MS1b: Statistical Data Mining

MSc Mini-project

Deadline: 12 noon, Monday 22 April 2013 (Week 1 Trinity Term).

Location: Hand in at the Statistics Department, 1 SPR.

For this miniproject, we will explore methods 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/datamining/X.txt"
http://www.stats.ox.ac.uk/%7Eteh/teaching/datamining/Y.txt"
```

You can get further information at:

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

You will at least need packages `MASS`, `nnet`, `rpart`, `randomForest`.

Use `install.packages` to install missing packages.

Typical lengths of projects are about 3000 words, with a maximum of 7000 words.

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.

Be concise but clear.

1 Pre-processing (10 marks)

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 project you should remove these two extreme outliers.
3. For questions 3 onwards, you should also normalise the dataset. Explain the normalisation you use.

2 PCA (15 marks)

1. Compute the PCA of the data. Compare the output of PCA using the covariance and the correlation matrices. For the rest of the question, use PCA computed using the correlation matrix.
2. How many principal components do you need to retain to explain 90% and 99% of the variance in the data respectively?
3. Present a biplot of the data.

3 Receiver Operating Curve (ROC) (20 marks)

For this question, quantise the response variable Y so that values less than 10 are placed in one class, and values above or equal to 10 in another.

1. Compare LDA, QDA and logistic regression. Splitting the dataset into equal sized training and test sets, train the models on the training set, and use the test set to produce ROC curves.
2. Which method has better performance?
3. Now try the same procedure but with a 90% training set, 10% test set split. Do the curves look different? Why is this?
4. If you were to use Fisher's LDA to reduce the dimensionality of the data, how many discriminant directions would you use?

4 K-NN (15 marks)

For this question, we will quantise the response variable Y so that values less than or equal to 8 are in one class, 9 or 10 in another, and 11 or above in a third class.

- Use 10-fold cross-validation to optimise for K in a K-NN algorithm.
- Plot the generalisation error as a function of K. What conclusions can you draw from the plot?
- Which value of K would you choose?

5 Comparing Methods (30 marks)

For this question, we use the same quantisation as the previous one: quantise the response variable Y so that values less than or equal to 8 are in one class, 9 or 10 in another, and 11 or above in a third class.

1. Using 10-fold cross-validation to compare the methods that you have learnt in the course. Include at least LDA, QDA, logistic regression, K-NN, CART and random forests.
2. For K-NN, CART and random forests, you will need to set parameters like the number of neighbours and the regularisation parameter α in CART. For the estimates of the generalization performance to be valid, **you cannot use the test sets in the cross-validation splits above to optimise these parameters**. Why is this?
3. A simple (non-cross-validation) approach to estimating the parameters is to split the data into 3 sets: a training set, a test set, and a *development set*. During the *development* phase, the training set is used for training the models as usual, and the development set used to estimate best parameter settings for each method (this plays the role of the test set in the lectures). Once the parameters are set, both training and development sets are used to train the models again, and the test set used to estimate generalisation error.

   How would you extend the simple approach above to estimate these parameters within a cross-validation framework? There is no correct answer, be creative but justify your choices. Implement your ideas.
4. Analyse the estimates of generalisation errors. Which method(s) seem to perform better?

6 Extensions (10 marks)

1. Try other methods, e.g. neural networks, boosting, regularised QDA, Naïve Bayes. You can also try combinations of methods, e.g. first using PCA or LDA to reduce dimensionality and rotate the observation vectors, before applying K-NN or random forests.
2. Instead of quantising the age of the abalone and performing classification, consider directly trying to estimate the age. Use the squared loss. What method(s) would you use, and how do they perform?