STATS 760 A Survey of Modern Applied Statistics

Assignment 2 2017

Due Monday April 10. Email to Alan by 5pm.

Q1.

The data set red.txt on the web page contains chemical measurements on 1355 different red wines, along with a subjective quality score. Your task is to develop a prediction formula for predicting the score from the chemical measurements.

You should try the following four methods (a) linear regression, possibly with some form of variable selection, (b) generalised additive models, (c) regression trees, and (d) neural networks.

In addition, calculate an estimate of the prediction error for your methods using cross-validation. You can get information on cross-validation from Lecture 3 (given on April 7 but available on the web page now). You may wish to experiment with different sizes for the hidden layer in the neural network.

[30 marks]

Q2.

You may remember the cherry tree data from STATS 330; a copy is available in the R330 package. Fit a tree to these data, and draw a diagram of the way the height x diameter plane is partitioned by the fitted tree. Mark the value of the fitted function on each region.

[10 marks]