Exercise Sheet 1: Regression

From the previous set of exercises, we know that the abalone data set has a number of attributes that are well-correlated with each other. We will use regression to study models that predict other attributes.

The raw data is available from: `http://archive.ics.uci.edu/ml/datasets/Abalone`. A version in the Weka format is available at: `http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs910/exercises/abalone.arff`

Linear regression, multilinear regression and non-linear regression

1. Fit a simple linear regression model to give diameter as a function of length. Give the parameters of the model, and the correlation coefficient. Comment briefly on what the parameters of the model tell you about abalone (see `http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names` for a description of the attributes).

2. The dataset includes information about the total weight of each specimen, along with the weight of different pieces (e.g. shell). Fit a multilinear model to give the whole weight as a function of the shucked weight, viscera weight, and shell weight, and give its parameters.

   Common sense would suggest that the whole weight should be related to the sum of these weights. Looking at the model and the data, comment on whether this relation holds for the observations.

3. There is a relationship between the whole weight (fifth attribute) and diameter (third attribute). Try plotting these two attributes to see the shape. Based on this, fit the following models to the data, and for each report the correlation coefficient.

   (a) A simple linear model, weight = $a\cdot$ diameter + b

   (b) A quadratic model, weight = $a\cdot$ diameter + $b\cdot$ diameter^2 + c

   (c) A cubic model without lower order or constant terms, weight = $a\cdot$ diameter^3

   (d) An exponential model, $\log$ (weight) = $a\cdot$ diameter + b

   Based on these results, and the meaning of the model for the data, which would you pick to model this dependency and why? You may find it useful to plot the models over the data.

Continued overleaf.

Logistic Regression

4. The male and female abalone are quite hard to tell apart, so for this question we will try to build a model to tell whether a specimen is an infant (I) or an adult (M/F).

   Build a *logistic regression* model to predict this feature based on the following attributes:

   (a) Length only

   (b) Whole Weight only

   (c) Class Rings only

   (d) Length, whole weight, and class rings together.

   For each model, give the accuracy (percentage of training examples predicted correctly).

   Hint. You may find it helpful to modify the input dataset to recode the new class value.

5. For the last question, we return to the familiar adult data set. (`http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs910/exercises/adult.arff`, or `http://archive.ics.uci.edu/ml/datasets/Adult`). Build a logistic regression model for the attribute sex (M/F) using combinations of attributes from adult.data (try adding and removing attributes to see what happens). The aim is to find a model that balances simplicity with accuracy, so try to include as few variables as possible while giving an accurate result. Describe the final model you obtain, the steps you followed to reach it, and its accuracy for the task.

   (a) Which attributes can be removed from the data set without affecting the accuracy of the resulting model significantly (say, by at most 1%)? Give an argument why this might be the case for the attributes in question.

   (b) Why is relationship-status helpful?

   (c) Why is country=Holland weighted heavily?

Bonus question (for those wanting to explore further, not for credit):

What is the best regression model you can build to predict the number of rings (which is related to the age of the specimen) in the Abalone data set? Throwing all variables into multilinear regression obtains a regression coefficient of about 0.73: can you beat this? You may want to try transformations of some variables (logarithmic, polynomial) based on visualizing the data.