



Stat 133, Spring 2011

Homework 4: XML, Clustering, and Classification Methods

Due Friday, April 15 at 11:59PM

XML, Clustering, and Classification Methods

1. Two applications where XML is already widely used are RSS (really simple syndication), and the distribution of television listings for use in home theater systems (XMLTV). In this assignment, we're going to use R to extract information from these readily available XML documents.

- (a) RSS is a format that is used to provide information about news stories, blog entries, or any other information that can be broken down into individual items. When you see a symbol like this  or this  on a web page, it indicates a link to an RSS feed.

Write a **function**, which accepts the URL of an RSS feed, and returns a data frame with the headlines (titles), descriptions, and URLs (links) of the objects in the feed. You can use the RSS feed of Yahoo headlines available at <http://rss.news.yahoo.com/rss/topstories> to develop your function. Then find at least two other RSS feeds, preferably at a site you're interested in, and use your function on them. (You can find some feeds at <http://www.google.com/support/news/bin/answer.py?hl=en&answer=59255>. You can turn any [bing.com](http://www.bing.com) search into an RSS feed by taking the URL for the search and appending `&format=rss`. Another way to find RSS feeds is by adding `filetype:rss` to a Google search.)

- (b) At <http://www.stat.berkeley.edu/classes/s133/listings.xml.gz>, there is a gzipped file containing five days of television listings for over 200 channels. This XML file has two separate trees - one which describes the channels and the other which describes the programs. You'll need to use `xmlElementsByTagName` to separate them before processing.

From these TV listings, use the XML library in R to create a data frame with columns for starting time, channel, title and description of each of the programs represented in the listings. Notice that some of this information is stored as attributes in the XML tags, so you'll need to use `xmlAttrs` to extract this information.

Once you've got the data frame set up, choose a show or topic that you're interested in, and use R to print the rows of the data frame corresponding to those programs that match your criteria.

2. Using the builtin data set `iris`, use the `rpart` library in R to build a classification tree to predict the value of `Species` for each observation. Write a sentence which describes the prediction rule that your `rpart` analysis produced.

Next, using the `lda` function in the `MASS` library, classify the species of the observations using linear discriminant analysis. Finally, using the `knn.cv` function from the `class` library, use kth-nearest neighbor classification on the species of the observations.

For each of the three methods, calculate the percent of observations that were misclassified in order to determine which method was most effective in classifying the observations.

3. At <http://www.stat.berkeley.edu/classes/s133/data/teeth.csv>, there is a CSV file containing the names of animals and the numbers of various types of teeth (both top and bottom incisors, canines, premolars, and molars) that they have. Use cluster analysis to find groups of animals that are similar, based on the numbers of different kinds of teeth that they have.

Note: If there are other data sets that you are interested that are appropriate for classification or cluster analysis, you can substitute them for the data sets in parts 2 and 3.

Your submission should be contained in a *single* pdf file, containing all the R code you used for the assignment. Email this file to me (s133@stat.berkeley.edu), with your name in the subject of the email message and on the report, by 11:59PM (plus or minus a minute) on the due date. Make certain to save a copy of your email submission.