Extension Assignment 2

ENGSCI255 Semester 1, 2018

This lab is due by 5pm on Monday 28 May. For each question hand in the commands you used and the output generated.

Download and read in the file `spamdata.csv` from Canvas; this file contains 57 dependent attributes (`A1-A57`), and 1 dependent attribute (`IsSpam`) in column 58.

1. Set the seed of the random number generator in R to 99
   (a) Use k-means clustering to cluster the data into two groups using all of the independent attributes, performing 20 repetitions.
   (b) Generate a table for the clustering, showing how well the clusters separate spam email from non-spam emails.
   (c) If you used this clustering as the foundation for a spam filter, what would be the *accuracy*, *sensitivity*, *specificity*, and *precision* of the filter. (For the purposes of this assignment something being spam will be the *positive* result.)

Set the seed of the random number generator to 50, and then generate a training data set of 2400 data points (the remaining data will be the test set).

2. For this question you will explore how certain parameters affect the performance of classification trees.
   (a) Loss Matrices
      i. Using the `rpart.control(...)` arguments for `rpart`, set the termination criteria for generating the classification tree to be a max depth of 10. (Disable any other termination criteria.)
      ii. By modifying the *loss matrix*, generate 6 classification trees (using all of the independent attributes), which range from having no *false positives* to no *false negatives* in the training data.
      iii. For each tree give both the *in-sample* (training data) and *out of sample* (test set) confusion matrix.
      iv. On a 2D scatterplot show the sensitivity vs. specificity of each classification model, include both the in-sample and out-of-sample performance in different colours.

(b) Tree Depth

    i. Generate six classification trees by varying the *max depth* termination criteria from 5 to 30 in steps of 5.

    ii. Plot the tree for a depth of 5.

    iii. Using a line graph, plot the *accuracy* of the classification model against the max depth; you should plot two lines, one for the in-sample accuracy, and one for the out-of-sample accuracy.

(c) Discuss the results of the two experiments above.

3. For this question you'll need to ensure that the spam classification is treated as a `factor` rather than a number, e.g.

   ```
   > spam$IsSpam = as.factor(spam$IsSpam)
   ```

   You may use the whole dataset for this analysis; you do not need to use training / test sets.

   Set the seed of the random number generator to 42.

   (a) Generate random forests with 10, 100 and 1 000 trees in order to classify the `spam` data set.

   (b) For each random forest compute the confusion matrix, and estimate the *accuracy*, *sensitivity*, *specificity*, and *precision* of the filter, if these were used as a spam filter.

   (c) For the random forest with 1 000 trees, plot the *importance* of the different attributes.

   (d) Plot a scatterplot matrix of the top 5 attributes, coloured by whether the data point is spam, and comment on any interesting relationships.

4. It is not practical for a spam filter to compute all of these attributes of every email; in the file `scantimes.csv` (available on Canvas from Friday 18 May), is the expected number of milliseconds that it will take to compute the attribute for a standard email. We wish to design the best spam filter possible which takes on average 50ms per email to compute the attributes.

   (a) Formulate a binary integer program in Excel, using the importance of different attributes from 3(c) above, and their expected computation times, in order to determine what attributes to collect and use in the spam filtering process.

   (b) Generate a random forest of 1 000 trees, using only the attributes that satisfies the 50ms requirement above.

   (c) Compute the confusion matrix of this optimised filter, and estimate the *accuracy*, *sensitivity*, *specificity*, and *precision* of the filter, if these were used as a spam filter.

   (d) Compare your results to that from Question 3.