Assignment 2, Math 3346, 2007

Lecturer: John Maindonald

August 21, 2007

This exercise will work with a credit approval dataset from the UCI machine learning website. Two files are available – **crx.data** which holds the data, and **crx.names** which holds limited documentation.

1. Give, in a readily assimilable from, brief summary desciptions of each of the data attributes.
   [1 mark]
2. Compare attributes between classes. Comment on anything that seems interesting or unusual. Do any of the attributes show clear differences, e.g., strong enough to be reproducible under bootstrap sampling
   [2 marks]
3. Carry out discriminant analyses using
   (a) `lda()` and/or `qda()`;
   (b) `randomForest()`.

   Which method seems to have the greatest predictive power?
   [3 marks]
4. Carry out the following checks:
   (a) Are some explanatory attributes dispensible?
   (b) Should any of the continuous variables be modeled, for `lda()` or `qda()` using splines or other non-linear response functions?
   (c) For `lda()` and `qda()`, are there any evident interaction effects?
   (d) Does the data show evidence of subgroup effects, perhaps to the extent that some subgroups should be examined separately?

   [4 marks]
5. Provide plots, one or more from use of `lda()`, and one or more from use of `randomForest()`, that give a two-dimensional representation of the data. Comment on the adequacy of these plots. Check their stability under bootstrap sampling, and report (do not give the graphs) what you have observed.
   [3 marks]
6. 2 further marks will be given for presentation and organization of material.

Due Date: September 18, 2007, 5pm

In addition to any R code that may be included in the main document, please provide the R code separately from the output. Please provide assigments in a pdf file, either as hard copy or emailed to john.maindonald@anu.edu.au