Assignment 2, Math 3346, 2008

Lecturer: John Maindonald

August 29, 2008

This exercise will work with the data set `Satellite` in the *mlbench* package. Be sure to examine its help page. Data are for a (small) sub-area of a scene, consisting of 82 x 100 pixels. Each line has data corresponds for a 3x3 square neighbourhood of pixels completely from the 82x100 sub-area. There are four ASCII values for each pixel; one for each of four spectral bands.'

In each line of data the four spectral values for the top-left pixel are given first followed by the four spectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom. Thus, the four spectral values for the central pixel are given by attributes 17, 18, 19 and 20.

If the data set seems too large for your computer equipment, try working with a subsample of perhaps half the data.

1. Run the following code, using it to get a quick overview of] the distribution of variable values:

```
> form <- formula(paste("classes ~", paste(paste("x.", 1:36, sep = ""),
+     collapse = "+")))
> bwplot(form, data = Satellite, outer = TRUE, layout = c(3, 6,
+     2))
```

   You may need to experiment with the layout parameters to get settings that are suitable for showing the graphs on a printed page. Comment on what you observe.
   [2 marks]

2. First, try classification using the central pixel, i.e., using attributes 17,18,19 and 20.

   (a) Compare the performance of the `lda()`, `qda()`, and `randomForest()` classifiers.
   [2 marks]

   (b) Plot the scores from use of `lda()`, and comment on what you see. (NB: For this, you will need to run `lda()` with `CV=FALSE` (the default)). Does the plot of scores give useful visual indications of which scores might be "easy" to classify, and which "hard"?
   [2 marks]

   (c) How adequate are the first two sets of scores as a summary of the analysis?
   [1 mark]

   (d) Repeat the `lda()` and `qda()` analyses with several different bootstrap samples, and comment on the indications this gives of the accuracy of the predictive accuracy estimates. For `randomForest()`, repeat the analysis several times, and comment. Why, for `randomForest()`, is it unnecessary to take repeated bootstrap samples?
   [3 marks]

3. Choosing whichever of the three algorithms achieved the best discrimination, run the analysis for the pixels on the four corners in turn – i.e. with pixel 1 (attributes 1, 2, 3 and 4), then pixel 3, then pixel 7 (attributes 25, 26, 27 and 28), and finally pixel 9. Comment on the likely reason for differences in predictive accuracy from use of these corner pixels, as opposed to use of the central pixel.
   [3 marks]

4. Repeat the analysis using all pixels, and using the `lda()` and `randomForest()` classifiers.

   (a) Comment on the predictive accuracies that are now achieved.
   [2 marks]

   (b) Plot the scores from use of `lda()`, and comment on what you see. (NB: For this, you will need to run `lda()` with `CV=FALSE` (the default)).
   [1 mark]

   (c) How adequate are the first two sets of scores as a summary of the analysis?
   [2 marks]

   (d) The accuracy that is returned is a specific form of average accuracy. For what mix(es) of soil types would you expect the accuracy be mauch higher than this? For what mix(es) would you expect it to be much lower? Give an example, in each case, of such a mix, calculating the expected predictive accuracy.
   [2 marks]

[**TOTAL: 20 marks**]

Due Date: September 19, 2008, 5pm

In addition to any R code that may be included in the main document, please provide the R code separately from the output. Marks will be subtracted if the R code is not provided.

Please provide assigments in a pdf file, either as hard copy or emailed to john.maindonald@anu.edu.au