Getting started hints - for students

This section provides a few brief hints for the student in how to begin thinking about analysing the data.

Project 1 - Analysis of 2012 Presidential election data

The first problem is essentially a clustering or grouping problem to find clusters of states which share similar characteristics. One feature in the data is that a number of the measured variables are in percentages, so you may want to think about the scales of the different variables when considering fitting any statistical models. The second set of problems relates to classification, where you want to find combinations of variables that can accurately predict electoral outcome. For the advanced chapter think about the fact that not all the measured variables may be relevant for clustering or classification, and relationships between predictors and response variables may not be linear or even possible to represent through a standard parametric model.

Reading material

1. Presidential Election Process. https://www.usa.gov/election
2. A Comprehensive Survey of Clustering Algorithms. Xu, D. and Tian, Y. (2015) Annals of Data Science 2 (2), 165–193. https://link.springer.com/article/10.1007/s40745-015-0040-1
3. Supervised machine learning: A review of classification techniques. Kotsiantis, S. B. (2007) Informatica, 31, 249-268. http://www.informatica.si/ojs-2.4.3/index.php/informatica/article/view/148/140

Project 2 - Assessing the impact of socio-economic factors on Presidential election voting in the USA in 2016

The problem, to start with, is manipulating data sets with large numbers of variables and merging records across different data sets (electoral outcomes and socio-economic variables). It will be useful to keep in mind that certain counties listed in one file may or may not appear in the other; also there may be counties in which either electoral or census data may have to be discarded for some reason, for example, missingness in some variables. Groups of demographic/economic variables may be related by nature and must be handled carefully (e.g. you may not be able to use them simultaneously in a regression-type model). Some variables measured are on widely different scales of magnitude from others and care must be taken to ensure that any analysis undertaken is not affected by this. Preliminary exploratory analyses would be invaluable in determining which variables to look at more closely for the later statistical analysis. Note that you might want to use either the data on winning party,

or proportion of votes won, as a response- they may not give identical results. Use of the `maps` package might be useful for visualisation of the distribution of variables across counties/states. For the advanced chapter, one option would be to source data from the earlier election (see, for example, US Presidential election data, 2012, from `https://electionlab.mit.edu/data`) and corresponding census data (`http://www.census.gov`) and determine whether socio-economic changes over the intervening period can explain the differences in election results between the two years.

Reading material

1. Presidential Election Process. https://www.usa.gov/election
2. A Comprehensive Survey of Clustering Algorithms. Xu, D. and Tian, Y. (2015) Annals of Data Science 2 (2), 165–193. https://link.springer.com/article/10.1007/s40745-015-0040-1
3. Supervised machine learning: A review of classification techniques. Kotsiantis, S. B. (2007) Informatica, 31, 249-268. http://www.informatica.si/ojs-2.4.3/index.php/informatica/article/view/148/140