SMLDM HT 2014 - MSc Week 5 Practical

We will explore PCA, LDA, logistic regression and kNN for predicting the age of abalone (type of shellfish) from physical measurements.

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, can be used to predict the age instead. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem, but not included here. The data is at:

`http://www.stats.ox.ac.uk/%7Eteh/teaching/smldmHT2014/X.txt"`
`http://www.stats.ox.ac.uk/%7Eteh/teaching/smldmHT2014/Y.txt"`

You can get further information at:

`http://archive.ics.uci.edu/ml/datasets/Abalone`

You will need package `MASS`. Use `install.packages` to install it if it is not present on your system.

Include the R commands and programmes you wrote, as well as outputs and figures which help demonstrate the conclusions you drew from your analyses. Comment your code. Explain your answers, but be concise and clear.

1 Pre-processing

1. First we will identify extreme outliers in the dataset. Use a combination of `boxplot` and `scale` to visualise the distributions of the individual dimensions, hence identify the two extreme outliers in the dataset.
2. What are the indices of the two outliers? For the rest of the practical you should remove these two extreme outliers.

2 PCA

1. Compute the PCA of the data, using the correlation matrix instead of the covariance matrix.
2. How many principal components do you need to retain to explain 95% and 99% of the variance in the data respectively?

3 LDA, logistic regression and kNN

For the rest of the practical, quantise the age Y so that values < 10 are placed in one class, and values ≥ 10 in another.

1. Split the data into 50% training set, and 50% test set.
2. Train a LDA and a logistic regression model on the training set. Report the training and test confusion matrices of both methods.

3. Now apply kNN to the dataset. Try values of $k = 1, 3, 5, 11, 31, 51$. Report the training and test accuracies of the various kNN methods as well as LDA and logistic regression.

4 ROC Curve

1. Use the test set to produce ROC curves for both LDA and logistic regression. Which method has better performance?
2. Now try the same procedure but with a 90% training set, 10% test set split. Do the curves look different? Why is this?