MATH 6020 – Final Project

Due back on Sunday, 9th May [1].

-1- (50 points) Refer to the data file `salmon_full.csv` Recall, from HW2 that salmon fishery is a valuable resource for both the US and Canada. Each country however, should ideally catch salmon that originated from its own waters. To help regulate catches, samples of fish are taken during harvest time and identified as originating from Alaskan or Canadian waters. The fish carry some information about their birthplace in the growth rings on their scales, weights and lengths.

- Column 1: (1 = Alaskan, 2 = Canadian).
- Column 2: Gender
- Column 3: Diameter of rings for first year freshwater fish (in hundredths of an inch).
- Column 4: Diameter of rings for first year marine fish (in hundredths of an inch).
- Column 5: Weight (in pounds).
- Column 6: Length (in inches).

Note: All answers to this question should be presented as part of a typed document that is no more than 4 pages long summarizing results that cover the following:

-a- Thorough analysis of the dataset via logistic regression, deviance analysis through a partial likelihood approach for an appropriate sequence of hypotheses. Which variables have significant explanatory power?

-b- Division of the dataset into a training and test sets. Analysis via a classification tree and random forests. Which variables are important? Report of accuracy of the classification.

-c- Perform Fisher's Linear Discriminant Analysis on the continuous predictor variables.

-c- Comparison and contrast of results from these three procedures. Which procedure is preferable for this dataset and why?

-2- (50 points) Refer to the data file `Pheno.csv`. The file has phenotypic information for 110 patients of a clinic.

- Column 1: Height
- Column 2: Weight
- Column 3: Age
- Column 4: Average Systolic Blood Pressure
- Column 5: Total Blood Lipid concentration (mg/dl).
- Columns 6 & 7: Life style risk indices for heart disease. These are derived as continuous linear functions of variables that are not explicitly present in this dataset.

Note: Analyze the dataset thoroughly in a typed document that is no more than 4 pages long summarizing results that cover the following:

[1] Final Project Version: 2021-05-01 at 19:36

-a- Can we meaningfully reduce the dimension of the data set through PCA and SPCA? Can we rotate the results to help with interpretation.

-b- Are there clusters to the data? Suppose we know that there are 3 distinct clusters. Do any of the clustering methods we have studied enable you to identify them?