Machine Learning 1

Data Science

BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Summer Semester 2018 **Prof. Tim Downie**

Workshop 12

Tree Models: regression and classification trees

There are many different R packages and data mining software for fitting tree models. For example you have already come across classification trees using RapidMiner

We will be using the R package `rpart`. James et al. uses the `tree` package. They are similar but `rpart` is more intuitive, a bit more comprehensive and there is a further package `rpart.plot` which gives somewhat nicer plots.

Exercise 1

Work through the Lab sections 8.3.1 (classification) and 8.3.2 (regression) in James et al. Pages 323 to 328, but instead of using the package `tree` adapt the code to use the `rpart` package instead. The changes required are listed below.

- Start the packages

  ```
  > library(rpart)
  ```

  and

  ```
  > library(rpart.plot)
  ```

- Instead of using the function `tree()` to grow the tree use `rpart()`

- The commands `plot(tree.obj)` and `text(tree.obj)` also work for `rpart` objects, but the function `rpart.plot(tree.obj)` gives much nicer diagrams (in my opinion). Start by using both methods, then use the plotting function you prefer.

- With `rpart` the cross-validation is already processed, we just need to access the results. Instead of `cv.tree` use

  `> printcp(obj)` and `> plotcp(obj)`

- To prune the tree instead of `cv.tree(obj,FUN=prune.misclass)` use

  ```
  > prune(obj,cp=??)
  ```

  using the first cp value which falls below the dotted line.

In general the settings in `rpart()` lead to a tree with less nodes than using `tree()`, as a result the `rpart()` tree is closer to the optimal tree, and pruning is not as essential. In these two example the prediction accuracy on the test data is slightly worse for the pruned tree, you should not assume that this is always the case.

Exercise 2 Brexit referendum results

In June 2016 the United Kingdom held a referendum to leave or remain in the European Union. The data have been published at a district level, the original data for each district is available from this website, but a user friendlier version is available in Moodle.

(a) Load the data using the function `load()`. The data frame is called `Brexit`.
Spend a few minutes exploring the data, for example:

  (i) How many districts are there?

  (ii) What are the variables in the data set?

  (iii) What proportion of voters voted leave and what proportion voted remain?

  (iv) What proportion of districts voted leave (over 50% of the votes in that district for Leave)?

  (v) Obtain a Boxplot for percentage of leave votes by Region.

(b) Split the data into a training and a test data set with 300 observations in the training set.

(c) Use the variables `Region`, `Electorate`, `Pct_Turnout`, `Votes_Cast` and `Pct_Rejected` to fit a regression tree fore the variable `Pct_Leave`. Plot the full tree, prune the tree and plot the pruned tree. Find the mean squared error for the pruned tree using the test data.

(d) Fit a classification tree for Leave `Status`. Repeat the steps in part (c) and obtain a confusion matrix for the test data.

Exercise 3 Written Exercise

Do Exercise 4 in James et al. on page 332 and 333.