weight. Below is the deviance likelihood, and p-values of the different parameter models.

	dev	llkhd	pval
6 parameter Model:	"3.211260589954e-09"	"-1.60562940720865e-09"	""
5 parameter Model:	NA	"-27.6445789615931"	NA
4 parameter Model:	"55.3360835406292"	"-27.6680417703146"	NA
3 parameter Model:	"55.4635505787448"	"-27.7317752893724"	"0.721072893602857"
2 parameter Model:	NA	"-39.4279760988912"	NA
1 parameter Model:	"190.954250488444"	"-95.4771252442219"	NA

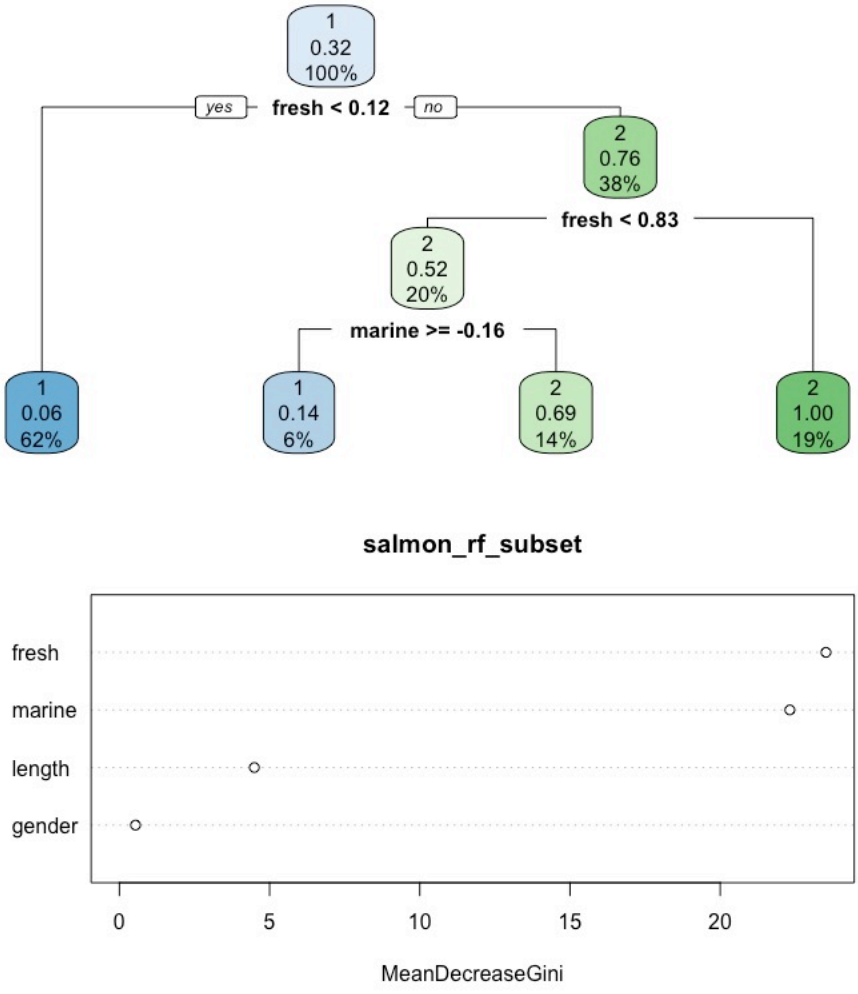
The second question involved splitting the data into a test and training set (with 70% in the training set, and 30% in the test set) and performing a classification tree, random forest and linear discriminant analysis. When all variables are included, weight was the overwhelming predictor of whether a salmon was

from Canada or Alaska. This can be seen by the fact that my classification tree has only one split, and weight has the highest GINI index.



When using the random forest created with all variables, predicting country against my test dataset was 100% accurate. This was also true for the linear discriminant analysis prediction as well.

So, I removed the weight variable. At first, I removed length and gender, and was left with marine, fresh and weight, but again, weight was the only variable that mattered. My classification tree still only had one split, weight had the highest GINI index, and my accuracy was 100%. If weight remained in the data set, it was going to overpower any other analysis, which is why it was removed. With weight removed, my tree became more interesting.

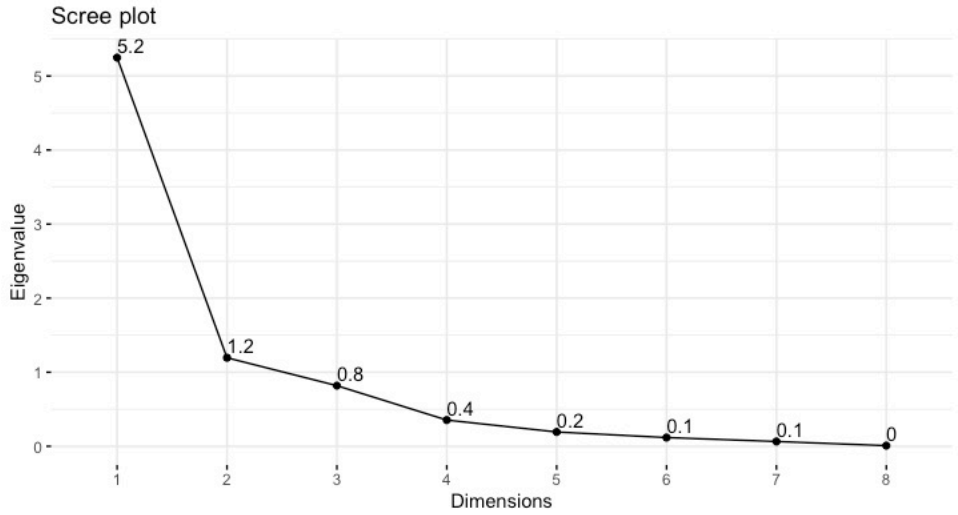


Completing the random forest prediction, the accuracy of the predictions was 87.9%. Completing a linear discriminant analysis without the weight variable, made my predictions 92.7% accurate.

In conclusion, weight was the most important variable in determining the origin of the salmon, followed by fresh and marine. Length and gender had less of an influence in determining country of origin. For this dataset, I think the classification tree is the most preferable, because it gives a lot of information at each layer of the variables.

Phenotype Dataset

The second question gave us the phenotype data for 110 patients. The first part of the question looked at PCA and SPCA analysis, and if rotation would help reduce the dimension of the data. Upon the PCA analysis, it can be determined that 80% of the variation was explained by the first two PCs. This is demonstrated in the scree plot:

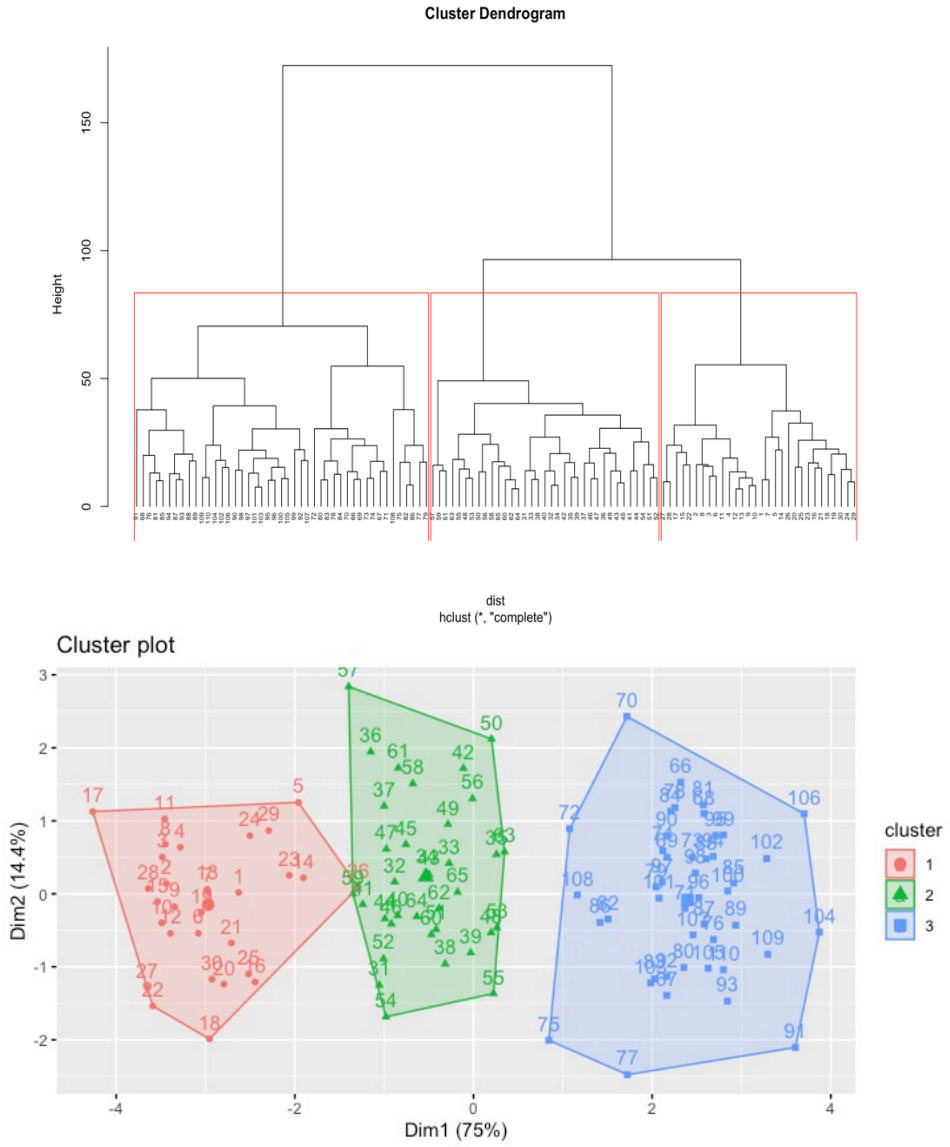


Varimax rotation of the two PCs indicated that Average Systolic Blood Pressure and one of the lifestyle risk indices had strong relationships with the loadings. This is evident by the PC loading summary:

	PC1	PC2
X	0	0
Height	0	0
Weight	0	0
Age	0	0
Avg_Systole	1	0
Lipid	0	0
Sln1	0	0
Sln2	0	1

Performing a SPCA on the two loadings accounted for 70% of the variation in the data, a loss of 10% variation. This indicates that PC analysis is preferable in this case.

The second part of the question detailed the clustering of the data. Given there were three clusters, agglomerative hierarchical clustering and K-means clustering was used with the following results:



From the dendrogram and the cluster graph, the data is split pretty evenly between the three clusters. The cluster analysis gives a better visual that there is not much overlap within the clusters. As far as what each cluster represents, it's very hard to distinguish without being very familiar with the data.

Conclusions

In conclusion, to determine the origin of a salmon, whether Canadian or Alaskan, look at the weight of the fish. If you want a meaty dinner, choose a Canadian salmon. Regarding the phenotype data, use the first two PCs if dimension reduction is the goal, and three clusters gives a good starting point for cluster analysis.

One last note about this class, I really enjoyed it. I learned a lot this semester, and I thought the material was really interesting, with very practical applications. Thank you Jyothsna for being a great professor!