

Classification Trees: Exercises/Assignments

We are going to use the R package `rpart`.

Exercise 1

Work through Sections 1 to 5 of Therneau & Atkinson 1997, An Introduction to Recursive Partitioning using the RPART routines, and reproduce their Examples 3.3 and 5.3. The data is available as `stagec.csv` on the course website. Information on the data is in `stagec.info`.

Note, because ploidy takes values in $\{1, 2, 3\}$, `rpart` assumes it is an ordinal variable and only considers the splits $\{1\}$ & $\{2, 3\}$ or $\{1, 2\}$ & $\{3\}$. It is however categorical, so to get `rpart` to consider the split $\{1, 3\}$ & $\{2\}$ we need make ploidy a factor. The appropriate command is

```
stagec$ploidy <- factor(stagec$ploidy, levels=1:3,
                       labels=c('diploid', 'tetraploid', 'aneuploid'))
```

Assignment 2

Answer the following questions about the model fitted to the stage C prostate cancer data in Therneau & Atkinson 1997.

1. What criteria were used to stop growing the tree?
2. How many cross-validation groups were used?
3. If you re-fit the model will the cross-validation errors change? Why?
4. Which impurity measure was used?
5. What is the misclassification rate for the model (on the training data)?
6. If you prune using $\alpha = 0.03$ how many leaves does the tree have?
7. What is the misclassification rate for the pruned tree?
8. What is the 1 - s.e. rule for choosing the pruning parameter?
9. How many surrogate splits are used to deal with missing data?
10. Using the pruned tree, what is your decision rule for whether or not the disease is likely to progress?

Assignment 3: Spam filtering

The data for this problem is in the file `spam.data` on the course website. Documentation is in the files `spam.info` and `spam.names`. The dataset consists of 4601 emails, 1813 of which are classified as spam (unsolicited commercial email). Each record has a 0-1 response (spam 1 or non-spam 0) and 57 predictor variables. The goal of the assignment is to develop a spam filter using a classification tree.

The file `spam.r` contains some useful R commands.

- A) Using the data set `spam.data`, construct a spam classification tree. Use 10-fold cross-validation to choose an optimal pruning parameter. What is your cross-validation estimate of the misclassification rate of the optimal tree? Plot a subtree of the optimal tree that has at most 8 terminal nodes and thus identify some of the more important variables for spam classification.

- B) The data `spam.data` has been split into a training data set `spam.train` with 3067 records and a test data set `spam.test` with 1534 records. Fit a classification tree using only the training set, using 10-fold cross-validation to choose the pruning parameter.
- 1) What is your estimated misclassification rate based on cross-validation?
 - 2) What is the misclassification rate of emails in the test set? (Compare with the rate above.)
 - 3) What proportion of spam was misclassified in the test set (false negatives)?
 - 4) What proportion of non-spam was misclassified in the test set (false positives)?
- C) Using your classifier as a spam filter. A desirable property of a spam filter is that it err on the side of caution: deleting a valid email is much worse than letting through a spam email. To construct such a spam filter we penalize the misclassification of non-spam emails more heavily than the misclassification of spam emails, using loss matrices. For the following two loss matrices build a classifier using the training data set then test it using the test data set.

$$L_1 = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, L_2 = \begin{pmatrix} 0 & 1 \\ 10 & 0 \end{pmatrix}$$

For each classifier report:

- 1) The total misclassification rate on the test data;
- 2) The spam misclassification rate;
- 3) The non-spam misclassification rate.

Of the three classifiers you have constructed on the training data set, which do you prefer, and why?

This exercise is adapted from the notes of J. Friedman,
<http://www.stanford.edu/class/stats202/>, accessed 10 August 2006.